# Forecasting Industrial Production Growth in Europe with Bayesian Autoregressive Models

Alexander Haider*[1]

[1]The New School for Social Research, New York, NY.

October 2019

**Abstract**

I study the forecasting performance of three different classes of univariate autoregressive models by predicting growth rates of industrial production in eight European Monetary Union economies. The models are estimated via Bayesian techniques and can be divided into autoregressive models with constant variance, autoregressive models with stochastic volatility, and a smooth transition autoregressive model. I evaluate the forecasting ability of a total of ten models – including a linear prediction pool – in terms of point forecasts and density forecasts. Although it has been argued in the literature that the parsimonious autoregressive model with constant variance cannot be improved upon with respect to industrial production point forecasts, I show that the other two model classes can improve point forecasts for certain countries and forecast horizons. Furthermore, density forecasts of autoregressive models with stochastic volatility can improve forecasts substantially for all countries.

*Keywords:* Bayesian Inference, Growth Forecasting, Non-linear Modelling, Stochastic Volatility
*JEL Classification:* C22, C53, E32, E37

*haidera@newschool.edu

# 1 Introduction

Forecasting the future path of an economy poses a complex task for economists and policy makers alike. Many decision problems have to be informed by accurate forecasts. At the same time uncertainty surrounding these forecasts has to be accounted for as well. Extracting signals from high-frequency data often constitutes an important first step in an efficient decision making process. A widely used indicator series of this kind is given by an index of industrial production. *Industrial Production Indices* (IPI) are collected by statistical agencies around the world as they provide relevant information for business cycle analysis. For example, Golinelli and Parigi (2007) show that monthly IPI are decisive variables in forecasting GDP for the G7 countries.

Analyzing and forecasting industrial production has a long tradition in empirical business cycle research and is still drawing a lot of attention. Recent studies very often present 'horse races' between models of varying degrees of complexity or different estimation methods; see, for example, Avdoulas and Bekiros (2018), Bulligan et al. (2010), Siliverstovs and van Dijk (2003), Silva et al. (2018) and Teräsvirta et al. (2005). The approach taken here also compares different models of varying degrees of complexity for forecasting monthly industrial production data. The benchmark model is given by an autoregressive model ($AR$) with a time-invariant variance. The main class of competitor models, on the other hand, allows for time-varying volatilities in the error structure. More precisely, I focus on autoregressive models with stochastic volatility due to their outstanding forecasting performance as documented in Alessandri and Mumtaz (2017) and Clark and Ravazzolo (2015).

Regarding forecasts for IPI it is conjectured that autoregressive models with stochastic volatility will deliver high quality forecasts. The volatile nature of industrial production indices is pointed out by Thury and Witt (1998) in their study of industrial production in Austria and Germany. At the same time, many studies analyze the evolution of IPI via non-linear time series models. In fact, in his seminal work on modelling smooth transition autoregressive ($STAR$) models, Teräsvirta (1994) analyzes industrial production in Germany via a logistic smooth transition autoregressive model ($LSTAR$). Before applying the non-linear model he tests and rejects linearity for the time series. As a consequence a linear autoregressive model is misspecified for the German industrial production index with possibly adverse effects for forecasting the series.

However, Marcellino (2004) finds that in cases where linear models outperform non-linear models in their forecasting ability, the latter usually yield forecasts of very low quality. In regime switching models this result may be due to a misclassification of observations leading

to inappropriate dynamic behavior in forecasts. As a consequence forecast errors may be larger than in a linear model, even if there exists strong evidence in favor of a non-linear model being the more accurate description of the data generating process (Dacco and Satchell, 1999).[1] Moreover, applying a non-linear model involves additional cost in implementation and estimation due to their higher degree of complexity. Thus, applying a non-linear model for forecasting represents a risky strategy with potentially high cost in case of failure.

Recently Avdoulas and Bekiros (2018) used a threshold autoregressive model for modelling IPI for three European countries – Germany, Italy, and Spain. They report positive forecasting results compared to an autoregressive model. I therefore investigate the forecasting potential of a non-linear model as well. Following the approaches of Chen and Lee (1995) and mainly Lopes and Salazar (2006), I estimate a $LSTAR$ model as an additional contender model.[2] Lastly, an $AR$ model with a rolling ten-year window (112 observations) is estimated as well. The rolling window autoregressive model constitutes a simple way of accounting for changing volatilities. Accordingly it offers a cost efficient alternative to modeling a volatile variance structure.

I concentrate on IPI of eight European Monetary Union (EMU) countries: Austria, Germany, Spain, France, Greece, Italy, the Netherlands, and Portugal and evaluate the forecasting performance of the models just described over an expanding data window starting in November 2002 and running until June 2019. I evaluate forecasts 1, 3, 6 and 12 month(s) ahead which results in 200 forecasts for each forecast horizon.

As stated above, forecasts are subject to uncertainty and this level of uncertainty has to be taken into account when evaluating them. A Bayesian approach lends itself to such a problem setting and allows for point and density forecasts in a natural way. Point forecasts of the cumulative IPI growth rates are evaluated via the Root Mean Square Error ($RMSE$). Density forecasts are assessed by Logarithmic Scores ($LogS$) and the Continuous Ranked Probability Score ($CRPS$).

Concentrating on univariate models entails certain advantages, but also certain risks. By focusing on past values of IPI only, useful information provided by other variables will be omitted. On the other hand, it is well established in the forecasting literature that parsimonious models work well, especially in the short run. I therefore focus on univariate models which allow for high quality forecasts – at least in the short run – while also being

[1] See Teräsvirta (2006) for potential pitfalls of forecasting economic time series with non-linear models.

[2] Note that Chen and Lee (1995) discuss threshold models with the indicator function as the transition function, while Lopes and Salazar (2006) focus on $LSTAR$ models. See Hubrich and Teräsvirta (2013) for an overview of non-linear models in macroeconomic research.

relatively easy to implement.

This paper contributes to the current IPI forecasting literature by testing Bayesian state-of-the art univariate time series models for a group of heterogeneous European economies.[3] Changes in volatility – due to possible changes in the structure of the World economy – are explicitly modeled by one class of contender models. Smooth regime switches in the conditional mean growth rate are accounted for by the $LSTAR$ model. Parameter uncertainty and the probabilistic nature of forecasts are taken into account as well. In line with previous research, I find that the $AR$ model performs well when evaluating point forecasts. However, the $AR$ model can be improved upon by more complex models with respect to point forecasts for certain countries and forecast horizons. Even more, concerning density forecasts the autoregressive models with stochastic volatility improve forecasts substantially. On the other hand, the $LSTAR$ model only improves forecasts episodically. Lastly, a linear prediction pool is formed based on the autoregressive models with stochastic volatility and the $LSTAR$ model. The prediction pool often improves forecasts over the $AR$ model, especially for density forecasts. At the same time the linear prediction pool is outperformed by the best country models most of the time.

The paper is organized as follows. The next section provides an overview of the data. Descriptive statistics and figures on industrial production and its monthly growth rate already indicate some volatility of the IPI. Section 3 presents the models used in this study. An evaluation of their forecasting abilities is presented in Section 4. Finally, Section 5 concludes.

## 2 Dataset

The dataset used in this study is provided by the OECD Main Economic Indicators (MEI). I use monthly data on *production of total industry* which represents an index with 2015 as its base year. The data starts in December 1969 and ends in June 2019, yielding 595 observations.[4]

The data is displayed in Figure 1. As can be seen the series are clearly trending. Therefore I follow the standard approach for forecasting IPI and compute monthly growth rates based on the first difference of the logarithm of the series. After computing the growth rates the

[3]Following De Santis and Cesaroni (2016) and Semmler and Haider (2018) the eight countries may be divided into European Monetary Union core countries – Austria, Germany, France, and the Netherlands – and periphery countries. The periphery group would consist of Spain, Greece, Italy, and Portugal.

[4]Note that for some countries IPI provided by the OECD start already in 1960. However, for other countries the data on industrial production starts considerably later. Therefore I disregard some observations at the beginning of the sample.

data is demeaned.[5] Figure 2 depicts the growth rates starting in January 1970 and ending in June 2019.

Furthermore summary statistics for all countries are provided in Tables 1-3. Here I split up the data in decades and provide statistics on minimum, maximum, median, mean, and standard deviation.

Tables 1-3 show that no uniform pattern exists across countries and decades. For example, while the range of growth rates – defined by minimum and maximum growth rates – is relatively small for Germany for the first sample period, its minimum and maximum growth rates are relatively large in the 1980s, when compared with other countries. The same pattern is observed for the Netherlands. Large changes in growth rates were seen in the 1970s in Austria, Spain, Italy and Portugal, while growth rates are relatively stable in France for all periods. In the 1980s the largest swings in growth rates are observed for Germany, Greece, and the Netherlands, while Greece stands out in the 1990s with respect to this metric. For the period between 2000 and the end of 2009 the Netherlands, Germany, and Greece show the largest negative growth rates. Portugal exhibits large swings in growth rates during this period. The same holds for Greece. Since 2010 Greece has shown again the largest swings in IPI growth rates. This doesn't come as a surprise given the Global Financial Crisis starting in 2007/2008 and the European debt crisis starting in 2009/2010. Large changes in industrial production are also observed in the Netherlands and in Portugal for this time period. Taking a look at the right part of Table 3, France stands out at the country with the smallest variation in growth rates over the whole sample. The other seven countries show much larger minimum and maximum values. Especially Greece and Italy stand out in this respect. But also the European core countries besides France – Austria, Germany, and the Netherlands – exhibit strong fluctuations in IPI growth over the sample.

The right part of Table 3 shows that median growth rates are relatively close to the mean for all countries when analyzing the whole sample period, indicating limited skewness of the data for the eight countries. Given that the data used in this study has been demeaned, column *Mean* represents the mean growth rates relative to the whole sample period for each decade. As can be seen from Table 1 relative growth rates were above average for all countries in the 1970s, but were below average for half of them in the 1980s. In the 1990s a similar pattern emerges with four countries showing negative relative growth rates. Relative growth rates turn negative for all countries in the the period 2000-2009. Again, this pattern

[5]This is mainly done to avoid zero values in the growth rates which can lead to problems for the stochastic volatility sampler discussed below. If zero values still exist after demeaning the series (or during the sampling procedure), then a small value, $\varepsilon = 0.0001$, is added to the data.

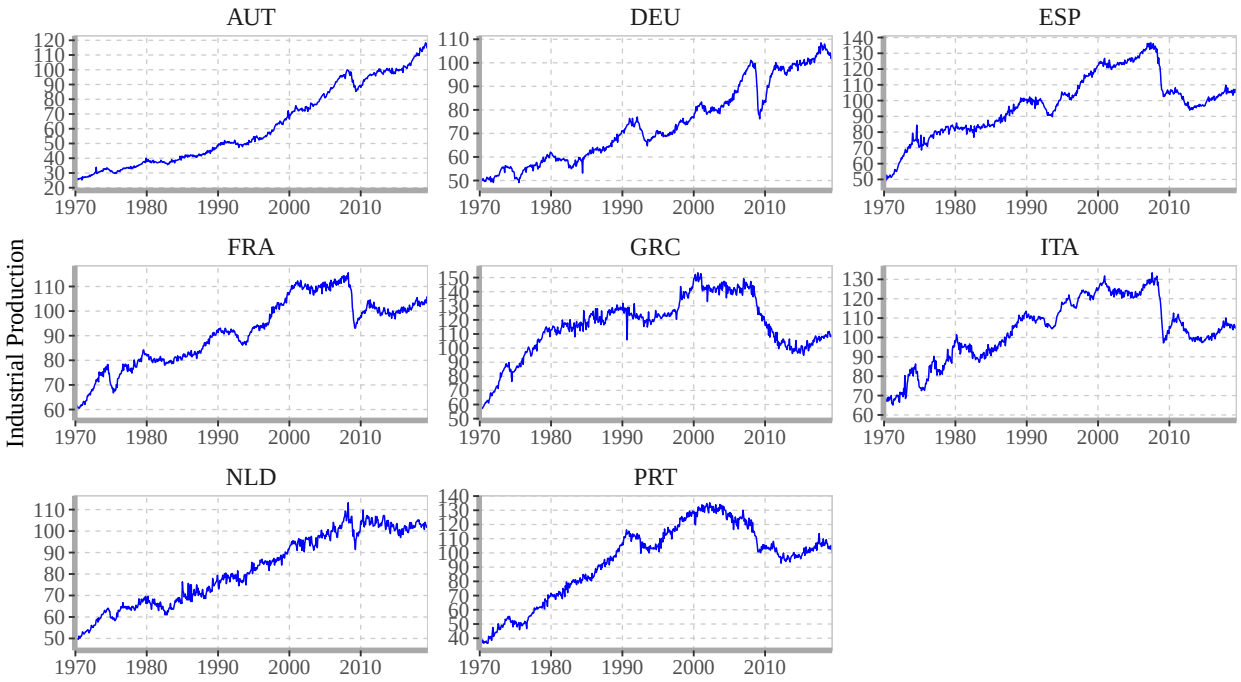


**Figure 1:** Industrial Production Indices; Countries: Austria (AUT), Germany (DEU), Spain (ESP), France (FRA), Greece (GRC), Italy (ITA), the Netherlands (NLD), and Portugal (PRT); Base Year = 2015; Source: OECD MEI.

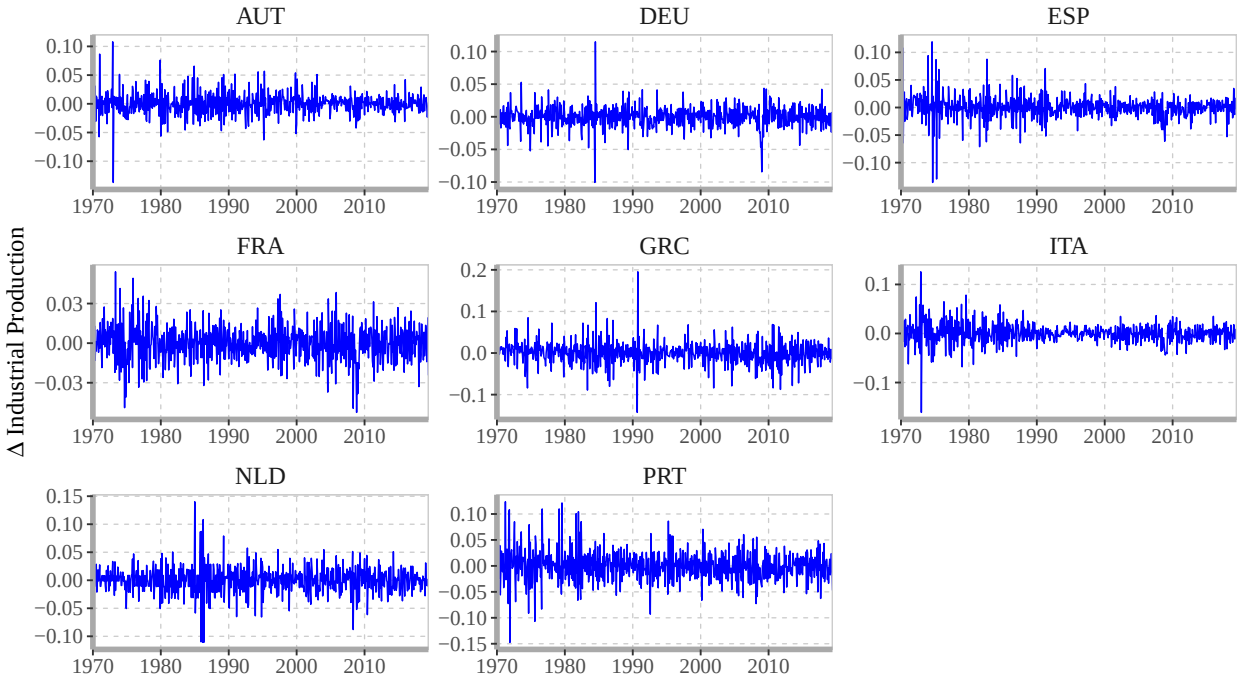


**Figure 2:** Log Growth Rates of Industrial Production (demeaned); Countries: Austria (AUT), Germany (DEU), Spain (ESP), France (FRA), Greece (GRC), Italy (ITA), the Netherlands (NLD), and Portugal (PRT); Base Year = 2015; Source: Author's own calculation based on OECD MEI.

| | 1970-1979 | | | | | 1980-1989 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Country | Min | Median | Max | Mean | Stdev | Min | Median | Max | Mean | Stdev |
| AUT | −13.700 | −0.092 | 10.795 | 0.193 | 2.651 | −5.627 | −0.196 | 6.555 | −0.107 | 2.292 |
| DEU | −5.221 | 0.114 | 5.278 | 0.079 | 1.682 | −10.064 | 0.158 | 11.488 | −0.009 | 2.080 |
| ESP | −13.588 | 0.507 | 11.911 | 0.281 | 3.340 | −7.108 | 0.036 | 8.761 | 0.053 | 2.362 |
| FRA | −4.889 | 0.142 | 5.405 | 0.191 | 1.755 | −3.211 | −0.096 | 2.525 | −0.002 | 1.190 |
| GRC | −8.404 | 0.617 | 8.489 | 0.498 | 2.454 | −8.914 | 0.073 | 12.114 | −0.011 | 3.087 |
| ITA | −16.091 | 0.028 | 12.576 | 0.307 | 3.553 | −6.317 | 0.067 | 5.839 | 0.027 | 2.117 |
| NLD | −5.054 | −0.127 | 4.705 | 0.147 | 1.709 | −11.083 | −0.127 | 14.007 | 0.023 | 3.612 |
| PRT | −14.735 | 0.271 | 12.355 | 0.346 | 4.254 | −6.619 | 0.036 | 10.435 | 0.180 | 3.052 |

**Table 1:** Descriptive Statistic for demeaned Industrial Production Growth in percentage; 1970-1979 and 1980-1989

| | 1990-2000 | | | | | 2000-2009 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Country | Min | Median | Max | Mean | Stdev | Min | Median | Max | Mean | Stdev |
| AUT | −6.290 | −0.127 | 5.673 | 0.013 | 2.029 | −4.221 | 0.094 | 5.099 | −0.053 | 1.580 |
| DEU | −3.429 | −0.122 | 4.156 | −0.055 | 1.368 | −8.441 | 0.103 | 4.368 | −0.058 | 1.794 |
| ESP | −5.179 | 0.046 | 7.064 | 0.031 | 1.763 | −6.123 | −0.007 | 3.791 | −0.252 | 1.572 |
| FRA | −2.684 | 0.006 | 3.695 | 0.024 | 1.178 | −5.246 | −0.135 | 3.837 | −0.174 | 1.531 |
| GRC | −14.275 | 0.060 | 19.529 | −0.017 | 3.182 | −8.401 | −0.326 | 6.228 | −0.281 | 2.665 |
| ITA | −3.099 | −0.093 | 2.304 | −0.018 | 0.966 | −4.395 | −0.013 | 3.649 | −0.258 | 1.659 |
| NLD | −6.553 | 0.177 | 5.745 | −0.012 | 2.359 | −8.796 | −0.078 | 5.470 | −0.046 | 2.309 |
| PRT | −9.300 | 0.095 | 8.626 | 0.007 | 2.609 | −7.267 | −0.456 | 7.055 | −0.334 | 2.792 |

**Table 2:** Descriptive Statistic for demeaned Industrial Production Growth in percentage; 1990-1999 and 2000-2009

| | 2010-2019 | | | | | Full Sample | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Country | Min | Median | Max | Mean | Stdev | Min | Median | Max | Mean | Stdev |
| AUT | −3.133 | −0.178 | 4.209 | −0.050 | 1.239 | −13.700 | −0.076 | 10.795 | 0 | 2.024 |
| DEU | −4.374 | −0.022 | 4.210 | 0.046 | 1.420 | −10.064 | 0.036 | 11.488 | 0 | 1.687 |
| ESP | −5.294 | −0.040 | 3.464 | −0.119 | 1.355 | −13.588 | 0.078 | 11.911 | 0 | 2.205 |
| FRA | −3.294 | 0.061 | 3.131 | −0.040 | 1.258 | −5.246 | 0.008 | 5.405 | 0 | 1.402 |
| GRC | −8.747 | −0.316 | 6.859 | −0.199 | 2.919 | −14.275 | 0.107 | 19.529 | 0 | 2.877 |
| ITA | −4.264 | 0.109 | 3.497 | −0.062 | 1.442 | −16.091 | 0.015 | 12.576 | 0 | 2.145 |
| NLD | −6.133 | −0.033 | 5.098 | −0.118 | 1.875 | −11.083 | −0.080 | 14.007 | 0 | 2.463 |
| PRT | −6.271 | −0.095 | 4.832 | −0.210 | 2.256 | −14.735 | −0.050 | 12.355 | 0 | 3.076 |

**Table 3:** Descriptive Statistic for demeaned Industrial Production Growth in percentage; 1990-1999 and Full Sample

is not surprising given the Global Financial Crisis in this period. Taking a step back and investigating Figure 1 again, it can be seen that industrial production decreased dramatically with the onset of the financial crisis. For the period starting in 2010, Table 3 shows that only Germany recovered in the sense that it was capable of achieving positive relative mean growth rates, while all other countries exhibit smaller than average growth rates for this period.

Lastly, a look at the standard deviations reported in Tables 1-3 reveals that *volatility* decreased for most countries over time. In fact, the standard deviation is smaller for all countries for the period 2010-2019 compared with the whole sample, with the exception of Greece. For Austria and Spain the standard deviation decreased monotonically over the five sample periods summarized in Tables 1-3, despite the increased volatility immediately after the financial crisis. This pattern does not hold for the other countries; in France the standard deviation increased in the period 2000-2009 compared with the 1990s, but started decreasing again afterwards. The Netherlands, on the other hand show very low levels of volatility at the beginning of the sample, which increased during the 1980s, but started falling again thereafter. Still, the period from 1970 until the end of 1979 remains the period with the lowest volatility level for the Netherlands. For Italy and Portugal industrial production was also more volatile in the early years on average. Finally, volatility is rather subdued in Germany over the entire sample, while Greece shows no decrease in volatility levels over time. Over the entire sample period Portugal and Greece exhibit the highest volatility levels. Interestingly, the Netherlands show the third highest level in volatility over the full sample, ahead of Spain and Italy. The lowest volatility levels for the whole sample are observed in Germany and France.

Given the changing nature of the volatility reported in Tables 1-3, models taking these changes into account are expected to render high quality forecasts. However, changes in IPI growth may also be driven by regime switches. Models accounting for these possibilities are discussed in the next section, while Section 4 investigates their forecasting performance in detail.

# 3 Forecasting Models

The benchmark model is an $AR$ model:

$$y_t = b_0 + \sum_{i=1}^{P} b_i y_{t-i} + v_t, \quad v_t \sim \mathcal{N}(0, \sigma^2). \tag{1}$$

I use a conjugate prior structure for the autoregressive parameters, $\boldsymbol{b} = (b_0, \ldots, b_P)^T$, and the error variance. More precisely, the prior is given by $p(\boldsymbol{b}, \sigma^2) = p(\boldsymbol{b}|\sigma^2)p(\sigma^2)$ with $p(\boldsymbol{b}|\sigma^2) \sim \mathcal{N}(\boldsymbol{b_0}, \sigma^2 \boldsymbol{B_0})$ and $p(\sigma^2) \sim \mathcal{IG}(c_0, C_0)$, where $\mathcal{IG}(\cdot, \cdot)$ is the inverse-gamma distribution with shape parameter $c_0$ and scale $C_0$. The prior distribution is non-informative with $\boldsymbol{b_0} = (0, \ldots, 0)^T$, $\boldsymbol{B_0} = \text{diag}\{10^2, \ldots, 10^2\}$, $c_0 = 1$, and $C_0 = 0.01$.[6] Using this conjugate prior yields the conditional posterior distributions

$$\boldsymbol{b}|\boldsymbol{y}, \sigma^2 \sim \mathcal{N}(\boldsymbol{b_1}, \boldsymbol{B_1}), \tag{2}$$

$$\sigma^2|\boldsymbol{y}, \boldsymbol{b} \sim \mathcal{IG}(c_1, C_1), \tag{3}$$

with $\boldsymbol{y} = (y_1, \ldots, y_T)^T$. The parameters of equations (2) and (3) are given by $\boldsymbol{b_1} = (\boldsymbol{X^T X} + \boldsymbol{B_0^{-1}})^{-1}(\boldsymbol{X^T y} + \boldsymbol{B_0^{-1} b_0})$, $\boldsymbol{B_1} = \sigma^2(\boldsymbol{X^T X} + \boldsymbol{B_0^{-1}})^{-1}$, $c_1 = c_0 + T/2 + P/2$, and $C_1 = C_0 + 0.5 \times ((\boldsymbol{y} - \boldsymbol{Xb})^T(\boldsymbol{y} - \boldsymbol{Xb}) + (\boldsymbol{b} - \boldsymbol{b_0})^T \boldsymbol{B_0^{-1}}(\boldsymbol{b} - \boldsymbol{b_0}))$. Matrix $\boldsymbol{X}$ is defined the usual way as a matrix including a column of ones and the lagged values $\boldsymbol{y_{t-1}}, \ldots, \boldsymbol{y_{t-p}}$.Given the convenient form of the conditional posterior distributions a Gibbs-sampling procedure can be used to sample iteratively from equations (2) and (3), where the starting value is chosen as $\sigma^2 = 1$.

As stated above, I estimate two versions of the $AR$ model. The first model, $AR_{rec}$, estimates the autoregressive model on an expanding (recursive) data window. The second model, $AR_{roll}$, utilizes a rolling data window containing data of the ten most recent years. Descriptive statistics in Section 2 revealed that the variances of the growth rates of the industrial production indices change over time. As a consequence, model $AR_{rec}$ might overestimate (underestimate) the variance of the industrial production growth rate, if observations in the more distant past are characterized by a larger (smaller) variance. The $AR_{roll}$ model can be regarded as a simple way of taking these changes in the variance structure into account. In

[6] I also tested shrinkage priors akin to the Minnesota prior for the $AR$ model. However, forecasting performance deteriorated.

addition, a rolling window might also work as a remedy against structural breaks which affect many macroeconomic time series; see Stock and Watson (1996) for the US and Marcellino (2002) for Euro area evidence.

The first class of contender models considered in this study takes changes in the variance structure explicitly into account. The autoregressive model with stochastic volatility (AR-SV) can be written as

$$\begin{aligned} y_t &= b_0 + \sum_{i=1}^{P} b_i y_{t-i} + v_t, \\ v_t &= \epsilon_t \; exp(h_t/2), \\ h_t &= \mu + \phi(h_{t-1} - \mu) + u_t, \end{aligned} \tag{4}$$

with $\epsilon_t \sim \mathcal{N}(0,1)$, $u_t \sim \mathcal{N}(0, \sigma_u^2)$, and $\epsilon_t$ and $u_s$ independent for $t, s \in \{1, \ldots, T\}$. Equation system (4) delineates a linear autoregressive model with a time-varying variance structure. The volatility terms, $exp(h_t/2)$, are modeled as a non-linear state space model with the logarithm of the squared volatilities – the latent states – following an $AR(1)$ process. Lastly, initial state, $h_0$, is assumed to be distributed according to the stationary distribution of the stochastic volatiliy process, i.e. $h_0|\mu, \phi, \sigma \sim \mathcal{N}(\mu, \sigma_u^2/(1-\phi^2))$.

In addition to the parameters already described for the $AR$ model, parameters $\boldsymbol{\theta} = (\mu, \phi, \sigma_u)^T$, as well as the volatility terms have to be sampled. As in Kim et al. (1998) and Kastner and Frühwirth-Schnatter (2014) I use an independent prior for $p(\boldsymbol{\theta})$ such that $p(\boldsymbol{\theta}) = p(\mu)p(\phi)p(\sigma)$. The priors of the individual components are given by

$$\begin{aligned} p(\mu) &\sim \mathcal{N}(b_\mu, B_\mu), \\ p\left(\frac{\phi+1}{2}\right) &\sim \mathcal{B}(a_\phi, b_\phi), \\ p(\sigma_u^2) &\sim \mathcal{G}(g_0, G_0), \end{aligned} \tag{5}$$

with $\mathcal{B}(\cdot,\cdot)$ being the beta distribution and $\mathcal{G}(\cdot,\cdot)$ representing the gamma distribution. I use non-informative priors for $\mu$ and $\sigma_u^2$ which are implemented by setting $b_\mu = 0$, $B_\mu = 1000$, $g_0 = 1/2$, and $G_0 = 1/2$. As pointed out by Kastner and Frühwirth-Schnatter (2014) the prior choices for $\mu$ and $\sigma_u^2$ are not influential in applied work and using a non-informative prior therefore uncontroversial. However, the prior choice for $\phi$ may influence the posterior

| Model | $a_\phi$ | $b_\phi$ | $E(\phi)$ | $VAR(\phi)$ |
|---|---|---|---|---|
| AR-SV1 | 22 | 0.5 | 0.956 | 0.061 |
| AR-SV2 | 62 | 5 | 0.851 | 0.064 |
| AR-SV3 | 95 | 13.5 | 0.751 | 0.063 |
| AR-SV4 | 120 | 25 | 0.655 | 0.063 |

**Table 4:** Hyper-parameters and implied expected values and variances for prior distribution of $\phi$.

distribution of the parameter to a large extent, especially if the data is not very informative about the persistence of the variance (Kim et al., 1998). Given the small to moderate sample size of most macroeconomic time series this poses a serious problem and setting appropriate hyper-parameters becomes crucial. I therefore follow the approach of Clark and Ravazzolo (2015) and estimate five different AR-SV models with relatively tight prior distributions for $\phi$ for the first four models. The hyper-parameters, as well as the implied expected values and variances, for $\phi$ for the first four models are summarized in Table 4. Finally, I also estimate a fifth autoregressive model with stochastic volatility (AR-SV5) with $a_\phi = b_\phi = 1$ which implies a uniform prior for $\phi$ on $(-1, 1)$.

The five AR-SV models are estimated based on Gibbs-sampling as well. Starting values have to be provided for $\boldsymbol{b}, \boldsymbol{\theta}$ and the logarithm of the squared volatilities, $\boldsymbol{h} = (h_1, \ldots, h_T)$. For $\boldsymbol{b}$ starting values are simply set equal to zero for all elements of the vector, while $\boldsymbol{\theta} = (0, 0.9, 0.1)^T$. Lastly, $\boldsymbol{h}$ is a zero vector of length $T$ initially.

Given these starting values, regression residuals – based on the first equation in system (4) – are computed and used for drawing $\boldsymbol{h}$ as well as $\boldsymbol{\theta}$. I follow the approach of Kastner and Frühwirth-Schnatter (2014) for drawing $\boldsymbol{h}$ and $\boldsymbol{\theta}$. They devise an algorithm for sampling the stochastic volatility terms and $\boldsymbol{\theta}$ based on the ancillarity-sufficiency interweaving strategy (ASIS) of Yu and Meng (2011) which yields an efficient sampler for the stochastic volatility terms. In addition, their MCMC algorithm samples all stochastic volatility terms instantaneously without using a loop over the stochastic volatility terms. Compared with single-move algorithms this allows for faster convergence to the stationary distribution as the correlation between samples decreases.

Given a sample for $\mathbf{h}$, I obtain a model with a known form of heteroscedasticity which can be removed by diving each element of $\boldsymbol{y}$ and each row of $\boldsymbol{X}$ by $exp(h_t/2)$. Using the same prior distribution for the autoregressive parameters as before and the normalized data

from the last step, the posterior distribution for $\boldsymbol{b}$ is given again by equation (2).[7]

Lastly, I also estimate the autoregressive model with stochastic volatility where the logarithm of squared volatilities follow a random walk (AR-SV-RW), i.e. $h_t = h_{t-1} + u_t$ and $u_t \sim \mathcal{N}(0, \sigma_u^2)$. With respect to the state equation described in equation system (4), the random walk assumption implies $\mu = 0$ and $\phi = 1$. Thus using the random walk model, only a prior for $\sigma_u$ has to be defined. It is given by $p(\sigma_i) \sim \mathcal{IG}(g_0, G_0)$. Again, an uninformative prior is used by setting $g_0 = 1$ and $G_0 = 0.01$. The Gibbs sampling algorithm is the same as for the AR-SV model, i.e. after sampling $\boldsymbol{h}$, heteroscedasticity is removed from the dependent variable and the predictors and the autoregressive coefficients can be sampled as in the $AR$ model. However, in case of the AR-SV-RW the single move independet Metropolis-Hastings algorithm of Jacquier et al. (1994) is applied for sampling the latent terms. In addition, I use different starting values for the stochastic volatility terms. They are given by the first difference of $\boldsymbol{y}$ squared. As the approach of Jacquier et al. (1994) utilizes a starting value two additional values have to be added for the starting values of $\boldsymbol{h}$ (the second additional value has to be provided because one value is lost by using first differences). These two values are simply set equal to the first and last value of the first difference of $\boldsymbol{y}$ squared. Finally, a small value, equal to the standard deviation of the data divided by 1000, is added to each observation to avoid zero values. Moreover, a prior has to be provided for the first observation when using the sampling approach by Jacquier et al. (1994). The prior is set as $h_0 \sim N(\mu_h, \sigma_h^2)$ with $\mu_h = -9$ and $\sigma_h^2 = 100$. The prior for $h_0$ is uninformative, reflecting the uncertainty regarding the stochastic volatility terms.

The last model considered here is the logistic smooth transition autoregressive model ($LSTAR$). The model can be written as[8]

$$y_t = (1 - G(\gamma, z, s_t)) \left[ b_0 + \sum_{i=1}^{P} b_i y_{t-i} \right] + G(\gamma, z, s_t) \left[ c_0 + \sum_{i=1}^{P} c_i y_{t-i} \right] + v_t, \quad v_t \sim \mathcal{N}(0, \sigma^2). \tag{6}$$

Here I focus on a model with two 'extreme' regimes, associated with parameters $b_0 +$

[7] Note that the autoregressive model with stochastic volatility samples an individual variance for each observation. Consequently equation (3) is not used here.

[8] It should be noted here that formulation (6) is only used for presentation. The model is estimated using the more convenient re-parameterization $y_t = w_t^T \eta + v_t$, where $w_t^T = (1, \ldots, y_{t-p}, G(\gamma, z, s_t)[1, \ldots, y_{t-p}])^T$, $\eta^T = (b_0, \ldots, b_p, \tilde{c}_0, \ldots, \tilde{c}_P)$ and $\tilde{c}_i = b_i + c_i$ as in Lopes and Salazar (2006). The re-parameterization won't affect the forecasts.

$\sum_{i=1}^{P} b_i y_{t-i}$ and $c_0 + \sum_{i=1}^{P} c_i y_{t-i}$, respectively.[9] I fix the number of extreme regimes to two, while the number of right-hand side variables is fixed to be equal to $P$ in both regimes. Note that these assumptions simplify the estimation procedure and could be relaxed.

The $LSTAR$ model is estimated due to the possibility of regime switches in the dynamic behavior of an economy. It is assumed that contractions and expansions are distinguished by differences in their dynamic behavior. This is accomplished by conditioning on past values, $y_{t-d}$, with respect to possible regime switches. The transition function, $G(\gamma, z, s_t)$, is given by the logistic function

$$G(\gamma, z, s_t) = \{1 + exp(-\gamma(s_t - z))\}^{-1}, \tag{7}$$

which is a continuous function bounded between 0 and 1. Parameter $\gamma > 0$ represents the smoothness parameter as it controls the steepness of the transition slope. As $\gamma$ approaches zero the model collapses to a linear model, while a large value of $\gamma$ results in an almost instantaneous transition, approaching the threshold autoregressive model in the limit. Variable $s_t$ is the transition variable and is given by $s_t = y_{t-d}$ with $d > 0$ being the delay parameter. For convenience it is assumed that $d \leq P$. Finally, $z$ is the threshold parameter.

Compared with the $AR$ model of equation (1), the $LSTAR$ model adds additional parameters $c_0, \ldots, c_P, \gamma, d, z$. The additional autoregessive parameters, $c_0, \ldots, c_P$, use the same priors already introduced above. Regarding the other priors I follow Lopes and Salazar (2006): $\gamma \sim \mathcal{G}(a_\gamma, b_\gamma), z \sim \mathcal{N}(m_z, \sigma_z^2)$ and the prior for $d$ is given by the discrete uniform distribution on $1, \ldots, P$. I set $a_\gamma = 5$ and $b_\gamma = 0.25$ for all countries and time periods which results in a rather uninformative prior with $E(\gamma) = 20$ and $VAR(\gamma) = 80$. For the threshold variable I set $m_z$ equal to the mean of a given sample. Thus, $m_z$ differs for each country and sample period, while $\sigma_z^2$ is set equal to the variance of the given sample.

The Gibbs sampling procedure for the $LSTAR$ model uses again a starting value of $\sigma^2 = 1$. Furthermore, starting values are needed for $\gamma, z, d$. The starting value for $\gamma$ is drawn from $\mathcal{G}(a_\gamma, b_\gamma)$, while the starting value for $d$ is drawn from the discrete uniform distribution described above. Lastly, the starting value for $z$ is equal to the mean of the sample. Given these starting values $s_t = y_{t-d}$, $G(\gamma, z, s_t)$ and the regressors for both regimes are determined

[9] I refer to extreme regimes here because in contrast to many other non-linear models, the STAR model does note exhibit hard thresholds. For example, in equation (6) if $G(\gamma, z, s_t) \in (0, 1)$ for some $t \in T$, then both extreme regimes are activated and the current regime is an affine combination of them. On the other hand, the model remains in the first extreme regime if $G(\gamma, z, s_t) = 0$, and in the second extreme regime, if $G(\gamma, z, s_t) = 1$.

for the first iteration.

The Gibbs sampler starts by drawing $b_0, \ldots, b_P$ and $c_0, \ldots, c_P$. After drawing the autoregressive parameters, $\sigma^2$ is sampled. The first two steps of the Gibbs sampler are again based on equations (2) and (3). In the next step values for $\gamma$ and $z$ are drawn jointly. This conditional bivariate distribution is of unknown form and a random-walk Metropolis step is used within the Gibbs sampler to draw and evaluate new values for $\gamma$ and $z$.

The candidate values, $\gamma_c$ and $z_c$, are drawn as $(\gamma_c, z_c)^T \sim \mathcal{N}\left((\gamma, z)^T, \left(\begin{smallmatrix} \sigma_c^2 & 0 \\ 0 & \sigma_c^2 \end{smallmatrix}\right)\right)$ with $\gamma, z$ representing the current values for the smoothness parameter. The values on the diagonal of the covariance matrix are initialized as $\sigma_c = 0.005$ and adjusted on the fly: if the acceptance probability is greater than 50%, then the covariance matrix of the proposal density is multiplied by 1.01. On the other hand, if the acceptance probability falls below 15%, then the covariance matrix of the proposal density is multiplied by 0.99. This adjustment procedure starts after iteration 100 and the adjustment frequency is 50. However, it is well known that such an adjustment procedure destroys the Markovian property of the chain. Therefore it is 'turned off' 100 iterations before the end of the burn-in phase.

Regarding the sampling procedure for the delay parameter, Chen and Lee (1995) show for the threshold model that the discrete uniform prior leads to a multinomial posterior distribution with probabilities equal to the likelihoods under $d = 1, \ldots, P$.

Finally, I also investigate the forecasting performance of a linear prediction pool. Geweke and Amisano (2011), among others, show that linear prediction pools often deliver high quality forecasts. This result holds for prediction pools with optimized weights, as well as for pools with constant weights. I therefore evaluate a linear prediction pool with equal (constant) weights for each model. I include the five AR-SV, the AR-SV-RW, and the $LSTAR$ model in the prediction pool. Thus weights are set equal to 1/7 for all models and do not change over time.

# 4 Empirical Results

As discussed in the last section all models are evaluated via a Gibbs sampling approach to facilitate comparison between the models. Each model is estimated based on 20,000 iterations with 15,000 burn-in iterations. Thus I retain 5,000 forecasts for each country and time period. Forecasts are obtained recursively based on the simulation results. The lag length $P$ is set equal to 12 for each country model. Using the same number of lags for all models might result in over-parameterization for the more complex models. Especially the $LSTAR$ model

might suffer from over-parameterization given that the number of autoregressive parameters is $2 \times (P+1)$ instead of $P+1$. Teräsvirta (1994) proposes starting with a large $STAR$ model and removing redundant lags during the specification stage. However, given the number of periods and countries evaluated here, this approach is not feasible. In addition, Teräsvirta et al. (2005) found some evidence that these zero restrictions might impair the forecasting performance of the model. I therefore estimate LSTAR models with 12 lags in each regime.

Before analyzing the forecasting performance of the nine models and the linear prediction pool, a short description of the forecast evaluation methods is provided. As stated in the introduction, I evaluate point and density forecasts. Point forecasts are evaluated via the root mean square error, defined as $RMSE_h^m = N^{-1} \sum_{t=1}^{N} \sqrt{(y_{t,h}^{obs} - \bar{y}_{t,h}^m)^2}$ for model $m$, horizon $h = \{1, 3, 6, 12\}$ and $t = 1, \ldots, N$ with $N = 200$ being the number of forecasts. Here $y_{t,h}^{obs}$ stands for the observed value in period $t+h$, while $\bar{y}_{t,h}^m$ represents the mean over the 5,000 retained forecasts for model $m$, forecast horizon $h$ and period $t$.

As already mentioned in the introduction, accounting for uncertainty of the forecasts is crucial. The Gibbs sampling approach used for all nine model allows for probabilistic forecasts based on the simulated predictive distribution $p(y_{t+k}|y_t)$. I use two proper scoring rules to assess these probabilistic forecasts; Logarithmic Scores, $LogS$, and the Continuous Ranked Probability Score, $CRPS$.[10] The logarithmic score for period $t$, forecast horizon $h$, and model $m$ is defined as

$$LogS(y_{t,h}^m) = -log(f_{t,h}^m(y_{i,t}^{obs})), \tag{8}$$

with $f_{t,h}^m(\cdot)$ denoting the predictive density for model $m$, horizon $h$, and time period $t$. Thus the observed value is evaluated based on the predictive density of the simulated forecats. The predicitve density has to be estimated based on the data obtained from the Gibbs algorithm. As in Krüger et al. (2019) I use kernel density estimation based on a Gaussian kernel with the rule-of-thumb bandwidth selection of Silverman (1986) to estimate the predictive density. Note that $LogS$ is defined as the negative of the log predictive density, yielding a negatively oriented penalty term.

However, logarithmic scores based on kernel density estimates are sensitive to outliers and therefore may result in distorted penalties if the observed value falls into the tail of the estimated distribution; see Krüger et al. (2019) for more details. The $CRPS$ is used as an

[10] Evaluating density forecasts is discussed in detail in Gneiting and Raftery (2007).

alternative for evaluating probabilistic forecasts. It is defined as

$$CRPS(F_{t,h}^m, y_t^{obs}) = \int_{-\infty}^{\infty} \left(F_{t,h}^m(z) - \mathbb{1}\{y_{i,h}^{obs} \leq z\}\right)^2 dz. \tag{9}$$

Here $\mathbb{1}\{\cdot\}$ is the indicator function and $F(\cdot)$ stands for the predictive CDF which is easily obtained from the simulated samples. As before in the case of the $RMSE$, I compute the average of the individual scores for $LogS$ and $CRPS$ over all 200 forecast samples. Forecasts are then evaluated by comparing each model with the benchmark model. The results are summarized in Tables 5-16.

Tables 5-8 report the $RMSE$ for the benchmark $AR_{rec}$ model and relative $RMSE$ for all the other models. The relative $RMSE$ for model $m$ is defined as $rRMSE_m = RMSE_m/RMSE_{AR_{rec}}$. Thus a value smaller than one signals a forecast improvement over the benchmark model.

Moreover, I compare the forecast performance of the eight contender models with the benchmark model via the modified Diebold-Mariano test of Harvey et al. (1997). Note that Monte Carlo evidence in Clark and McCracken (2001) shows that Diebold-Mariano test is undersized for nested models. Unfortunately this result may affect the autoregressive models with stochastic volatility as well as the $LSTAR$ model. Clark and West (2007) suggest a simple procedure to adjust the Diebold-Mariano test for such a setting.

However, my focus lies on evaluating forecast accuracy. Inference about which model may represent the more accurate description of the data generating process is not of interest here. In such a setting the Diebold-Mariano test remains useful as pointed out by Diebold (2015). I therefore apply the test by Clark and West (2007)[11] and the Diebold-Mariano test. Both tests are formulated as one-sided tests with the alternative hypothesis stating that the given contender model is outperforming the benchmark model. The results of the two tests are indicated in Tables 5-8. Significant results for the Diebold-Mariano test are shown next to the relative $RMSE$ in the superscript, while the results of the Clark-West test are shown as subscripts.

First of all, Tables 5-8 show that the $AR_{rec}$ model is a competitive forecasting model with respect to point forecasts for IPI. The more complex models do not outperform it substantially with respect to $rRMSE$. In fact, for Germany the recursive autoregressive model is the best one month ahead forecasting model, while the rolling window autoregressive

[11] The estimate of the variance of the test is based on the HAC estimator of Newey and West (1994).

| Model | AUT | DEU | ESP | FRA | GRC | ITA | NLD | PRT |
|---|---|---|---|---|---|---|---|---|
| $AR_{rec}$ | 0.010 | **0.012** | 0.011 | 0.011 | 0.019 | 0.013 | 0.016 | 0.019 |
| $AR_{roll}$ | **0.965** | 1.082 | **0.930**$^{*}$ | 1.042 | 1.043 | 1.008 | 1.007 | 1.026 |
| AR-SV1 | 0.995 | 1.009 | $0.979_{**}$ | 0.996 | **0.993** | $\mathbf{0.958}^{***}_{***}$ | $0.990_{*}$ | 1.002 |
| AR-SV2 | 0.995 | 1.010 | $0.979_{***}$ | 0.998 | 0.996 | $0.959^{***}_{***}$ | $0.988_{*}$ | 1.002 |
| AR-SV3 | 0.999 | 1.006 | $0.977_{***}$ | 0.996 | 0.997 | $0.962^{***}_{***}$ | $0.987_{*}$ | 1.003 |
| AR-SV4 | 0.995 | 1.004 | $0.978_{***}$ | 0.997 | 0.996 | $0.964^{***}_{***}$ | $\mathbf{0.983}^{*}_{**}$ | 1.004 |
| AR-SV5 | 0.999 | 1.005 | $0.980_{***}$ | **0.993** | 0.996 | $0.959^{***}_{***}$ | $0.989_{*}$ | 1.004 |
| AR-SV-RW | 1.008 | 1.020 | $0.983_{**}$ | 0.994 | 1.000 | $0.958^{***}_{***}$ | $0.998_{*}$ | $\mathbf{0.995}_{*}$ |
| LSTAR | 1.003 | 1.001 | 0.993 | 1.001 | 1.005 | $0.988^{**}_{**}$ | $0.996_{**}$ | 0.999 |
| Linear Pool | 0.998 | 1.005 | 0.980 | 0.996 | 0.997 | 0.961 | 0.989 | 1.001 |

Diebold-Mariano Test: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Clark-West Test: $_{***}p < 0.01$, $_{**}p < 0.05$, $_{*}p < 0.1$

**Table 5:** Relative mean square errors ($rRMSE$) **one month ahead**. The first line shows the average $RMSE$ for $AR_{rec}$ over the whole sample. All other rows show $rRMSE$; relative to the benchmark $AR_{rec}$ model. Values smaller than one signal an improvement over the benchmark model. Line *Linear Pool* presents the results of a pool with equal weights for models AR-SV1, AR-SV2, AR-SV3, AR-SV4, AR-SV5, AR-SV-RW, LSTAR. Bold entries show the best model for a country (excluding the pooled model). Diebold-Mariano test excludes the linear pool. Clark-West test excludes $AR_{roll}$ and the linear pool.

model outperforms all other models for Austria for the same horizon. The same result holds for Spain, where $AR_{roll}$ is the best forecasting model averaged over the 200 pseudo out-of-sample forecasts. For the other five countries the autoregressive models with stochastic volatility perform best in the very short run. The $LSTAR$ model, on the other hand, outperforms the linear model in Spain, Italy, the Netherlands, and Portugal for the one month ahead forecast. Lastly, the linear prediction pool delivers better forecasts than the linear model for all countries but Germany and Portugal with respect to $rRMSE$ for one month ahead forecasts.

As already mentioned above, I also use the Diebold-Mariano test and the Clark-West test to evaluate forecasting performances of the forecasting models.[12] Here a similar picture emerges: the benchmark model performs very well. In fact, for Austria, Germany, and France the contender models cannot outperform the benchmark model at any horizon with respect to the two tests. This result does not hold for the other five countries. As can be seen from Table 5, $AR_{roll}$ and the autoregressive models with stochastic volatility perform significantly better in Spain than the $AR_{rec}$ model. For Italy the AR-SV models perform very well in the short run as well with respect to the two tests. Significant test results are also reported for all

[12]I do not evaluate the Diebold-Mariano test and the Clark-West test for the linear prediction pool. In addition, the Clark-West test is not evaluated for the rolling window autoregressive model.

models for the Netherlands with the exception of the rolling window autoregressive model. In the last column of Table 5 it can be seen that the AR-SV-RW model improves forecasts for Portugal one month ahead over the benchmark model according to the Clark-West test.

| Model | AUT | DEU | ESP | FRA | GRC | ITA | NLD | PRT |
|---|---|---|---|---|---|---|---|---|
| $AR_{rec}$ | 0.016 | 0.018 | 0.015 | 0.014 | 0.023 | 0.017 | 0.022 | 0.024 |
| $AR_{roll}$ | 1.015 | 1.093 | 1.043 | 1.077 | 1.037 | 1.061 | $1.047^{**}$ | $1.001^{**}$ |
| AR-SV1 | **0.999** | 0.990 | $0.990^{*}_{**}$ | 0.992 | 0.989 | $\mathbf{0.958^{**}_{***}}$ | $1.017_{*}$ | $1.000_{**}$ |
| AR-SV2 | 1.003 | 0.991 | $\mathbf{0.987^{*}_{*}}$ | 0.990 | 0.989 | $0.959^{**}_{***}$ | $1.018_{*}$ | $1.003_{*}$ |
| AR-SV3 | 1.007 | 0.991 | $0.988^{*}_{*}$ | 0.989 | 0.990 | $0.961^{**}_{***}$ | $1.019_{**}$ | 1.003 |
| AR-SV4 | 1.003 | 0.987 | $0.988^{*}_{*}$ | 0.992 | 0.988 | $0.966^{**}_{**}$ | 1.020 | 1.005 |
| AR-SV5 | 1.005 | **0.986** | $0.989^{*}_{*}$ | **0.989** | **0.987** | $0.959^{**}_{***}$ | 1.021 | 1.003 |
| AR-SV-RW | 1.009 | 0.992 | $0.990^{**}_{**}$ | 0.991 | 0.998 | $0.960^{**}_{***}$ | 1.014 | $\mathbf{0.995^{*}_{**}}$ |
| LSTAR | 1.005 | 0.998 | 0.999 | 1.001 | 1.001 | 0.991 | **0.997** | 1.012 |
| Linear Pool | 1.004 | 0.990 | 0.989 | 0.990 | 0.991 | 0.962 | 1.015 | 1.002 |

Diebold-Mariano Test: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Clark-West Test: $_{***}p < 0.01$, $_{**}p < 0.05$, $_{*}p < 0.1$

**Table 6:** Relative mean square errors ($rRMSE$) **three months ahead**. The first line shows the average $RMSE$ for $AR_{rec}$ over the whole sample. All other rows show $rRMSE$; relative to the benchmark $AR_{rec}$ model. Values smaller than one signal an improvement over the benchmark model. Line *Linear Pool* presents the results of a linear pool with equal weights for models AR-SV1, AR-SV2, AR-SV3, AR-SV4, AR-SV5, AR-SV-RW, LSTAR. Bold entries show the best model for a country (excluding the pooled model). Diebold-Mariano test excludes the linear pool. Clark-West test excludes $AR_{roll}$ and the linear pool.

Table 6 shows the point forecast performance for a three months ahead forecast horizon. Here the AR-SV models perform best with respect to $rRMSE$ with the exception of the Netherlands, where the $LSTAR$ model yields superior forecasts. The linear prediction pools perform well for most countries, but do not improve over the best country model for any country.

Turning to the two formal prediction test it can be seen that significant forecasting improvements are again provided for Spain, Italy, and the Netherlands by the AR-SV models. For Portugal, $AR_{roll}$ and certain AR-SV models show significant results. A counter-intuitive result is reported for the Netherlands in Table 6: the $LSTAR$ model delivers the best results for $rRMSE$, but fails to show significant improvement with respect to the Diebold-Mariano and Clark-West tests. On the other hand, the Diebold-Mariano test shows significant results for $AR_{roll}$ and the Clark-West test reports significant improvements for AR-SV1, AR-SV2, and AR-SV3. Moreover, $rRMSE$ is larger than one for all four models. A similar result is observed for the three month horizon for Portugal. Here $AR_{roll}$, AR-SV1 and AR-SV2 show significant test results although their $rRMSE$ are larger than one.

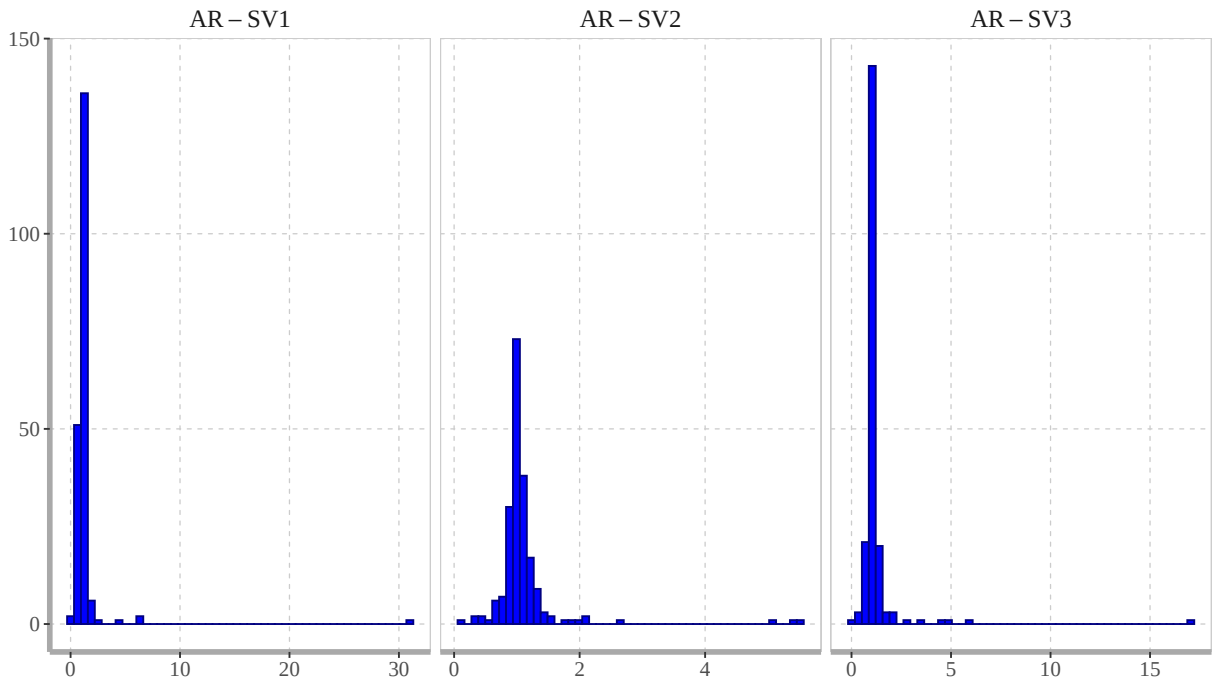


**Figure 3:** rRMSE for models AR-SV1, AR-SV2, and AR-SV3; **three months ahead** for the Netherlands.

This result is explained by a small number of large outliers for the $AR_{roll}$ and the AR-SV models which affect $rRMSE$ more strongly than the two forecast performance tests. For illustration purposes Figure 3 shows a histogram of $rRMSE$ for models AR-SV1, AR-SV2, and AR-SV3 for the Netherlands three months ahead. Here I take all 200 pseudo out-of-sample forecasts for the three autoregressive models with stochastic volatility and divide them by the forecasts of $AR_{rec}$. As can be seen from the Figure, all three models exhibit a small number of large outliers. As a consequence the average $rRMSE$ of the three models increases. Note that in Tables 5-8 average $RMSE$ are computed first for all models before the $rRMSE$ ratios are computed. This approach mitigates the effects of extreme outliers as shown in Figure 3, but cannot eliminate them. Therefore the Mariano-Diebold and Clark-West tests are applied as well to allow for improved inference regarding the point forecasts.

Table 7 shows the results for six months ahead forecasts. Again the autoregressive models with stochastic volatility dominate for most countries with respect to $rRMSE$. The only exceptions are Austria, where $AR_{rec}$ has the smallest $RMSE$, and the Netherlands where the $LSTAR$ model predicts best. The Diebold-Mariano test reports significant forecasting improvements for the autoregressive models with stochastic volatility for Spain, Greece, and Italy. The results for the Clark-West test are similar; improvements are shown for the

| Model | AUT | DEU | ESP | FRA | GRC | ITA | NLD | PRT |
|---|---|---|---|---|---|---|---|---|
| $AR_{rec}$ | **0.021** | 0.028 | 0.023 | 0.019 | 0.026 | 0.025 | 0.026 | 0.031 |
| $AR_{roll}$ | 1.061 | 1.134 | 1.143 | 1.145 | 1.007 | 1.153 | 1.061 | 1.009 |
| AR-SV1 | 1.007 | 0.979 | $0.990_{*}$ | 0.992 | 0.998 | $0.950^{**}_{**}$ | 1.032 | 0.991 |
| AR-SV2 | 1.009 | 0.980 | $0.991_{*}$ | 0.993 | 0.999 | $0.954^{**}_{*}$ | 1.037 | 0.990 |
| AR-SV3 | 1.013 | **0.978** | 0.989 | 0.996 | $0.997_{*}$ | $0.961^{*}_{*}$ | 1.038 | 0.994 |
| AR-SV4 | 1.010 | 0.981 | $0.992_{*}$ | 0.998 | **0.997** | $0.966^{*}_{*}$ | 1.036 | 1.001 |
| AR-SV5 | 1.009 | 0.978 | $0.992^{*}_{*}$ | 0.996 | 1.000 | $0.953^{**}_{**}$ | 1.038 | 1.000 |
| AR-SV-RW | 1.009 | 0.978 | $\mathbf{0.987^{*}_{*}}$ | **0.988** | $1.012^{***}_{***}$ | $\mathbf{0.950^{***}_{**}}$ | 1.020 | **0.982** |
| LSTAR | 1.007 | 0.994 | 1.002 | 0.995 | 1.007 | 0.997 | **0.998** | 1.024 |
| Linear Pool | 1.008 | 0.980 | 0.991 | 0.992 | 1.001 | 0.961 | 1.027 | 0.996 |

Diebold-Mariano Test: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Clark-West Test: $_{***}p < 0.01$, $_{**}p < 0.05$, $_{*}p < 0.1$

**Table 7:** Relative mean square errors ($rRMSE$) **six months ahead**. The first line shows the average $RMSE$ for $AR_{rec}$ over the whole sample. All other rows show $rRMSE$; relative to the benchmark $AR_{rec}$ model. Values smaller than one signal an improvement over the benchmark model. Line *Linear Pool* presents the results of a linear pool with equal weights for models AR-SV1, AR-SV2, AR-SV3, AR-SV4, AR-SV5, AR-SV-RW, LSTAR. Bold entries show the best model for a country (excluding the pooled model). Diebold-Mariano test excludes the linear pool. Clark-West test excludes $AR_{roll}$ and the linear pool.

autoregressive models with stochastic volatility for Spain, Greece, and Italy.[13] Finally, the last row shows the familiar picture that the linear prediction pool performs better for most countries than the benchmark model but does not improve upon the best model for any country.

Lastly, Table 8 summarizes the results for the twelve months ahead forecasts. In Austria the benchmark model maintains its superior forecasting performance, while in Greece the rolling window autoregressive model performs best. In the Netherlands, the $LSTAR$ model performs best as before. The $LSTAR$ model now also delivers the best predictions for Germany. For the other four countries the AR-SV models show the best results again. Once again, linear prediction pools cannot outperform the best country models for any country. Regarding the Diebold-Mariano test and the Clark-West test it can be seen that the number of occasions where the benchmark model is outperformed decreases. Significant results are reported for the Clark-West test for Greece for the AR-SV-RW model. Significant improvements with respect to the Clark-West test are also reported for two AR-SV models for Portugal. Finally, the $LSTAR$ model performs exceptionally well twelve months ahead for the Netherlands as can be seen in Table 8.

[13] Once again the counter-intuitive entry for Greece is explained by large outliers.

| Model | AUT | DEU | ESP | FRA | GRC | ITA | NLD | PRT |
|---|---|---|---|---|---|---|---|---|
| $AR_{rec}$ | **0.029** | 0.045 | 0.040 | 0.030 | 0.041 | 0.043 | 0.032 | 0.047 |
| $AR_{roll}$ | 1.166 | 1.218 | 1.259 | 1.309 | **0.983** | 1.250 | 1.114 | 0.978 |
| AR-SV1 | 1.032 | 0.997 | 0.998 | 0.994 | 1.004 | 0.962 | 1.082 | $0.973_{*}$ |
| AR-SV2 | 1.037 | 0.996 | 0.996 | 0.999 | 1.003 | 0.968 | 1.090 | $0.973_{*}$ |
| AR-SV3 | 1.036 | 0.994 | 0.996 | 1.008 | 0.999 | 0.975 | 1.088 | 0.981 |
| AR-SV4 | 1.031 | 0.996 | 0.997 | 1.013 | 1.003 | 0.984 | 1.083 | 0.992 |
| AR-SV5 | 1.031 | 0.996 | 0.995 | 1.000 | 1.007 | 0.964 | 1.084 | 0.994 |
| AR-SV-RW | 1.029 | 0.992 | **0.991** | **0.990** | $1.013_{**}$ | **0.958** | 1.047 | **0.957** |
| LSTAR | 1.008 | **0.991** | 0.995 | 0.997 | 1.004 | 0.998 | $\mathbf{0.989}^{***}_{**}$ | 1.057 |
| Linear Pool | 1.026 | 0.993 | 0.993 | 0.997 | 1.004 | 0.972 | 1.064 | 0.988 |

Diebold-Mariano Test: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Clark-West Test: $_{***}p < 0.01$, $_{**}p < 0.05$, $_{*}p < 0.1$

**Table 8:** Relative mean square errors ($rRMSE$) **twelve months ahead**. The first line shows the average $RMSE$ for $AR_{rec}$ over the whole sample. All other rows show $rRMSE$; relative to the benchmark $AR_{rec}$ model. Values smaller than one signal an improvement over the benchmark model. Line *Linear Pool* presents the results of a linear pool with equal weights for models AR-SV1, AR-SV2, AR-SV3, AR-SV4, AR-SV5, AR-SV-RW, LSTAR. Bold entries show the best model for a country (excluding the pooled model). Diebold-Mariano test excludes the linear pool. Clark-West test excludes $AR_{roll}$ and the linear pool.

Tables 9-12 report the logarithmic score of the benchmark model and the relative $LogS$ for all the other models. The relative logarithmic score, $rLogS$, for model $m$ is defined as $rLogS_m = LogS_m - LogS_{AR_{rec}}$. By subtracting the penalty term $LogS_{AR_{rec}}$ from $LogS_m$ a value smaller than zero defines an improvement in forecasting industrial production growth over the benchmark model.[14]

Table 9 shows that for the one month ahead forecast the AR-SV models perform best for all countries with respect to $LogS$. Improvements are substantial for Austria, Germany, Spain, France, and Italy. They are smaller for the Netherlands, Portugal, and especially Greece. The $LSTAR$ model on the other hand performs worse than the benchmark model for all countries, while the linear prediction pool outperforms the benchmark model for all countries. At the same time linear prediction pools cannot compete with the best country models. Finally, the rolling window autoregressive model does not predict well with respect to density forecasts. In Tables 5-8 it was shown that $AR_{roll}$ delivers high quality forecasts for most countries with respect to point forecasts. This result does not hold for density forecasts. The moving window autoregressive models cannot improve forecasts for any country at any

[14]Regarding density forecasts I only discuss relative scores without applying any formal tests. Note that the commonly used test by Amisano and Giacomini (2007) requires a rolling or fixed estimation window and is therefore not appropriate here.

| Model | AUT | DEU | ESP | FRA | GRC | ITA | NLD | PRT |
|---|---|---|---|---|---|---|---|---|
| $AR_{rec}$ | −2.534 | −2.509 | −2.483 | −2.615 | −2.231 | −2.453 | −2.329 | −2.197 |
| $AR_{roll}$ | 0.453 | 0.461 | 0.404 | 0.539 | 0.311 | 0.395 | 0.358 | 0.268 |
| AR-SV1 | −0.347 | −0.299 | **−0.419** | −0.238 | −0.084 | −0.334 | −0.149 | −0.110 |
| AR-SV2 | **−0.348** | **−0.301** | −0.410 | −0.261 | **−0.089** | −0.338 | −0.142 | −0.111 |
| AR-SV3 | −0.342 | −0.299 | −0.395 | −0.259 | −0.078 | −0.322 | −0.137 | −0.106 |
| AR-SV4 | −0.337 | −0.296 | −0.381 | **−0.266** | −0.087 | −0.311 | −0.137 | −0.096 |
| AR-SV5 | −0.332 | −0.291 | −0.410 | −0.255 | −0.084 | −0.341 | −0.139 | −0.105 |
| AR-SV-RW | −0.333 | −0.279 | −0.411 | −0.227 | −0.068 | **−0.341** | **−0.152** | **−0.129** |
| LSTAR | 0.011 | 0.005 | 0.004 | 0.009 | 0.009 | 0.0005 | 0.009 | 0.007 |
| Linear Pool | −0.317 | −0.276 | −0.377 | −0.252 | −0.086 | −0.310 | −0.136 | −0.106 |

**Table 9:** Relative logarithmic score ($rLogS$) **one month ahead**. The first line shows the average $LogS$ for $AR_{rec}$ over the whole sample. All other rows show $rLogS$; relative to the benchmark $AR_{rec}$ model. Values smaller than zero signal an improvement over the benchmark model. Line *Linear Pool* presents the results of a linear pool with equal weights for models AR-SV1, AR-SV2, AR-SV3, AR-SV4, AR-SV5, AR-SV-RW, LSTAR. Bold entries show the best model for a country (excluding the pooled model).

horizon with respect to $rLogS$. This is also the case for $rCRPS$ as shown below. The rolling window autoregressive model is therefore not further discussed with respect to density forecasts.

| Model | AUT | DEU | ESP | FRA | GRC | ITA | NLD | PRT |
|---|---|---|---|---|---|---|---|---|
| $AR_{rec}$ | −2.261 | −2.056 | −2.171 | −2.245 | −2.023 | −2.069 | −2.096 | −1.998 |
| $AR_{roll}$ | 0.628 | 0.537 | 0.560 | 0.605 | 0.370 | 0.502 | 0.464 | 0.336 |
| AR-SV1 | **−0.254** | 0.024 | −0.418 | −0.200 | −0.101 | −0.408 | −0.086 | −0.109 |
| AR-SV2 | −0.253 | 0.037 | −0.410 | −0.221 | −0.099 | **−0.414** | −0.075 | −0.094 |
| AR-SV3 | −0.247 | 0.212 | −0.380 | **−0.340** | −0.102 | −0.401 | −0.069 | −0.097 |
| AR-SV4 | −0.242 | **−0.058** | −0.358 | −0.309 | **−0.103** | −0.383 | −0.064 | −0.088 |
| AR-SV5 | −0.238 | 0.014 | −0.402 | −0.291 | −0.098 | −0.406 | −0.076 | −0.088 |
| AR-SV-RW | −0.242 | 0.664 | **−0.419** | −0.211 | −0.076 | −0.381 | **−0.101** | **−0.125** |
| LSTAR | 0.002 | 0.067 | 0.006 | 0.011 | 0.008 | −0.004 | 0.008 | 0.003 |
| Linear Pool | −0.236 | 0.070 | −0.375 | −0.331 | −0.095 | −0.374 | −0.078 | −0.098 |

**Table 10:** Relative logarithmic score ($rLogS$) **three months ahead**. The first line shows the average $LogS$ for $AR_{rec}$ over the whole sample. All other rows show $rLogS$; relative to the benchmark $AR_{rec}$ model. Values smaller than zero signal an improvement over the benchmark model. Line *Linear Pool* presents the results of a linear pool with equal weights for models AR-SV1, AR-SV2, AR-SV3, AR-SV4, AR-SV5, AR-SV-RW, LSTAR. Bold entries show the best model for a country (excluding the pooled model).

Results for the three months ahead forecasts for $LogS$ are very similar to one month ahead forecasts as shown in Table 10. Again, the AR-SV models perform best. However, for Germany the $AR_{rec}$ model is almost competitive compared with the best performing autoregressive model with stochastic volatility. Improvements are again substantial for Austria,

Spain, France, and Italy. They are somewhat smaller for Germany, Greece, the Netherlands, and Portugal. As before the linear prediction pools cannot compete with the best country models. In fact, in Germany, the linear prediction pool is even outperformed by the benchmark model. However, as stated above, for this particular horizon the $AR_{rec}$ model performs very well for Germany.

| Model | AUT | DEU | ESP | FRA | GRC | ITA | NLD | PRT |
|---|---|---|---|---|---|---|---|---|
| $AR_{rec}$ | −1.947 | −1.441 | −1.797 | −1.854 | −1.802 | −1.650 | −1.878 | −1.714 |
| $AR_{roll}$ | 0.674 | 0.390 | 0.621 | 0.609 | 0.383 | 0.542 | 0.512 | 0.336 |
| AR-SV1 | −0.237 | 0.190 | −0.328 | −0.176 | −0.142 | −0.191 | −0.099 | −0.099 |
| AR-SV2 | **−0.239** | 0.689 | −0.342 | −0.137 | −0.143 | −0.395 | −0.089 | −0.093 |
| AR-SV3 | −0.228 | −0.066 | −0.323 | −0.021 | **−0.146** | −0.388 | −0.047 | −0.085 |
| AR-SV4 | −0.227 | 0.265 | −0.291 | −0.207 | −0.142 | −0.372 | −0.073 | −0.071 |
| AR-SV5 | −0.224 | 0.961 | −0.336 | **−0.228** | −0.139 | **−0.396** | −0.089 | −0.078 |
| AR-SV-RW | −0.217 | 4.673 | **−0.347** | −0.176 | −0.144 | 0.062 | **−0.107** | **−0.129** |
| LSTAR | 0.001 | **−0.082** | 0.007 | 0.008 | 0.014 | −0.013 | 0.002 | 0.012 |
| Linear Pool | −0.230 | 0.249 | −0.324 | −0.380 | −0.133 | −0.369 | −0.092 | −0.089 |

**Table 11:** Relative logarithmic score ($rLogS$) **six months ahead**. The first line shows the average $LogS$ for $AR_{rec}$ over the whole sample. All other rows show $rLogS$; relative to the benchmark $AR_{rec}$ model. Values smaller than zero signal an improvement over the benchmark model. Line *Linear Pool* presents the results of a linear pool with equal weights for models AR-SV1, AR-SV2, AR-SV3, AR-SV4, AR-SV5, AR-SV-RW, LSTAR. Bold entries show the best model for a country (excluding the pooled model).

Table 11 reports results on $rLogS$ for the six months ahead forecasts. As for the two shorter forecasting horizons discussed before, the AR-SV models perform best for most countries. However, for Germany the best model is now given by the $LSTAR$ model, which outperforms the AR-SV3 model by a small margin. With respect to the linear prediction pool results it can be seen that a prediction pool forecasts badly for Germany where its performance is worse than the benchmark model. At the same time, the linear prediction pool finally outperforms the best country model for one country, namely France. The improvement in forecasting ability is not negligible. While the $rLogS$ is $-0.228$ for the best individual model in France, the AR-SV5 model, the relative score is $-0.38$ for the prediction pool, a substantial improvement.

Lastly, Table 12 summarizes information about the forecasting performance 12 months ahead with respect to $rLogS$. Again, the AR-SV models perform best for seven out of eight countries. The only exception is given by Germany as before. Here the benchmark model perform best now, although the $LSTAR$ model forecasts are almost of equal quality. The autoregressive models with stochastic volatility forecasts are of low quality for Germany. Especially the AR-SV-RW model shows a large relative logarithmic score. Regarding the

| Model | AUT | DEU | ESP | FRA | GRC | ITA | NLD | PRT |
|---|---|---|---|---|---|---|---|---|
| $AR_{rec}$ | −1.534 | **−1.185** | −1.273 | −1.383 | −1.379 | −1.172 | −1.604 | −1.302 |
| $AR_{roll}$ | 0.695 | 0.607 | 0.634 | 0.603 | 0.370 | 0.582 | 0.553 | 0.295 |
| AR-SV1 | −0.216 | 0.498 | −0.108 | 0.482 | −0.108 | 0.631 | **−0.085** | −0.097 |
| AR-SV2 | **−0.225** | 0.361 | 0.106 | **−0.207** | −0.110 | 0.022 | −0.074 | −0.088 |
| AR-SV3 | −0.198 | 0.315 | **−0.241** | 0.203 | **−0.115** | −0.164 | −0.047 | −0.070 |
| AR-SV4 | −0.216 | 0.249 | −0.094 | 0.263 | −0.106 | **−0.283** | −0.068 | −0.050 |
| AR-SV5 | −0.133 | 0.701 | −0.206 | 0.186 | −0.108 | 0.086 | −0.068 | −0.072 |
| AR-SV-RW | −0.087 | 5.137 | 0.254 | 0.306 | −0.095 | 1.881 | −0.061 | **−0.112** |
| LSTAR | 0.003 | 0.007 | −0.011 | 0.007 | 0.011 | −0.026 | −0.005 | 0.032 |
| Linear Pool | −0.226 | −0.204 | −0.269 | −0.384 | −0.103 | −0.319 | −0.080 | −0.077 |

**Table 12:** Relative logarithmic score ($rLogS$) **twelve months ahead**. The first line shows the average $rLogS$ for $AR_{rec}$ over the whole sample. All other rows show $rLogS$; relative to the benchmark $AR_{rec}$ model. Values smaller than zero signal an improvement over the benchmark model. Line *Linear Pool* presents the results of a linear pool with equal weights for models AR-SV1, AR-SV2, AR-SV3, AR-SV4, AR-SV5, AR-SV-RW, LSTAR. Bold entries show the best model for a country (excluding the pooled model).

linear prediction pools, it is shown that for a forecasting horizon of twelve months ahead they finally outperform the best country models for Germany, Austria, Spain, France, and Italy.

Table 12 suggests that taking changes in volatility into account does not improve long term forecasts for Germany. However, as already mentioned above, logarithmic scores can become very sensitive to outliers. In fact, this seems to be the case here. While the AR-SV-RW model is the worst performer with respect to $rLogS$, Tables 15 and 16 report that the AR-SV-RW model is the best model for forecasting IPI for Germany six and twelve months ahead with respect to $CRPS$. I therefore investigate twelve months ahead forecasts for Germany in more detail. The results are depicted in Figure 4. There the forecasts of the AR-SV-RW, the $AR_{rec}$, and the $LSTAR$ model are illustrated. The minimum and maximum values of the forecasts are represented by grey shaded areas. For illustrative purposes the grey shaded areas are only shown until January 2010. In addition the light blue shaded areas depict the 90% forecast intervals for the three models, while the dark blue shaded areas show the 68% intervals for the simulated forecasts. The onset of the Great Financial Crisis is dated with September 2008 and shown in Figure 4 as a vertical dashed red line.

As can be seen from Figure 4 the forecasts for the $LSTAR$ model and the $AR_{rec}$ model are very similar for the entire forecasting period. This is also reflected in $rLogS$ for the $LSTAR$ model which is only slightly above the score for the recursive autoregressive model which performs best for twelve months ahead forecasts for Germany. Furthermore, it is remarkable that forecast intervals do not change much over time for Germany. Note that Figure 4 shows

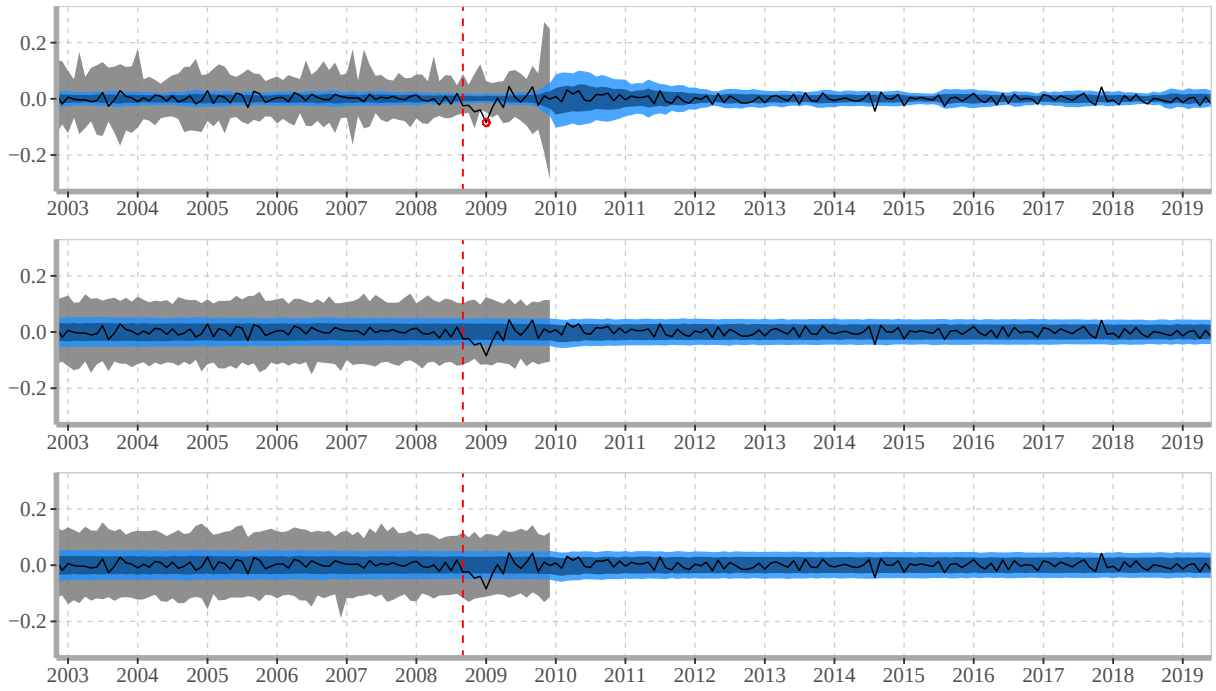


**Figure 4:** Quantiles for models AR-SV-RW (top), $AR_{roll}$ (middle), and $LSTAR$ (bottom); **twelve months ahead** for Germany. The grey shaded area depicts the minimum and maximum values of the sampled forecasts. The light blue shaded area shows the 90% forecast interval, the dark blue shaded area depicts the 68% forecast interval. The black line signals the realized values. The vertical dashed red line shows the onset of the Great Financial Crisis in September 2008. The red dot signals the observed value for January 2009.

the results of 200 estimated models. On the other hand, it was shown in Section 2 that Germany is distinguished by low levels of volatility in its IPI growth rates. Only the 1980s are standing out as a more volatile time period in Germany. As a consequence forecast intervals drawn from 200 posterior predictive distributions do not change much over time.

Taking a look at the first subfigure in Figure 4 it can be seen that the 68% and 90% posterior predictive intervals for the AR-SV-RW model are tighter than for the other two models. A very important feature of the AR-SV-RW model is shown as a red dot for January 2009. It can be seen that the observed value is smaller than smallest draw from the posterior predictive distribution. It has been argued above that logarithmic scores are very sensitive to such outliers. Indeed, the AR-SV-RW fails dramatically for the period after the financial crisis in Germany as volatility increased substantially. This result is not surprising given the tight forecasting interval for Germany due to Germany's low volatility levels in IPI growth before the crisis.

The other two models depicted in Figure 4 have severe problems in accouting for the steep drop in industrial production for January 2009 as well; the drop in IPI growth falls outside

the 90% forecast interval for both models, although the $LSTAR$ model performs slightly better than the $AR_{rec}$ model. However, due to their larger forecasting interval dispersion, the observed value does not fall below the smallest predicted value, which leads to improved logarithmic scores compared with the AR-SV-RW model. At the same time, $CRPS$ rewards concentrated forecast distributions, while outliers are not penalized as stronly as for $LogS$ (Krüger et al., 2019). Consequently the AR-SV-RW model performs better for $rCRPS$ than for $rLogS$ in Germany for the twelve months ahead forecasts.

Finally, I investigate the predictive capabilities for all models with respect to the second scoring rule, namely $CRPS$. The relative $CRPS$ for model $m$ is given by $rCRPS_m = CRPS_m/CRPS_{AR_{rec}}$. Similar to the case of $rRMSE_m$, a value smaller than one for $rCRPS_m$ signals an improvement in forecasting over the recursively estimated autoregressive model. The results for the $CRPS$ are reported in Tables 13-16.

| Model | AUT | DEU | ESP | FRA | GRC | ITA | NLD | PRT |
|---|---|---|---|---|---|---|---|---|
| $AR_{rec}$ | 0.009 | 0.010 | 0.010 | 0.009 | 0.014 | 0.010 | 0.012 | 0.014 |
| $AR_{roll}$ | 1.387 | 1.389 | 1.318 | 1.474 | 1.199 | 1.299 | 1.238 | 1.165 |
| AR-SV1 | **0.837** | 0.887 | **0.819** | 0.897 | **0.964** | 0.842 | 0.943 | 0.959 |
| AR-SV2 | 0.837 | 0.886 | 0.819 | 0.895 | 0.966 | **0.841** | 0.943 | 0.959 |
| AR-SV3 | 0.838 | **0.886** | 0.822 | 0.896 | 0.969 | 0.846 | 0.942 | 0.959 |
| AR-SV4 | 0.839 | 0.886 | 0.825 | 0.896 | 0.966 | 0.851 | **0.941** | 0.962 |
| AR-SV5 | 0.841 | 0.887 | 0.820 | **0.895** | 0.967 | 0.842 | 0.944 | 0.960 |
| AR-SV-RW | 0.848 | 0.899 | 0.821 | 0.897 | 0.974 | 0.842 | 0.946 | **0.951** |
| LSTAR | 1.007 | 1.003 | 1.002 | 1.007 | 1.006 | 0.996 | 1.001 | 1.004 |
| Linear Pool | 0.848 | 0.888 | 0.826 | 0.895 | 0.966 | 0.848 | 0.944 | 0.958 |

**Table 13:** Relative CRPS ($rCRPS$) **one month ahead**. The first line shows the average $CRPS$ for $AR_{rec}$ over the whole sample. All other rows show $rCRPS$; relative to the benchmark $AR_{rec}$ model. Values smaller than one signal an improvement over the benchmark model. Line *Linear Pool* presents the results of a linear pool with equal weights for models AR-SV1, AR-SV2, AR-SV3, AR-SV4, AR-SV5, AR-SV-RW, LSTAR. Bold entries show the best model for a country (excluding the pooled model).

Results for the one month ahead forecasts for $rCRPS$ are similar to $rLogS$. Table 13 shows that the AR-SV models perform very well. The $LSTAR$ model predicts worse than the recursively estimated autoregressive model for all countries, although the relative forecasting performance is almost as good as for the linear model. Still, as discussed in the introduction, given the higher cost of implementing and estimating the $LSTAR$ model this result hardly justifies using the $LSTAR$ model for density forecasts with few exceptions as discussed above for $rLogS$ and below for $rCRPS$. Regarding the linear prediction pool it can be seen again

that the pool cannot outperform the best country models. For France the linear prediction pool performs equally well as the AR-SV5 which dominates for one month ahead forecasts in France.

| Model | AUT | DEU | ESP | FRA | GRC | ITA | NLD | PRT |
|---|---|---|---|---|---|---|---|---|
| $AR_{rec}$ | 0.012 | 0.015 | 0.014 | 0.012 | 0.017 | 0.015 | 0.016 | 0.017 |
| $AR_{roll}$ | 1.591 | 1.549 | 1.503 | 1.603 | 1.246 | 1.451 | 1.335 | 1.202 |
| AR-SV1 | **0.897** | 0.875 | 0.827 | 0.850 | 0.955 | 0.813 | 0.983 | 0.961 |
| AR-SV2 | 0.898 | 0.876 | 0.829 | 0.850 | 0.956 | **0.812** | 0.987 | 0.965 |
| AR-SV3 | 0.901 | 0.876 | 0.838 | 0.851 | 0.955 | 0.818 | 0.988 | 0.967 |
| AR-SV4 | 0.903 | 0.878 | 0.845 | 0.854 | **0.954** | 0.826 | 0.989 | 0.971 |
| AR-SV5 | 0.903 | 0.877 | 0.829 | **0.849** | 0.954 | 0.814 | 0.988 | 0.969 |
| AR-SV-RW | 0.903 | **0.865** | **0.824** | 0.852 | 0.971 | 0.815 | **0.980** | **0.954** |
| LSTAR | 1.004 | 1.000 | 1.002 | 1.007 | 1.005 | 0.995 | 1.001 | 1.009 |
| Linear Pool | 0.903 | 0.877 | 0.838 | 0.855 | 0.958 | 0.826 | 0.982 | 0.965 |

**Table 14:** Relative CRPS ($rCRPS$) **three months ahead**. The first line shows the average $CRPS$ for $AR_{rec}$ over the whole sample. All other rows show $rCRPS$; relative to the benchmark $AR_{rec}$ model. Values smaller than one signal an improvement over the benchmark model. Line *Linear Pool* presents the results of a linear pool with equal weights for models AR-SV1, AR-SV2, AR-SV3, AR-SV4, AR-SV5, AR-SV-RW, LSTAR. Bold entries show the best model for a country (excluding the pooled model).

Turning to three months ahead forecasts, Table 14 reports again that the AR-SV models perform best. The $LSTAR$ model show some promise in Italy, but is outperformed by autoregressive models with stochastic volatility as before. Neither can the linear prediction pool improve forecasts over the best country models. Regarding the six months forecasts Table 15 shows similar results to Table 14; the autoregressive models with stochastic volatility perform best. The $LSTAR$ model cannot compete with them and only outperforms the benchmark model for Italy. The linear prediction pool shows high quality forecast results but does not perform as well as the best country models.

Finally, Table 16 reports the results for $rCRPS$ for the twelve months ahead forecasts. Here the AR-SV-RW model dominates for six countries. Only for France and Greece two other AR-SV models outperform the AR-SV-RW model although the margins are small. Again, the $LSTAR$ model shows promising results for Italy where it outperforms the benchmark model, although it cannot compete with the AR-SV-RW model. This time the $LSTAR$ model also performs better than the benchmark model in the Netherlands. In fact, the $LSTAR$ model is the second best model for the Netherlands for twelve months ahead forecasts. Its forecast performance is almost on a par with the AR-SV-RW model.

| Model | AUT | DEU | ESP | FRA | GRC | ITA | NLD | PRT |
|---|---|---|---|---|---|---|---|---|
| $AR_{rec}$ | 0.017 | 0.023 | 0.020 | 0.018 | 0.020 | 0.023 | 0.019 | 0.023 |
| $AR_{roll}$ | 1.696 | 1.636 | 1.615 | 1.662 | 1.265 | 1.561 | 1.411 | 1.199 |
| AR-SV1 | **0.903** | 0.880 | 0.859 | 0.818 | 0.936 | 0.824 | 0.988 | 0.958 |
| AR-SV2 | 0.906 | 0.880 | 0.865 | **0.817** | 0.936 | 0.826 | 0.991 | 0.960 |
| AR-SV3 | 0.910 | 0.880 | 0.872 | 0.819 | **0.935** | 0.837 | 0.995 | 0.965 |
| AR-SV4 | 0.910 | 0.884 | 0.882 | 0.823 | 0.936 | 0.847 | 0.992 | 0.973 |
| AR-SV5 | 0.910 | 0.883 | 0.865 | 0.819 | 0.940 | 0.827 | 0.993 | 0.972 |
| AR-SV-RW | 0.904 | **0.871** | **0.857** | 0.821 | 0.944 | **0.824** | **0.974** | **0.945** |
| LSTAR | 1.002 | 0.999 | 1.003 | 1.003 | 1.008 | 0.996 | 1.001 | 1.019 |
| Linear Pool | 0.909 | 0.883 | 0.871 | 0.825 | 0.942 | 0.841 | 0.984 | 0.965 |

**Table 15:** Relative CRPS ($rCRPS$) **six months ahead**. The first line shows the average $CRPS$ for $AR_{rec}$ over the whole sample. All other rows show $rCRPS$; relative to the benchmark $AR_{rec}$ model. Values smaller than one signal an improvement over the benchmark model. Line *Linear Pool* presents the results of a linear pool with equal weights for models AR-SV1, AR-SV2, AR-SV3, AR-SV4, AR-SV5, AR-SV-RW, LSTAR. Bold entries show the best model for a country (excluding the pooled model).

| Model | AUT | DEU | ESP | FRA | GRC | ITA | NLD | PRT |
|---|---|---|---|---|---|---|---|---|
| $AR_{rec}$ | 0.025 | 0.038 | 0.034 | 0.029 | 0.032 | 0.037 | 0.025 | 0.035 |
| $AR_{roll}$ | 1.809 | 1.692 | 1.702 | 1.723 | 1.243 | 1.676 | 1.497 | 1.160 |
| AR-SV1 | 0.902 | 0.915 | 0.888 | **0.817** | 0.953 | 0.853 | 1.007 | 0.952 |
| AR-SV2 | 0.904 | 0.914 | 0.893 | 0.818 | 0.952 | 0.858 | 1.014 | 0.953 |
| AR-SV3 | 0.906 | 0.914 | 0.898 | 0.825 | **0.951** | 0.871 | 1.017 | 0.963 |
| AR-SV4 | 0.905 | 0.919 | 0.907 | 0.828 | 0.955 | 0.886 | 1.013 | 0.974 |
| AR-SV5 | 0.907 | 0.917 | 0.892 | 0.820 | 0.958 | 0.855 | 1.013 | 0.973 |
| AR-SV-RW | **0.895** | **0.898** | **0.883** | 0.819 | 0.966 | **0.852** | **0.983** | **0.934** |
| LSTAR | 1.004 | 0.998 | 0.997 | 1.007 | 1.005 | 0.999 | 0.994 | 1.042 |
| Linear Pool | 0.906 | 0.913 | 0.894 | 0.827 | 0.958 | 0.870 | 1.000 | 0.965 |

**Table 16:** Relative CRPS ($rCRPS$) **twelve months ahead**. The first line shows the average $CRPS$ for $AR_{rec}$ over the whole sample. All other rows show $rCRPS$; relative to the benchmark $AR_{rec}$ model. Values smaller than one signal an improvement over the benchmark model. Line *Linear Pool* presents the results of a linear pool with equal weights for models AR-SV1, AR-SV2, AR-SV3, AR-SV4, AR-SV5, AR-SV-RW, LSTAR. Bold entries show the best model for a country (excluding the pooled model).

The results presented in Tables 5-16 are similar to other studies testing the forecasting ability of different model classes for IPI. With respect to a general forecasting context, early studies by de Gooijer and Kumar (1992) and Ramsey (1996) find that non-linear models do not provide any improvement over linear models with respect to point forecasts.[15] At the same time, Pesaran and Potter (1997) already show that non-linear models may be helpful in describing the uncertainty of forecasts more efficiently.

More specifically, with respect to forecasting IPI, Siliverstovs and van Dijk (2003) confirm these results. They find that linear autoregressive models deliver high quality point forecasts, but they also show that non-linear models help improving density forecasts for the G7 countries. Marcellino (2002) finds that linear models deliver the best forecasts performance for IPI in the EMU. However, he does not investigate the uncertainty surrounding the forecasts and focuses on point forecasts only. In line with those studies, I find that autoregressive models with a constant variance perform very well in terms of point forecasts. Still, Tables 5-8 show that it is possible to improve upon point forecasts of $AR$ models by applying AR-SV and/or $LSTAR$ models for certain countries and forecast horizons.

Results change when analyzing density forecasts. Tables 9-16 clearly show that autoregressive models with stochastic volatility improve forecasts with respect to $LogS$ and $CRPS$ for most countries and forecast horizons. At the same time the $LSTAR$ model only improves forecasts for certain countries and horizons, especially for Italy and the Netherlands. However, it should be pointed out that non-linear models like the $LSTAR$ might need a more careful modelling strategy to deliver high quality forecasts as pointed out by Teräsvirta et al. (2005). However, given the large number of models estimated here a more sensible modelling procedure for the $LSTAR$ model was not feasible.

# 5 Conclusion

I investigated the potential of nine univariate time series models estimated via Bayesian techniques for forecasting industrial production growth in eight EMU countries. The forecasting performance was evaluated for point forecasts and density forecasts. Furthermore, I also analyzed the forecasting performance of a linear prediction pool with equal weights for all autoregressive models with stochastic volatility and the logistic smooth transition

[15] Teräsvirta et al. (2005), on the other hand, compare point forecasts for the $LSTAR$ and $AR$ model (and neural networks) for the G7 countries using seven monthly time series (including industrial production) and find that carefully construced $LSTAR$ models improve forecasting peformance over linear models in some cases. However, the results do not indicate a dominant model in general.

autoregressive model.

It has been argued in the literature – see, for example, Marcellino (2002) and Siliverstovs and van Dijk (2003) – that IPI forecasts by linear models cannot be improved upon by more complex univariate models, at least with respect to point forecasts. In fact, Siliverstovs and van Dijk (2003) argue that non-linear models only outperform linear models with respect to density forecasts for industrial production growth.

Indeed, I show that the autoregressive models with constant variance perform very well with respect to point forecasts. At the same time I find that the model can be improved upon – at least at shorter horizons – by autoregressive models with stochastic volatility for some countries. $LSTAR$ models improve point forecasts for certain countries and forecast horizons as well. Finally, regarding linear prediction pools results were rather disappointing. They only helped improving density forecast for longer horizons with respect to $LogS$. On the other hand, they did not improve point forecasts or density forecasts with respect to $CRPS$.

In this paper I focused on autoregressive models with stochastic volatility and the $LSTAR$ model for forecasting. Of course there exist many more univariate forecasting models which could be used for forecasting industrial production growth, for example the unobserved components model with stochastic volatility of Stock and Watson (2007). Bos and Koopman (2010) report positive results for such a model for US industrial production growth. Furthermore, the lag length was fixed for all countries and models in this study. Alternatives to using a fixed lag length are numerous. In fact, Lopes and Salazar (2006) extend the basic $LSTAR$ algorithm with a reversible jump MCMC procedure to account for uncertainty of the lag length as well. However, computation time for eight countries and 200 forecasts is already substantial. The implementation of additional models and the extension of the applied models is therefore left for future work.

Finally, all algorithms for the models described in Section 3 were written **R** and **C++** and integrated via the **Rcpp** package (Eddelbuettel and François, 2011; Eddelbuettel and Balamuta, 2017). The algorithm by Kastner and Frühwirth-Schnatter (2014) was taken from package **stochvol** (Kastner, 2016). The code is available at `https://github.com/alexhaider/Bayes_VARs`.[16]

[16]The functions are provided as-is with very little documentation so far. They are not available as a **R** package.

# References


Alessandri, Piergiorgio and Haroon Mumtaz (2017). "Financial conditions and density forecasts for US output and inflation". In: *Review of Economic Dynamics* 24, pp. 66–78.

Amisano, Gianni and Raffaella Giacomini (2007). "Comparing Density Forecasts via Weighted Likelihood Ratio Tests". In: *Journal of Business & Economic Statistics* 25.2, pp. 177–190.

Avdoulas, Christos and Stelios Bekiros (2018). "Nonlinear Forecasting of Euro Area Industrial Production Using Evolutionary Approaches". In: *Computational Economics* 52.2, pp. 521–530.

Bos, Charles S. and Siem Jan Koopman (2010). *Models with Time-varying Mean and Variance: A Robust Analysis of U.S. Industrial Production*. Tinbergen Institute Discussion Papers 10-017/4. Tinbergen Institute.

Bulligan, Guido, Roberto Golinelli, and Giuseppe Parigi (2010). "Forecasting monthly industrial production in real-time: from single equations to factor-based models". In: *Empirical Economics* 39, pp. 303–336.

Chen, Cathy W. S. and Jack C. Lee (1995). "Bayesian Inference of Threshold Autoregressive Models". In: *Journal of Time Series Analysis* 16.5, pp. 483–492.

Clark, Todd E. and Michael W. McCracken (2001). "Tests of equal forecast accuracy and encompassing for nested models". In: *Journal of Econometrics* 105.1. Forecasting and empirical methods in finance and macroeconomics, pp. 85–110.

Clark, Todd E. and Francesco Ravazzolo (2015). "Macroeconomic Forecasting Performance under Alternative Specifications of Time-Varying Volatility". In: *Journal of Applied Econometrics* 30.4, pp. 551–575.

Clark, Todd E. and Kenneth D. West (2007). "Approximately normal tests for equal predictive accuracy in nested models". In: *Journal of Econometrics* 138.1. 50th Anniversary Econometric Institute, pp. 291–311.

Dacco, Robert and Steve Satchell (1999). "Why do regime-switching models forecast so badly?" In: *Journal of Forecasting* 18.1, pp. 1–16.

de Gooijer, Jan G. and Kuldeep Kumar (1992). "Some recent developments in non-linear time series modelling, testing, and forecasting". In: *International Journal of Forecasting* 8.2, pp. 135–156.

De Santis, Roberta and Tatiana Cesaroni (2016). "Current Account 'Core–Periphery Dualism' in the EMU". In: *The World Economy* 39.10, pp. 1514–1538.

Diebold, Francis X. (2015). "Comparing Predictive Accuracy, Twenty Years Later: A Personal Perspective on the Use and Abuse of Diebold–Mariano Tests". In: *Journal of Business & Economic Statistics* 33.1, pp. 1–1.

Eddelbuettel, Dirk and James Joseph Balamuta (2017). "Extending R with C++: A Brief Introduction to Rcpp". In: *PeerJ Preprints* 5, e3188v1.

Eddelbuettel, Dirk and Romain François (2011). "Rcpp: Seamless R and C++ Integration". In: *Journal of Statistical Software* 40.8, pp. 1–18.

Geweke, John and Gianni Amisano (2011). "Optimal prediction pools". In: *Journal of Econometrics* 164.1. Annals Issue on Forecasting, pp. 130–141.

Gneiting, Tilmann and Adrian E Raftery (2007). "Strictly Proper Scoring Rules, Prediction, and Estimation". In: *Journal of the American Statistical Association* 102.477, pp. 359–378.

Golinelli, Roberto and Giuseppe Parigi (2007). "The use of monthly indicators to forecast quarterly GDP in the short run: an application to the G7 countries". In: *Journal of Forecasting* 26.2, pp. 77–94.

Harvey, David, Stephen Leybourne, and Paul Newbold (1997). "Testing the equality of prediction mean squared errors". In: *International Journal of Forecasting* 13.2, pp. 281–291.

Hubrich, Kirstin and Timo Teräsvirta (2013). "VAR Models in Macroeconomics – New Developments and Applications: Essays in Honor of Christopher A. Sims". In: ed. by Thomas B. Fomby, Lutz Kilian, and Anthony Murphy. Vol. 32. Advances in Econometrics. Emerald. Chap. Thresholds and Smooth Transitions in Vector Autoregressive Models, pp. 273–326.

Jacquier, Eric, Nicholas G Polson, and Peter Rossi (1994). "Bayesian Analysis of Stochastic Volatility Models". In: *Journal of Business & Economic Statistics* 12.4, pp. 371–89.

Kastner, Gregor (2016). "Dealing with Stochastic Volatility in Time Series Using the R Package stochvol". In: *Journal of Statistical Software* 69.5, pp. 1–30.

Kastner, Gregor and Sylvia Frühwirth-Schnatter (2014). "Ancillarity-Sufficiency Interweaving Strategy (ASIS) for Boosting MCMC Estimation of Stochastic Volatility Models". In: *Computational Statistics & Data Analysis* 76, pp. 408–423.

Kim, Sangjoon, Neil Shephard, and Siddhartha Chib (1998). "Stochastic Volatility: Likelihood Inference and Comparison with ARCH Models". In: *The Review of Economic Studies* 65.3, pp. 361–393.

Krüger, F., S. Lerch, T.L. Thorarinsdottir, and T. Gneiting (2019). *Predictive inference based on Markov chain Monte Carlo output,* research rep. Heidelberg Institute for Theoretical Studies,

Lopes, Hedibert F. and Esther Salazar (2006). "Bayesian Model Uncertainty In Smooth Transition Autoregressions". In: *Journal of Time Series Analysis* 27.1, pp. 99–117.

Marcellino, Massimiliano (2002). *Instability and non-linearity in the EMU*. Working Papers 211. IGIER (Innocenzo Gasparini Institute for Economic Research), Bocconi University.

— (2004). "Forecasting EMU macroeconomic variables". In: *International Journal of Forecasting* 20.2. Forecasting Economic and Financial Time Series Using Nonlinear Methods, pp. 359–372.

Newey, Whitney K. and Kenneth D. West (1994). "Automatic Lag Selection in Covariance Matrix Estimation". In: *The Review of Economic Studies* 61.4, pp. 631–653.

Pesaran, M.Hashem and Simon M. Potter (1997). "A floor and ceiling model of US output". In: *Journal of Economic Dynamics and Control* 21.4, pp. 661–695.

Ramsey, James B. (1996). "If Nonlinear Models Cannot Forecast, What Use Are They?" In: *Studies in Nonlinear Dynamics & Econometrics* 1.2, pp. 1–24.

Semmler, Willi and Alexander Haider (2018). "Cooperative Monetary and Fiscal Policies in the Euro Area". In: *Southern Economic Journal* 85.1, pp. 217–234.

Siliverstovs, B. and D.J.C. van Dijk (2003). *Forecasting industrial production with linear, nonlinear, and structural change models*. Econometric Institute Research Papers EI 2003-16. Erasmus University Rotterdam, Erasmus School of Economics (ESE), Econometric Institute.

Silva, Emmanuel Sirimal, Hossein Hassani, and Saeed Heravi (2018). "Modeling European industrial production with multivariate singular spectrum analysis: A cross-industry analysis". In: *Journal of Forecasting* 37.3, pp. 371–384.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman & Hall.

Stock, James H. and Mark W. Watson (1996). "Evidence on Structural Instability in Macroeconomic Time Series Relations". In: *Journal of Business & Economic Statistics* 14.1, pp. 11–30.

— (2007). "Why has U.S. Inflation Become Harder to Forecast?" In: *Journal of Money, Credit and Banking* 39, pp. 3–33.

Teräsvirta, Timo (1994). "Specification, Estimation, and Evaluation of Smooth Transition Autoregressive Models". In: *Journal of the American Statistical Association* 89.425, pp. 208–218.

Teräsvirta, Timo (2006). "Forecasting economic variables with nonlinear models". In: *Handbook of Economic Forecasting*. Ed. by G. Elliott, C.W.J. Granger, and A. Timmermann. Vol. 1. Elsevier. Chap. 8, pp. 413–457.

Teräsvirta, Timo, Dick van Dijk, and Marcelo C. Medeiros (2005). "Linear models, smooth transition autoregressions, and neural networks for forecasting macroeconomic time series: A re-examination". In: *International Journal of Forecasting* 21.4. Nonlinearities, Business Cycles and Forecasting, pp. 755–774.

Thury, Gerhard and Stephen F. Witt (1998). "Forecasting industrial production using structural time series models". In: *Omega* 26.6, pp. 751–767.

Yu, Yaming and Xiao-Li Meng (2011). "To Center or Not to Center: That Is Not the Question—An Ancillarity–Sufficiency Interweaving Strategy (ASIS) for Boosting MCMC Efficiency". In: *Journal of Computational and Graphical Statistics* 20.3, pp. 531–570.