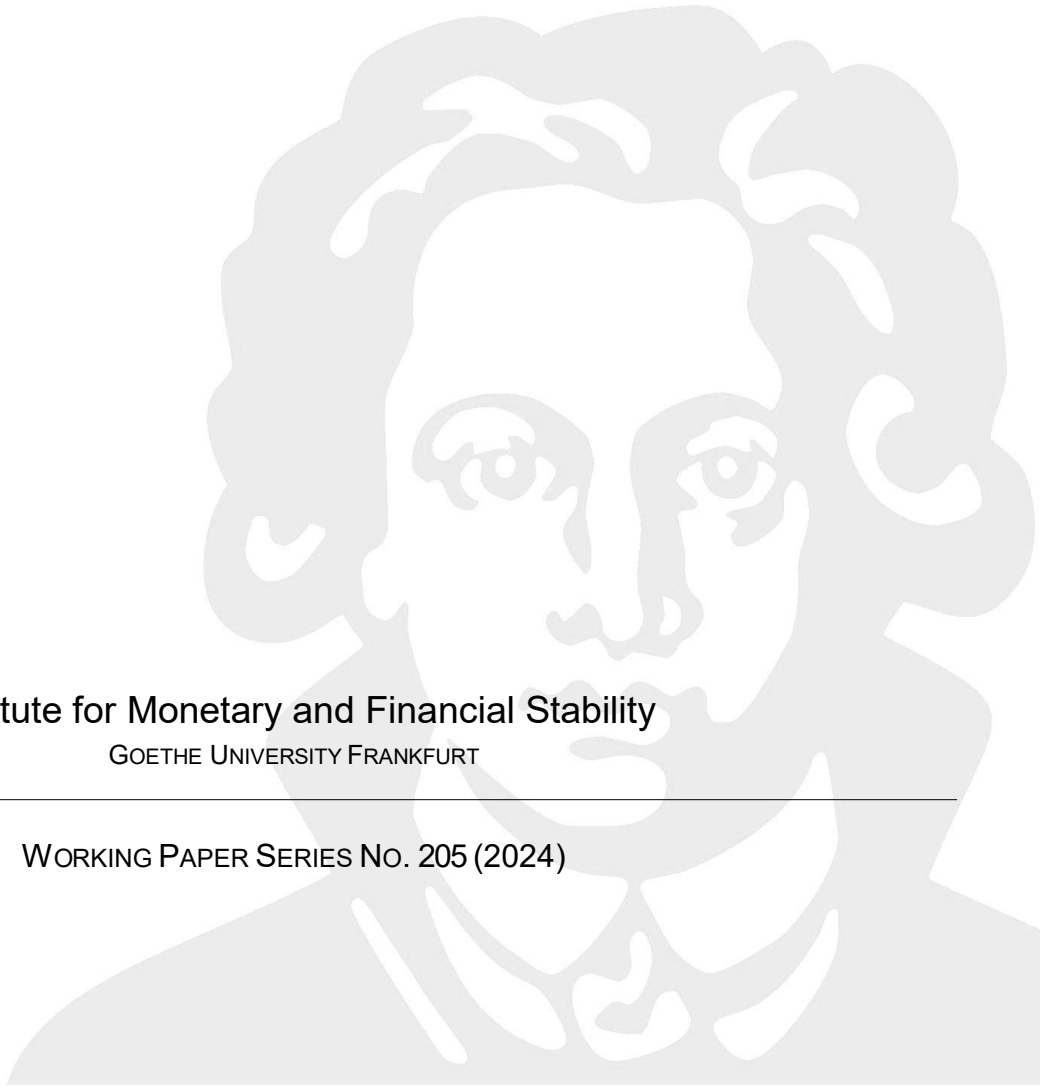
ALINA TÄNZER

Multivariate Macroeconomic Forecasting:

From DSGE and BVAR to Artificial Neural Networks



Institute for Monetary and Financial Stability
GOETHE UNIVERSITY FRANKFURT

WORKING PAPER SERIES NO. 205 (2024)

Institute for Monetary and Financial Stability
Goethe University Frankfurt
House of Finance
Theodor-W.-Adorno-Platz 3
D-60629 Frankfurt am Main
www.imfs-frankfurt.de | info@imfs-frankfurt.de

Multivariate Macroeconomic Forecasting: From DSGE and BVAR to Artificial Neural Networks

Alina Tänzer[a,]

[a]*Institute for Monetary and Financial Stability, Goethe University Frankfurt, Theodor-W.-Adorno Platz 3, 60323 Frankfurt, Germany*

Abstract

This paper contributes a multivariate forecasting comparison between structural models and Machine-Learning-based tools. Specifically, a fully connected feed forward nonlinear autoregressive neural network (ANN) is contrasted to a well established dynamic stochastic general equilibrium (DSGE) model, a Bayesian vector autoregression (BVAR) using optimized priors as well as Greenbook and SPF forecasts. Model estimation and forecasting is based on an expanding window scheme using quarterly U.S. real-time data (1964Q2:2020Q3) for 8 macroeconomic time series (GDP, inflation, federal funds rate, spread, consumption, investment, wage, hours worked), allowing for up to 8 quarter ahead forecasts. The results show that the BVAR improves forecasts compared to the DSGE model, however there is evidence for an overall improvement of predictions when relying on ANN, or including them in a weighted average. Especially, ANN-based inflation forecasts improve other predictions by up to 50%. These results indicate that nonlinear data-driven ANNs are a useful method when it comes to macroeconomic forecasting.

Keywords: Artificial Intelligence; Machine Learning; Neural Networks; Forecast Comparison/ Competition; Macroeconomic Forecasting; Crises Forecasting; Inflation Forecasting; Interest Rate Forecasting; Production, Saving, Consumption and Investment Forecasting; (JEL: C45, C53, E47, E27)

1. Introduction

Predicting macroeconomic variables during business cycles is one of the key challenges, economists face. Especially against the background of the Great Moderation's decay with the financial crisis 2007, and the return of increased volatility, this issue gained even more attention. In the last decades, forecasting in economics was mainly conducted with either structural models - such as Dynamic Stochastic General Equilibrium (DSGE) models -, which rely on the implementation of theoretical knowledge,

*Corresponding author *Email address*: alina.taenzer@web.de

or with purely data-driven methods, such as Vector Autoregression (VAR). While conventional methods constantly improved over the last fifty years, their forecasting abilities are still criticized (Edge and Gürkaynak, 2010). Contrary, artificial neural networks (ANNs) are flexible and powerful systems, that demonstrated particularly impressive progress in recent years, but are by now rarely used in the field of macroeconomics. With this paper, new evidence on multivariate macroeconomic forecasting is provided, challenging standard representative DSGE and Bayesian VAR models by ANNs. The results indicate an overall gain in forecast precision using neural networks, which underpins their value for macroeconomic forecasting.

In principle, a macroeconomist can choose from a large pool of econometric models for her forecast. They differ with respect to the number of treated variables (univers us multivariate), the assumption about the structure of the underlying interrelation (linear versus nonlinear), as well as the setup which either relies on economic theory or simply exploits correlations in the data. However, there are several undesirable properties that conventional approaches bring along, which range from high sensitivity regarding model specifications to high data requirements and time-consuming estimation procedures. In this regard, neural networks are beneficial, as they do not require a choice regarding the underlying structure to be linear or nonlinear, as the universal function approximation property ensures that any underlying interrelation can be approximated to an arbitrary degree (Hornik et al., 1989). Further, except for the selection of input variables, the model does not rely on economic theory and hence no parametric cross-requirements can hinder the forecast from being precise. The number of included variables is virtually unlimited as is the data to be processed. Further, no a-priori beliefs are required and thus also periods of unknown economic behavior are likely to be predictable. Building on these advantages, conventional approaches are challenged by ANNs to thereby exploit their forecasting abilities, which - to the best of the authors' knowledge - has not been done before.

Precisely, this paper contributes a multivariate inspection of ANNs, opposed to a post-crisis structural DSGE model by Del Negro et al. (2015) and a Bayesian vector autoregression (BVAR) using priors as Giannone (2016). Since they challenge their model's forecasting performance in various setups and improve upon predictions by other DSGE models (Binder et al., 2021), Del Negro et al. (2015) constitutes a well established theoretical model which is suitable for this forecasting comparison. Further, employing a BVAR with priors as in Giannone et al. (2015) can improve upon regular (B)VARs by optimally setting hyperparameters. On the other side, a parsimonious and fundamental ANN is generated to limit the parameters to be estimated and to keep the complexity of the network rather small. Hence, this paper contributes a sound comparison of precursor DSGE and BVAR models versus a benchmark ANN model. Furthermore, official forecasts by the Greenbook and SPF predictions are added. Moreover, the forecasting performance during crisis times is investigated, where a new Machine Learning (ML)-based clustering (k-means) approach is provided to identify the respective subsamples. Model estimation and forecasting is based on quarterly U.S. real-time

data from 1964Q2 to 2017Q4. Incorporating 8 macroeconomic time series (GDP, inflation, federal funds rate, spread, consumption, investment, wage, hours worked), a forecasting comparison based on expanding window estimations is conducted.

The results show that using ANNs is advantageous, since they prove to be a robust forecasting tool for a variety of variables. In general, the ANNs' informativeness varies over time and increases the more recent the data. The crisis clustering reveals similarities to NBER recessions and provides further evidence for the robustness of ANN-based forecasts with respect to disruptive times. Particularly ANN-based inflation forecasts are precise and can compete with and improve upon official forecasts. The BVAR is also shown to be a good forecasting tool which outperforms the DSGE, in addition the weighted average of all models constitutes a robust method. These results suggest that ANNs can be a very useful addition to the time series forecasting toolbox. Due to its rather simple setup, the findings can be interpreted as a lower bound of ANN-based forecasting power.

Assigned to the field of *Supervised Learning*, one of the three main ML branches, the employed ANNs can serve well in a variety of applications. Besides pattern recognition to identify objects or signals in speech, vision and control systems, time-series prediction is a core power. However, these ML-methods are not frequently used to solve macroeconomic problems.

This paper relates to the small branch of literature, dealing with ML implementations in a macroeconomic setup. To name some of these rare applications, Zhang et al. (1998) and Kaastra and Boyd (1996) are primal adaptors of ANNs for economic time-series forecasting. Also Swanson and White (1997) conduct a comparison of ANN to several (non)-adaptive and (non)-linear models and provide evidence for their forecasting superiority. Nowadays, due to increased data availability and improved technical facilities, machine-learning based methods can exploit their full potential. While there exists a bunch of forecasting projects in the finance area (see e.g. Fadlalla and Lin (2001) and Dutta et al. (2006)), there is still less research conducted on ML-based macroeconomic forecasting (especially in comparison to conventional methods). Smalter Hall and Cook (2017) for example compare the forecast performance of several deep neural networks to that of the professional forecasters' survey (SPF). These *deep* networks are characterized by a structure with many hidden layers and neurons (see Section 2.1) which is equivalent to a large number of parameters to be estimated. The authors find their ANNs to be superior with respect to short-term unemployment predictions and one of their networks improving at all forecast horizons. Another research project by Verstyuk (2020) uses networks with memory of various sizes to predict US Data on five key macroeconomic variables. The authors contrast their multivariate predictions to VARs and provide evidence for them to generate better forecasts. Showing impulse response functions, they also enhance the interpretability of the networks and provide evidence for them to be able to discover several macroeconomic regimes. Marcellino (2004) provide another rich comparative exercise, which predicts 15 European data series using ANNs and other linear and nonlinear methods. His findings are heterogenous with

respect to variables and models, complex models work well for some variables and worse for others. Furthermore, multivariate inflation forecasting is done by Medeiros et al. (2021), who employ various ANNs as well as another ML-technique called random forests[1]. Contrasting their results to a BVAR, the authors provide evidence for random forests to produce the best predictions.

These projects have in common that results outlay benefits from using novel methods for macroeconomic predictions. However, the variety of ML-methods in general and specifically network types is large and complemented by unlimited possibilities of network sizes and setups. In this regard, a fairly small and simple network is employed which can be seen as a parsimonious and fundamental benchmark and enhances comparability to conventional methods. On this basis, a broad ANN-based forecasting project is contributed which further improves upon existing research due to the multivariate macroeconomic setup, its high number of forecasted windows and the crisis examination.

Furthermore, the paper implements a comparative aspect bringing together ANN-based results and predictions by DSGE and BVAR. Hence, it also relates to the branch of literature dealing with theoretical model-based and conventional empirical forecasting. DSGE models constitute the *state of the art* theory-based models, emphasizing intertemporal decision making and the role of expectations. This class of models is very popular and employed in many institutions as - building on sound theoretical foundations - it delivers an internally consistent interpretation of the current state and future trajectories of the economy and allows for well-grounded analyses of policy scenarios (Del Negro and Schorfheide, 2013). In their seminal work, Smets and Wouters (2007) apply Bayesian estimation to DSGE models and prove good forecasting performance[2]. In general, Bayesian inference produces posterior predictive distributions which reflect uncertainty regarding parameters, state variables and the realization of future shocks conditional on the information available at that time. On the other hand, VARs are widely used for macroeconomic forecasting. As to Karlsson (2013), their popularity stems from the relative simplicity, flexibility, ability to fit the data and from their forecasting accuracy. While the rich parametrization of VAR models brings with it high data requirements or the risk of overfitting the data, Bayesian VARs offer a theoretically grounded way to impose judgmental information and a-priori beliefs in the model. Thereby, the number of parameters shrinks towards a stylized representation of data, reducing parameter uncertainty and thus improving forecasts.

Many economists compared the forecasting performance of DSGE models with (B)VARs or professional forecasts (see for example Del Negro and Schorfheide (2013) who use an adapted Smets and Wouters (2007) model). In a subsequent study, Del Negro et al.

[1]Random Forests are an ensemble learning method, suitable for regression and forecasting, that operates by constructing a multitude of decision trees linking training inputs to outputs.

[2]There are several papers providing reviews on this topic such as An and Schorfheide (2007) and Del Negro and Schorfheide (2011).

(2015) improve upon their previous work, adding detailed financial frictions and thus improve the model's performance during the financial crisis starting 2008. Multiple other projects provide forecasting comparisons, finding differing results depending on the underlying model, the data used for estimation and the forecasting periods[3]. With this project, a comparison of these conventional models and ANNs is provided in order to counteract macroeconomists' skepticism against ML-methods and to broaden their perception with respect to the great potential of such novel tools.

This paper is structured as follows: First, the models used for estimation and forecasting are explained. Second, the data and forecasting strategy are presented. Then, several pseudo out-of-sample forecasts are presented and evaluated in different subsamples. The results are discussed by integrating them in the current stance of literature.

2. Forecasting Approaches

2.1. Artificial Neural Network

First of all, the ANN as an increasingly popular tool of artificial intelligence is introduced, which is perfectly suitable for timeseries forecasting. As stated before, the idea of ANNs dates back to the 1940's, however in recent years, they attracted even more attention, which can be explained by the increasing amount of data availability and equivalently the increased processing power of computer technology. Through continuous innovations, training efficiency could be increased and the risk of overfitting reduced, such that model training turns out to be a feasible task. While this led to widespread applications of ANNs in a variety of fields, the area of macroeconomics has so far refrained to a great extent from integrating them into their tool set. In order to change this with respect to macroeconomic forecasting, a challenge is provided which directly links the prediction power of ANNs to conventional methods.

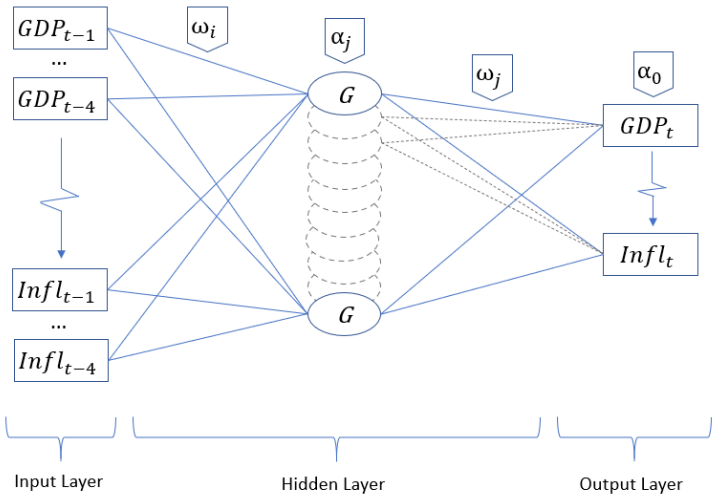
2.1.1. Basic Concept

At first, the basic concept of neural networks is introduced, which can be stylized by directed graphical models, in which information flows from inputs to outputs via a specific structure. This structure consists of nodes, that conduct predefined operations

[3]Del Negro et al. (2007) approximate a DSGE by a VAR (so-called DSGE-VAR) and find the misspecification of large-scale models is not small enough to be ignored. Rubaszek and Skrzypczyński (2008) employ a small size DSGE model and compare its forecasting performance to a trivariate VAR. The superiority of the DSGE in predicting GDP growth turns out to be insignificant. Further, most SPF's forecasts for inflation and the federal funds rate are better than those from DSGE and VAR models. Del Negro and Schorfheide (2013) provide a profound study, focussing on possible improvements of model-based forecasts when including additional information like nowcasts, interest rates and long-run inflation. They find that predictions can compete with Blue-Chip forecasts. Kolasa et al. (2012) provide evidence that GDP forecasts by a DSGE model is better than that of a BVAR and DSGE-VAR in the longer run.

on the data. Instead of specifying model equations and the relevance of data inputs, which economists are used to, when relying on neural networks one rather discusses the model architecture. This refers to the configuration of the network structure, i.e. the number of nodes, the number of layers, the respective interconnections, and the type of operations performed at each node.

Figure 1: Neural Network Scheme



Three distinct sets of nodes, collected in one layer each, form the basic architecture of a neural network. First, there is the Input Layer which merges all model inputs, then there is the Hidden Layer which consists of the set of computational nodes, and last the third set of nodes constitutes the Output Layer. A basic network structure is schematically displayed in Figure 1. It shows that the Input Layer contains all variables of interest and the corresponding lags, which are weighted by parameters - also called weights - collected in the vector ω_i. Feeding this weighted sum of inputs, complemented by a bias α_j, to J nodes (also called neurons), transformations according to the transfer function $G(\cdot)$ are performed in the Hidden Layer. The results are further processed by applying a vector of weights ω_j and assembled to produce the final outputs by adding another bias term α_0. Deep learning models, which are the prevailing neural networks for complex tasks, are scaled-up versions of the introduced structure, containing multiple Hidden Layers with an arbitrary number of parameters. For network estimation, input and target output data is provided to determine weights and biases such that a loss criterion is minimized (e.g. the mean squared error). While economists might recognize the basic network structure as being similar to a Generalized Linear Model (GLM), the training process is conceptually different to conventional econometric approaches employing e.g. maximum likelihood estimation. Contrary, neural networks are trained with non-parametric algorithms which rely on back-propagation and gradient descent for example.

Using a less schematical and more mathematical expression, a neural network is represented by the following equation:

$$y_t = \alpha_0 + \sum_{j=1}^{J} \omega_j G\left(\sum_{i=1}^{I} \omega_i' y_{t-i} + \alpha_j\right) + \varepsilon_t. \tag{1}$$

The vector y_t consists of all input variables entering with $i = 1, ..., I$ lags which are processed on $j = 1, ..., J$ nodes in the Hidden Layer. As before, ω_j, ω_i, α_j and α_0 are the parameters to be estimated.

A key advantage of neural networks is its capability to identify the information (nodes) which is relevant for the prediction throughout the training process. In turn, the researcher may be less selective regarding the data supplied to the model. Furthermore, ANNs have the major advantage of being universal function approximators (Hornik et al., 1989), which indicates that with a neural network, any underlying interrelation can be approximated to an arbitrary degree by linear combinations of the transfer functions $G(\cdot)$, such that $|H(y_t) - \sum_{j=1}^{J} \omega_j G(\cdot)| < \delta$ with J being finite and $\delta \in \mathbb{R}_{>0}$. This characteristic implies that using ANNs, no functional form has to be preselected and the resulting interrelation is purely data-driven. Naturally, neural networks also bring along some caveats, such as being only locally identified. Also the estimated parameters lack economic interpretability and hence they are mainly used for forecasting. In this context, the information *how* predictions are generated is of less interest than their precision. While this disadvantage is generally referred to as the *black-box* characteristic, one can circumvent it to some degree by taking partial derivatives or by calculating partial variable importance.

Consequently, neural networks differ with respect to the specific network architecture in terms of its size, meaning the number of neurons and hidden layers, the interconnection between nodes, the applied transformations, the training procedure employed to estimate weights and biases, and several additional settings which can be individually adjusted. In the following section, some details regarding the specific network employed for the comparison will be presented.

2.1.2. Network Architecture

For this forecasting comparison a fully connected feed forward nonlinear autoregressive neural network is considered. Here, the specification *fully connected* refers to connections which exist between all nodes, while *feed forward* indicates that data is only transferred in one direction, namely from Input to Output Layer (no reverse movement). *Autoregressive* is a standard expression indicating that lags of variables also enter the system and are understood as those. For this exercise, a lag length of four is selected, since this has been shown to be sufficient for quarterly data[4]. Furthermore, this lag structure is chosen to create a fair comparison between BVAR and ANN by allow-

[4]Also (Medeiros et al., 2021) choose a lag length of 4 in their ML-based forecasting exercise.

ing for the same information set. The network structure is selected because it is well suitable for timeseries analysis. Further, with the intention to reduce complexity and deviate as little as possible from conventional methods, the ANN is designed with one Hidden Layer only. As the forecasting performance of networks generally increases with their complexity and depth, the ANN at hand is expected to be a parsimonious and fundamental benchmark.

The neural network's weights and biases are estimated using the Levenberg-Marquardt algorithm with Bayesian regularization (Foresee and Hagan, 1997), which produces networks with excellent generalization capabilities. Typically, a network is trained using backpropagation, which relies on supervised learning, deploying a gradient descent method to reduce a chosen error function (e.g. mean squared error). One caveat of this technique is the potential for overfitting, which leads to a loss of generalization due to fitting of the noise. Bayesian regularization, developed by MacKay (1992) and transferred to neural networks by Foresee and Hagan (1997), is a technique to counteract this issue. While in general, the mean squared errors E_D are reduced, this method aims at also shrinking the employed weights by expanding the objective to $F = \beta E_D + \alpha E_W$, where E_W is the sum of squares of the network weights and biases. The parameters α and β determine the relative importance of function approximation and generalization and are optimized through Bayesian methods. The required Gauss-Newton approximation of the Hessian matrix is achieved applying the Levenberg-Marquardt optimization algorithm. This procedure reduces the potential for arriving at local minima and thereby further increases the generalizability of the network (Ticknor, 2013). One great advantage of Bayesian estimation compared to other regularization techniques, such as early stopping, is the missing needs of splitting the available data into training and validation sets. The resulting network therefore benefits from more training data.

Next, the employed transfer function $G(\cdot)$ is designed in a rectified linear form $ReLU(x) = max(0, x)$, which gained popularity in recent years (Glorot et al., 2011). Compared to sigmoid or similar activation functions, it allows faster and effective training of neural networks using complex datasets. For choosing the width J, which is equivalent to the number of nodes per layer, Hanin and Sellke (2018) provide a range for the minimum width such that in the setting with a single hidden layer and *ReLU* activation functions, the universal approximator property is fulfilled[5]. In this application, the minimum width is between 33 and 40 nodes. However, it still depends on the complexity of the underlying data. An important feature of Bayesian regularization is that it provides a measure of how many network parameters/nodes are effectively being used.

[5]According to Hanin and Sellke (2018),"feed-forward neural nets with a single hidden layer can approximate essentially any function if the hidden layer is allowed to be arbitrarily wide." This result holds for a variety of activation functions, including ReLU. In detail, any continuous function $f : [0,1]^{d_i n} \rightarrow \mathbb{R}^{d_{out}}$ can be approximated arbitrarily close by a ReLU net N with input dimension d_{in}, output dimension d_{out} and minimum layer width ν: $|f(x) - f_N(x)| \leqslant \varepsilon$, with $\varepsilon > 0$. Calculating ν according to $d_{in} + 1 \leqslant \nu_{min}(d_{in}, d_{out}) \leqslant d_{in} + d_{out}$, gives the minimum width for the universal function approximation property to hold.

This *number of effective parameters* is determined in a pre-training step with $j = 1, ..., 40$ nodes, prior to the actual estimation. As the network converges using $j = 20$ nodes (the number of effective parameters remains constant when increasing J), this is chosen as the optimal width for the network at hand.

During each training session, the network's initial weights and biases are randomly chosen, which may lead to different solutions for the same application. Hence, in order to reduce the variance of the individual ANNs, one can estimate multiple networks (in this case $n = 1, .., 30$) and average their outputs. This is especially useful for noisy data and small datasets and a common procedure when working with neural networks, see Cook and Smalter Hall (2017). All multi-step ahead forecasts are computed relying on the iterative method[6].

2.2. Conventional Models

2.2.1. Bayesian Vector Autoregression

One of the key empirical tools in modern macroeconomics is (Bayesian) Vector Autoregression. Their basic concept builds on linearly modeling multiple timeseries, capturing their joint dynamics. VARs are especially suitable for forecasting and policy analysis to shed light on the source of business cycle fluctuations. Building on their frequent and widespread usage, a Bayesian VAR (BVAR) is added to this forecasting competition, which allows to evaluate the ANN against such a well-known benchmark.

VARs are in principal multivariate generalizations of univariate autoregressive models. However, there are many authors who show the superiority of VARs with a prior structure, specifying the shape of the multivariate distribution of the parameters, over unconditional univariate and multivariate models (Canova, 2011). Through the choice of priors, possible overfitting and imprecise predictions caused by a large number of parameters to be estimated can be avoided. A reduced form representation of the VAR is given by:

$$y_t = \Phi_c + \sum_{i=1}^{I} \Phi_i y_{t-i} + \varepsilon_t \tag{2}$$

$$\varepsilon_t \sim N(0, \Sigma), \tag{3}$$

where y_t is the vector of endogenous variables, entering with $i = 1, ..., I$ lags, ε_t is a vector of one-step ahead forecast errors[7], Φ_c is a vector of constants, and Φ_i and Σ are matrices containing the parameters to be estimated. These parameters can be stacked into one vector as $\beta \equiv vec([\Phi_c, \Phi_i]')$. Estimating this vector requires the user to impose additional prior beliefs on the parameters. In this case, the conditional prior distribution

[6]For multistep ahead predictions, this *naive* recursive method is chosen - instead of a Monte-Carlo sampling - in order to keep the computational burden low.

[7]In order to characterize the conditional distribution of y_t given its history, a distributional assumption has to be made for ε_t, such as the normal distribution here.

of the coefficients is given by a conjugate prior of the following form

$$\beta|\Sigma \sim N(b, \Sigma \otimes \Omega\xi), \tag{4}$$

where b and Ω are known functions of hyperparameters, and ξ controls the degree of shrinkage and hence the tightness of the prior information[8]. In case that $\xi \to$ inf, posterior expectations coincide with ordinary least squares (OLS) since the prior becomes uninformative. Contrary, when $\xi \to 0$, the posterior approaches a dogmatic prior. Consequently, ξ is the key parameter for this Bayesian estimation.

The paper follows Giannone et al. (2015)[9] in determining the optimal informativeness of the priors, which is treated as an additional parameter in their approach. While the prior is centered around the Minnesota prior, the degree of shrinkage is chosen optimally, given the marginal likelihood of the data. Furthermore, the GLP prior improves long-run forecasting properties by optimally setting hyperparameters for the combination of the Minnesota prior, the sum of coefficients prior and the dummy-initial observations prior. The authors provide evidence for this procedure to be superior to naive benchmarks and flat-prior BVARs.

One hyperparameter which remains to be determined is the lag length I. As Karlsson (2013) summarizes, increasing the lag length only improves some variables' forecast, choosing it by maximizing the marginal likelihood leads to modest improvements for a majority of the variables. As Carriero et al. (2013) show, the gains of this procedure are rather small and in general, a lag length of four (for quarterly data) is an efficient choice. Hence, the maximum lag number is selected to be $I = 4$.

Based on the estimated model, iterative forecasts are conducted[10]

$$\hat{y}^j_{T+h} = \sum_{i=1}^{h-1} \Phi^j_i \hat{y}_{T+h-i} + \sum_{i=h}^{I} \Phi^j_i y_{T+h-i} + \Phi^j_c, \tag{5}$$

where the first sum indicates that previous forecasts are used for predictions further in the future and the second sum represents lags of the actual data (this term cancels as soon as $h > 4$). Hence, a sample of forecasts $(\hat{y}^j_{T+1}, ..., \hat{y}^j_{T+H})$ is generated from the joint posterior distribution of parameters and ξ, where $j = 10.000$. Calculating the mean (and median) of the predictive densities for each vintage and each forecasting horizon generates the forecasts. The BVAR optimizes the hyperparameters again for every window which categorizes it as an *adaptive* forecasting model.

[8]The conjugate prior implies a likelihood and posterior that come from the same family of distributions and hence makes Bayesian inference feasible also for a large number of parameters to be estimated.

[9]The replication codes are made publicly available by the authors (Giannone et al., 2014).

[10]Alternatively, one could do direct forecasts, with individual parameter sets for every h. For the sake of comparability with the DSGE model, the focus lies on iterative forecasts.

2.2.2. Dynamic Stochastic General Equilibrium Model

As stated before, DSGE models - based on modern macroeconomic theory - are another tool often used to explain and forecast the paths of aggregate time series over the business cycle. Within this class of models, decision rules of the involved agents are derived from assumptions about preferences, technologies, and fiscal and monetary policy regimes, by solving intertemporal optimization problems (see Christiano et al. (2010) for a review). Since also DSGE models are frequently used for forecasting, or at least implicitly rely on precise forecasts when used for modeling e.g. policy implications, one representative model is added to this comparison. The employed DSGE model (Del Negro et al., 2015), which was developed post-(financial)-crisis, is chosen because the authors provide several profound forecast evaluations themselves. Furthermore, Binder et al. (2021) include it in an extensive forecasting comparison, where the model succeeds against multiple other DSGE models.

The DSGE model by Del Negro et al. (2015) builds on work by Smets and Wouters (2007) and extends their work by detailed financial frictions which considerably improved the model's fit against the background of the financial crisis, as well as a time varying inflation target. It is of medium-scale and adds nominal price and wage rigidities, consumption habit formation and investment adjustment costs to the standard neoclassical growth model. In terms of aggregate demand, households maximize their lifetime utility choosing consumption and labor in a non-separable utility function. They are subject to an intertemporal budget constraint and preferences are characterized by habit persistence. Household have the monopoly on labor and stickiness of wages is introduced through a Calvo framework. The supply side is formed by monopolistically competitive firms on the one hand, which produce intermediate goods and sell these to another firm which aggregates them to a final consumption good. For the production, the firm chooses labor and capital inputs. The final good is sold at prices set according to Calvo as consumption or investment good. The financial sector consists of a financial intermediary, capital producers and entrepreneurs, and comprises frictions as designed by Bernanke et al. (1999). Households store their deposits at banks, which lend to entrepreneurs who use the funds together with their own wealth to acquire physical capital. The capital in turn is rented to intermediate goods producers. The entrepreneurs' ability to manage funds effectively is subject to disturbances which gives rise to a state-verification problem. This in turn leads to a spread above the risk-free rate.[11].

In general, before forecasts can be obtained from DSGE models, solving them with numerical methods is the first step to then estimate the models with Bayesian techniques (see e.g. Del Negro and Schorfheide (2011)). More precisely, the set of equations characterizing the models' equilibrium are brought into a system of linear rational ex-

[11]The model codes are extracted from the archive of macroeconomic models (Wieland et al., 2012) and (Wieland et al., 2016) and adapted as required.

pectations difference equations, which can be expressed in their state-space from:

$$s_t = \Phi_1^m(\theta)s_{t-1} + \Phi_\epsilon^m(\theta)\epsilon_t \tag{6}$$

$$y_t = \Psi_1(\theta)s_t + \Psi_0(\theta). \tag{7}$$

Here, s_t represents model inherent state variables, and the coefficient matrices Φ_1^m and Φ_ϵ^m are now functions of model parameters θ. The observational variables y_t are then linked to model variables through the measurement equation (7), which also contains model-dependent coefficient matrices $\Psi_1(\theta)$ and $\Psi_0(\theta)$. While the state variables of the DSGE model are already in a VAR-form, one can also rewrite the state-space representation (6) and (7) as an autoregressive process of the observable vector y_t. Sticking to the notation in Equation (2), we obtain

$$y_t = \Psi_0 + \sum_{i=1}^{I} \Phi_i(\Phi_1^m, \Phi_\epsilon^m, \Psi_1, \Psi_0)y_{t-i} + \Phi_\epsilon(\Phi_\epsilon^m, \Psi_1)\varepsilon_t. \tag{8}$$

Hence, one can see that the estimation of the DSGE model is at its core similar to the BVAR introduced in the previous section, however there are cross-coefficient restrictions which the VAR-parameters have to fulfill (see the dependency of Φ_i and Φ_ϵ on coefficient matrices of the state-space model and thereby on model parameters θ and their interrelations). The specific measurement equations are formulated following Del Negro et al. (2015).

In the next step, the analysis relies on Bayesian techniques to estimate the model parameters θ, which is known as a well-designed and robust method (see An and Schorfheide (2007)). As a prerequisite, the prior distribution of each parameter has to be specified, for which the paper leans on the authors' suggestions. The process of posterior sampling follows a Metropolis-Hastings algorithm and ultimately allows to make predictions following

$$\hat{y}_{T+h}^j = \Psi_1(\theta^j)s_{T+h}^j + \Psi_0(\theta^j), \tag{9}$$

which yields a sample of forecasts $(\hat{y}_{T+1}^j, ..., \hat{y}_{T+H}^j)$, generated from the posterior predictive distribution of the DSGE model with $j = 10.000$ draws. For multiple step-ahead forecasts, first the state variable paths are generated and thereby iterative predictions of the observables. This estimation and forecasting procedure is conducted for each window and the mean (and median) are calculated. All these settings are kept constant for each window over the expanding estimation scheme, it is hence a *non-adaptive* procedure. Please refer to the Appendix for a more detailed description of the algorithm and further settings.

2.3. Model Average

Since this is a common method when conducting forecasts with various models, also a model average (ModAv) is created, which is a simple equally weighted joint forecast:

$$y_{T+h}^{ModAv} = \frac{1}{3}\left(y_{T+h}^{ANN} + y_{T+h}^{BVAR} + y_{T+h}^{DSGE}\right). \tag{10}$$

Here, y_{T+h}^{ANN}, y_{T+h}^{BVAR} and y_{T+h}^{DSGE} are h-step ahead predictions for each variable from the respective models as stated above. This constitutes the fourth forecasting method.

2.4. Official Forecasts

Another popular type of non-structural forecasts are official forecasts conducted for example by the Federal Reserve Board of Governors, called *Greenbook* (GB) forecasts. These projections of multiple key economic indicators are produced prior to each meeting of the Federal Open Market Committee, and made publicly available after 5 years. Furthermore, a consensus forecast is regularly generated based on the Survey of Professional Forecasters (SPF). Here, several economists, companies and agencies produce individual forecasts of given macroeconomic indicators which forms an amount of values which often has the tendency to cluster near the realized value. This characteristic makes SPF forecast but also well-founded GB-forecasts valuable and frequently used benchmarks. Hence, in addition to the model-based predictions, both are included in the comparison.

3. Estimation Principle & Data

3.1. Estimation Principle

Employing the introduced models, an expanding window estimation is conducted. The whole quarterly dataset ranges from 1964Q1 to 2020Q3, and the initial estimation window is defined from 1964Q2 to 1987Q2. Each following window expands by one quarter by adding the subsequent data point to the sample[12]. Forecasts by the models are compared against each other, as well as against official forecasts. Hence, as convenient for forecast comparisons, vintage data is used to ensure that model estimates and competing forecasts only use information that was available at the time of the prediction. Consequently, real-time data is used for model estimation.

3.2. The real-time Data Set

For the estimation of each forecasting model, the paper follows Del Negro et al. (2015) by choosing 8 representative U.S. data series: Real gross domestic product (ROUT-

[12]The first estimation window contains data on 94 quarters, which is well above the recommendation by Fernandez-Villaverde and Rubio-Ramirez (2004), who determine the minimum number of observations for Bayesian estimation to be around 48 quarterly observations.

PUT), the GDP deflator (GDPDEF), nominal personal consumption expenditures (CONS)[13] and fixed private investment (FPI) are generated by the Bureau of Economic Analysis (BEA) and collected in the National Income and Product Accounts of the United States (NIPA). All of the aforementioned variables are real-time data series, providing a realistic environment to evaluate forecasts based exclusively on the information available at that point in time. The average weekly hours of production (AWHNONAG) is included in revised form as the difference between vintages is assessed to be neglectable. The hourly wage of the non-farm business sector (COMPNFB) is also revised, as real-time information would limit the length of the dataset substantially. Further, the effective federal funds rate (EFF) and the Spread (BAA10YM)[14] are included[15]. The civilian employment (CE16OV) and the population level (CNP16OV, in thousands of persons)[16] are required for variable transformation and used in revised from. After employing some transformations on the data (see Appendix), the variables represent the growth rates of output, consumption, investment, the real wage, furthermore the hours worked, inflation, the federal funds rate (FFR) and a spread.

Some considerations go into the selection of data for official forecasts. Greenbook data is extracted by collecting that of the last FOMC meeting of the respective quarter. Since the real time dataset for the estimation is constructed using data as it existed in the middle of each quarter (Croushore and Stark, 2001), this procedure ensures comparability between forecasts, being conservative in a sense that if at all, the official forecasts are too good as they might have somewhat larger information sets. The Greenbook predictions range from 4 to 8 (and more) quarters ahead and, amongst others, provide data on real GDP growth, the GDP deflator and the Federal Funds Rate. Since there is a lag of 5 years of the publication, data is available up to 2015 only. The SPF predictions are available 1 to 4 steps ahead, GDP deflator and real GDP growth are included and the dataset covers the whole analyzed period from 1964 up to 2020. Since the professional forecasters provide a range of predictions, their mean is used as the SPF measure.

4. Forecast Evaluation Strategy

The precision of forecasts crucially depends on the kind of evaluation measure as well as the dimension along which it is compared. In the following, the employed performance measures as well as two evaluation dimensions which refer to (a) different

[13] ROUTPUT, GDPDEF and CONS are extracted from the Real-Time Data Set for Macroeconomists (RTDSM) provided by the Federal Reserve Bank of Philadelphia (see Croushore and Stark (2001) for detailed information).

[14] The spread is defined as Moody's seasoned Baa corporate bond yield relative to the yield on a 10-year treasury bond with constant maturity.

[15] AWHNONAG, COMPNFB, EFF and BAA10YM are extracted from the database FRED offered by the Federal Reserve Bank of St. Louis

[16] FPI, CE16OV and CNP16OV are taken from the Archive for Federal Reserve Economic Data (ALFRED) by the Federal Reserve Bank of St. Louis.

time segments under investigation, and (b) the identification of the forecasted period as crisis time are described.

4.1. Performance Measures

RMSFE. For every estimation, iterative pseudo out-of-sample forecasts with horizons $h = 1, .., 8$ are calculated[17]. To evaluate them, revised data from the most current vintage (2020Q3) is used to determine root mean squared forecasting errors (RMSFE). Given a vector of variables to be forecasted y by method m, the real-time forecast is evaluated with

$$RMSFE(y_m^h) = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{y}_{m,t}^h)^2} \tag{11}$$

where $\hat{y}_{m,t}^h$ denotes predictions for each variable in y, for each step h, and T is the total number of vintages. Following Edge and Gürkaynak (2010), this metric is common for DSGE-based forecasting performance calculations and can easily be applied to BVAR and ANN-based methods. The results are given in two versions: First, the RMSFE is calculated for each variable and each forecasting horizon $h = 1, .., 8$ and then averaged over all 8 variables ($RMSFE_{all}$). Next, the average is taken over a reduced set of variables (GDP, inflation and federal funds rate; $RMSFE_{red}$), which are of special interest to macroeconomists.

Test for Superior Predictive Accuracy. To evaluate the statistical significance of the forecasting results, the test for superior predictive accuracy by Hansen (2005) is conducted. It tests whether the forecasts from distinct models are superior to a set of benchmark models and improves upon the reality check by White (2000) because it is more powerful and less sensitive to poor and irrelevant alternatives. Each model produces a sequence of losses $L_{m,t}$ which are the deviations from actual data in this case. Let the benchmark model be $m = 1$ and the alternative models be $m = 2, ..., M$, $(t = 1, .., T)$, then the relative performance can be defined as

$$X_{m,t} = L_{1,t} - L_{m,t}. \tag{12}$$

The null hypothesis to be tested is that the benchmark model is not inferior to other models, which can be formulated as $H_0 : \lambda_m = E(X_{m,t}) \leqslant 0$ for all $m = 1, .., M$ because $\lambda_k > 0$ means that model m is better than the benchmark. The test statistic $T_n^{SPA} = max_{m=2,..,M}\bar{X}_m/\bar{\omega}_{mm}$ where $\bar{X}_m$ is the m'th element of $\bar{X} = n^{-1}\sum_{t=1}^{n} X_t$ and $\hat{\omega}_{mm}^2$ is a consistent estimator of $\omega_{mm}^2 = lim_{n\to\infty} var(\sqrt{n}\bar{X}_{m,n})$. Thus, the question is whether the

[17]The number of forecasting horizons differs in the literature and varies between one year to up to three years. With the choice of a 2 year horizon the paper follows Edge and Gürkaynak (2010) who also compare conventional models' prediction performance; and multiple machine-learning projects such as Verstyuk (2020) and Paranhos (2021). The analysis thus contains all information for predictions *up to* 8 quarters ahead.

t-statistic T_n^{SPA} is too large for plausibility of $\lambda \leqslant 0$. Assuming strict stationarity of X_t, the test can be implemented using the stationary bootstrap of Politis and Romano (1994).

Hence, the Hansen (2005) test compares a benchmark model to a set of alternatives and answers the null hypothesis whether the benchmark is not inferior to any alternative forecast. This approach is applied to the forecasting comparison by using either a setup with the ANN-based predictions or the ModAv as the benchmark. Within each setup, the set of models is either the reduced one (comprising ANN/ModAv, BVAR and DSGE) or the full one also including the official forecasts. This allows to deduce the overall best predictions, but also the best model-based forecasts. Furthermore, forecast errors for each individual variable and also for the reduced (GDP, inflation, FFR) and full average are taken into account[18].

4.2. Subsamples

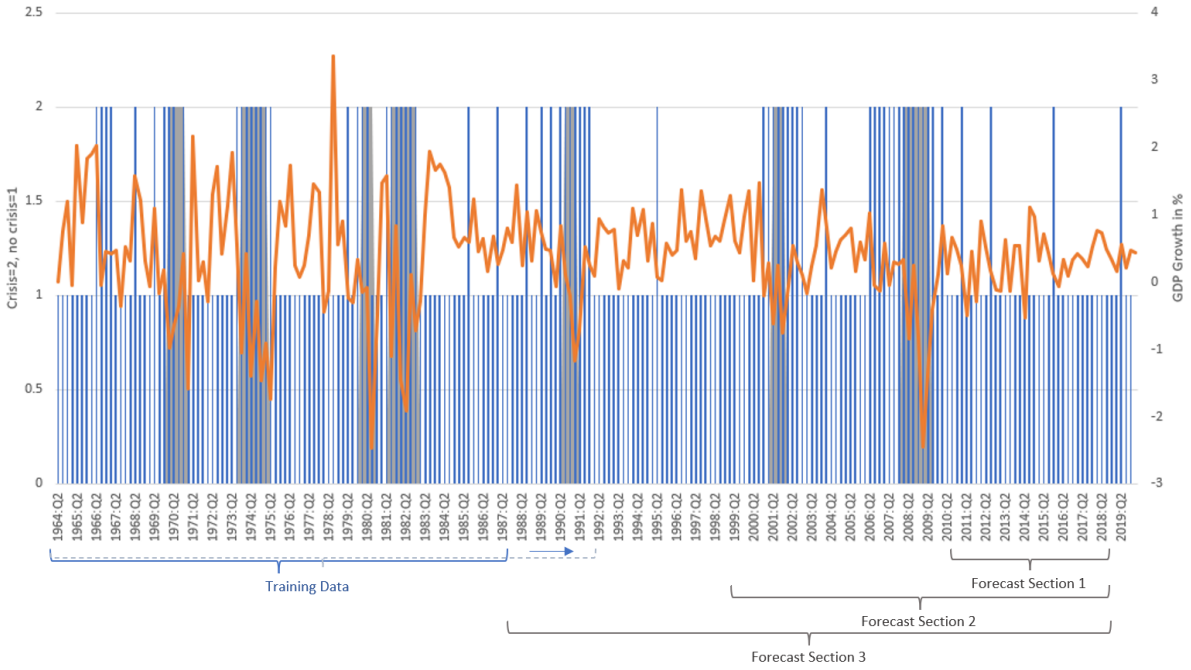
Subsamples According Time Segment. Multiple historic events and resulting fluctuations in the data motivate the contemplation of different forecast subsamples. The first one consists of forecasts during the most recent *post-crisis* time periods from 2010Q2 to 2017Q4 (*Section 1*). This is motivated by the interest in insights regarding contemporary forecasting power. To evaluate the performance over a larger, crisis-influenced but rather stable estimation subsample, it is further looked at forecasts from 1999Q1 to 2017Q4 (*Section 2*)[19]. During the great moderation (starting in the 1980's until 2007), the volatility of business cycle fluctuations declined, which increases the informativeness of the training sample and simplifies the forecasts. The epoch between 1964 and 1987 is embossed by high volatility in monetary policy as well as politics, expressed in manifold crises, which impede the training/estimation process. Hence, this estimation window is analyzed seperately with all 123 forecasts (1987Q3:2017Q4) evaluated as a whole (*Section 3*). The principle of using different forecasting subsamples is visualized at the bottom of Figure 2.

Subsamples According Crises Identification. In addition, in order to allow for a forecast evaluation over disruptive times, a *k-means* clustering of the 8 data series is performed. This method is assigned to the class of unsupervised learning. It follows Lloyd (1982) by assigning n observations to one of k clusters. These are defined by centroids which are randomly initialized and optimized until convergence by an iterative process (see Appendix for details). This happens in an *unsupervised* manner, i.e. no targets are determined, and hence the reason for datapoints to be assigned to a specific cluster may be manifold. Nevertheless, defining clusters $k = 2$, the multivariate dataset at hand seems to identify recessions very precisely. Figure 2 shows the *k-means* result in a column chart in blue (references to the left-hand axis). The clusters are indicated by 1, representing

[18]The codes for conducting this test are publicly available within a toolbox by Sheppard (2009)

[19]A comparable division is conducted by (Medeiros et al., 2021) and justified by a change in inflation volatility around the millennium.

Figure 2: Crises Identification & Estimation and Forecasting Scheme



Note: This figure depicts the mapping of all data series into two *k-means* clusters (left hand axis) against GDP growth (right hand axis). It is obvious that periods clustered in group 2 (indicated by blue bars at value = 2), coincide with periods of small GDP growth. Further, the grey shaded areas depict NBER-dated recessions, which also coincide with most of the *k-means*-dated crises periods. The principle of expanding window estimation and the evaluation according time-based subsamples is indicated at the bottom.

normal times, and 2, referring to recessions. Taking a look at GDP growth in the same figure (right hand side axis), one can clearly see the interrelation between downdrafts in GDP growth and crises identified through clustering. Further, these phases coincide with the grey shaded areas, which represent the NBER-defined recession indicators[20]. Moving along historic events, by this method one can identify the period of monetary tightening (1969/1970), the oil crises 1973-1975, 1979 and 1981/1982 and the golf war 1990/1991. Further, the burst of the dot-com bubble in 2001, the financial crisis (2007-2009) and the corona crisis (2020, not indicated here) can be tagged. It is obvious that the k-means approach allocates more quarters to the *recession* cluster than the NBER data. However, since these are centered around NBER-crises periods, it seems as if the k-means approach already identifies the onset of recessions as well as some separate recessive quarters. The paper takes advantage of this finding and further provides insights about the forecasting performance of each model during these crises.

[20]This data series (USRECQ) is extracted from FRED.

5. Forecasting Results

This section presents the forecasting results based on DSGE, BVAR and ANN models and compares them to the model average (ModAv) as well as official forecasts within several subsamples. First, results are compared along the time dimension using three previously defined subsamples (2010Q2:2017Q4, 1999Q1:2017Q4 and 1987Q3:2017Q4). In the Appendix, forecasts are further compared along periods identified as crises times (H.2).

5.1. Forecasting Results by Time Segment

5.1.1. 2010Q2 to 2017Q4 (Section 1)

The analysis starts by focussing on the moste recent forecasts. Having a look at the epoch after the financial crisis, starting in year 2010, the forecasting power of the ANN can be seen (see Table 1). Averaging the predictions of all 8 variables ($RMSFE_{all}$), the ANN produces the most exact forecasts for $h = 3 : 8$. In the very near term, $h = 1 : 2$, the BVAR succeeds the remaining models. Hansen's test for superior predictive accuracy supports this result, yielding large p-values for the ANN at respective forecast horizons (see Appendix Table 8)[21].

While the ANN outperforms the others concerning all variables for most of the forecasting horizons, it even wins at all horizons concerning the core variable set consisting of GDP, Inflation and the Federal Funds Rate ($RMSFE_{red}$). Even compared to the Greenbook and the SPF forecasts, the ANN's predictions are much closer to the data. The second best model is the BVAR, followed by the Model Average and the DSGE model.

This result is partly driven by an extraordinarily precise ANN-based inflation forecast for all forecast horizons h (Figure 3), which improves the Greenbook and other models' predictions by 20% to 50% and the SPF forecast by up to 5%. GDP growth is best predicted by the DSGE model in this section, however the results are very close to the other models and around 15% better than the Greenbook. There are no FFR forecasts available by the Greenbook for this section, hence the long-term superiority of the ANN is only shown against ModAv, BVAR and DSGE-based predictions (in the short and medium term, the BVAR seems to be slightly better). Figure 11 provides more detailed insights on the remaining variables' forecasts by each model. While ANN-based investment forecasts are improved and now superior from $h = 3$ on, consumption forecasts deteriorate in the long-run. The wage predictions still vary over horizons, whereas forecasts of hours worked seem to converge with the forecasting horizon. Furthermore, the ANN improves its *financial* (spread) forecasts and is now superior from $h = 4 : 8$.

While it is difficult to disentangle the causes for varying superiority in the predictive accuracy, it is clear that after the financial crisis there is less volatility in the data which in general facilitates forecasting. This is underlined by low RMSFEs over all horizons and

[21]The SPA test is either conducted with the ANN or the ModAv as benchmark. This is because the two models' results are very close and by this method one can draw more conclusions about whether the ANN or another model drives the superiority of the ModAv.

Table 1: RMSFE Section 1

Model	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=7$	$h=8$
av. all Vs.								
ANN	0.45	0.50	**0.48**	0.49	**0.47**	**0.49**	**0.51**	**0.51**
ModAv	0.44	0.46	0.48	**0.49**	0.48	0.50	0.53	0.52
BVAR	**0.43**	**0.46**	0.49	0.51	0.51	0.53	0.55	0.56
DSGE	0.53	0.56	0.57	0.58	0.58	0.59	0.62	0.61
av. red. Vs.								
ANN	**0.24**	**0.27**	**0.28**	**0.30**	**0.30**	**0.31**	**0.33**	**0.33**
ModAv	0.25	0.28	0.31	0.33	0.34	0.36	0.38	0.39
BVAR	0.25	0.27	0.30	0.31	0.31	0.33	0.34	0.36
DSGE	0.30	0.35	0.40	0.43	0.46	0.50	0.53	0.55
GB	0.32	0.35	0.37	0.36	0.35	0.35	0.34	0.37
SPF	0.31	0.31	0.30	0.31	NaN	NaN	NaN	NaN

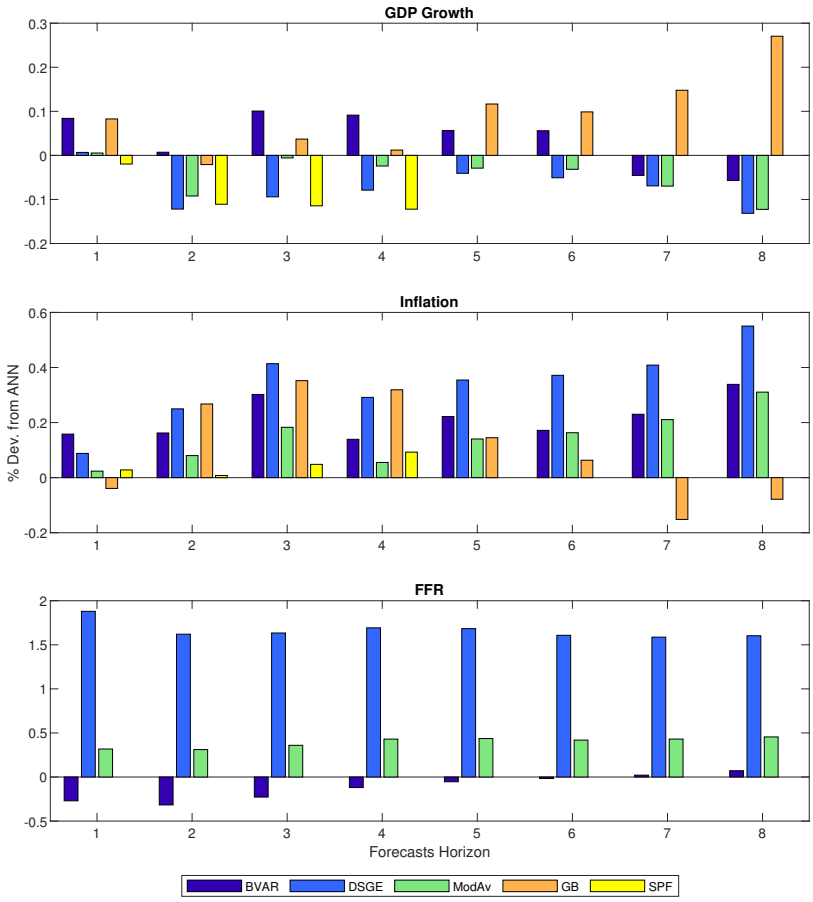
Note: The column *av. all Vs.* shows the RMSFE, of the respective subsample, averaged over all 8 variables. The column *av. red. Vs.* averages over GDP, inflation and the federal funds rate only. $h = 1, ..., 8$ gives the forecasting horizon. As there is no FFR data from the SPF available, it is averaged over GDP and Inflation forecasts only. The maximum forecast horizon of the SPF is 4. The GB forecasts are available up to 2015. Since there are no Greenbook FFR forecasts available for this section, the *av. red. Vs.* now refers to the average over GDP and inflation only, similar to the SPF.

models compared to section 2 and 3. However, GDP growth and inflation paths (Figure 8 and 9) still indicate a substantial degree of volatility after 2010. The improvement (especially in the ANN-based inflation predictions) could therefore also result from an enlarged training sample which now includes information on the financial crisis and the underlying mechanisms, which are incorporated in the post-crisis DSGE model through theoretical assumptions and the choice of priors. The results from section 1 hence constitute a fascinating result for the robustness of ANN-based forecasting performance over many variables and especially for inflation.

5.1.2. 1999Q1 to 2017Q4 (Section 2)

The second forecasting segment starts after the phase of high economic volatility before 1999 and averages over forecasts based on vintages ranging from 1999Q1 to 2017Q4. Table 2 shows the respective results measured again as RMSFEs. While we can see that the performance of the DSGE is now much more competitive, the lowest RMSFEs averaged over all variables are provided by the ModAv which is mainly driven by the BVAR during short-term predictions ($h = 1 : 4$), while the ANN again provides more predictive power during long-term forecasts ($h = 5 : 8$). The results from Hansen (2005) underline the overall superiority of the ModAv (see Table 11) with the respective drivers giving the largest p-values when omitting the averaged forecasts (see Table 10).

Figure 3: Relative RMSFE Section 1



Note: This figure shows RMSFEs of individual variables as percentage deviation from the ANN over all forecast horizons.

Focussing on the reduced variable set, the ANN already drives the ModAv's forecasts for $h = 3$ and even provides the most exact forecasts itself for $h = 4, .., 8$. It is of special interest, that the ANN's superiority also holds true comparing it to the Greenbook forecasts. Superior predictive accuracy of the ANN for the reduced variable set is also found by the test results in Table 10. Including all models and the official forecasts as alternatives, the test results still provide evidence that the ANN produces the most exact forecasts starting with forecast horizon $h = 3$ to $h = 8$. It is thus the preferred specification for medium- and long-term predictions in this section.

The individual variables' predictions in terms of relative RMSFEs (Figure 4) show in detail the source of the ANN's performance. GDP growth forecasts gain in precision compared to all other models and the officials' in the medium-term. Further, in the long run, the ANN is superior. Forecasted GDP growth paths of each model versus actual

Table 2: RMSFE Section 2

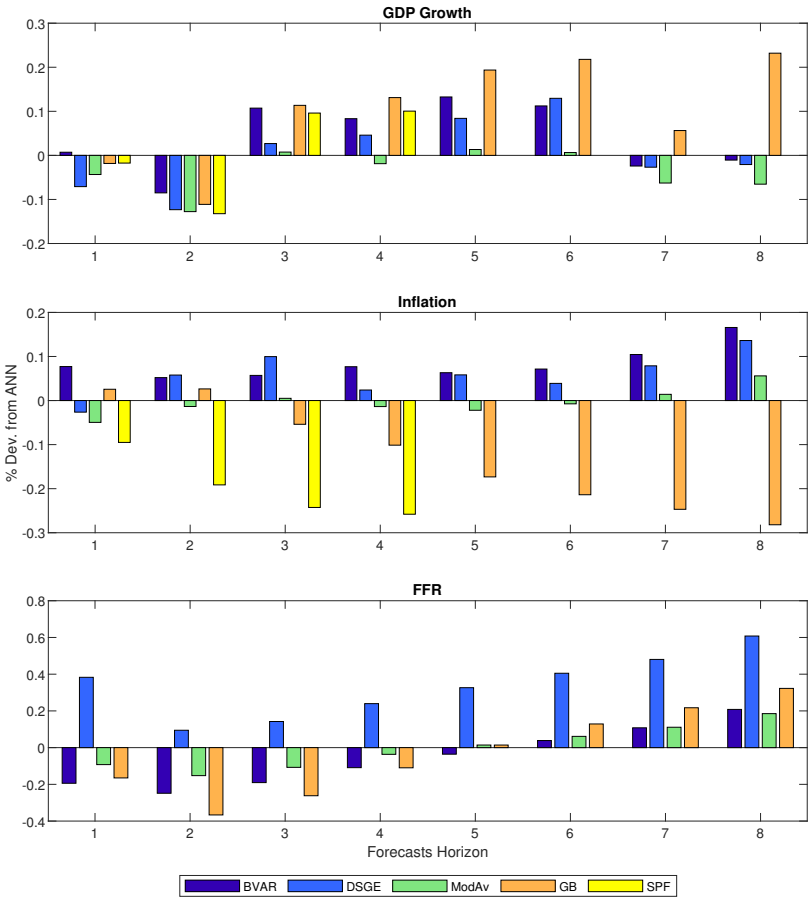
Model	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=7$	$h=8$
av. all Vs.								
ANN	0.62	0.75	0.75	0.78	0.77	0.77	0.82	0.83
ModAv	**0.57**	0.67	**0.68**	**0.71**	**0.74**	**0.75**	**0.79**	**0.81**
BVAR	0.58	**0.65**	0.70	0.75	0.80	0.81	0.83	0.85
DSGE	0.61	0.69	0.74	0.78	0.83	0.85	0.88	0.90
av. red. Vs.								
ANN	0.32	0.43	0.43	0.46	**0.47**	**0.49**	**0.52**	**0.51**
ModAv	**0.31**	**0.38**	**0.42**	**0.48**	0.48	0.50	0.53	0.54
BVAR	0.32	0.38	0.43	0.46	0.50	0.52	0.55	0.57
DSGE	0.33	0.42	0.47	0.51	0.55	0.59	0.61	0.64
GB	0.31	0.36	0.41	0.45	0.49	0.52	0.54	0.58
SPF	0.38	0.40	0.42	0.43	NaN	NaN	NaN	NaN

Note: The column *av all vars* shows the RMSFE, of the respective subsample, averaged over all 8 variables. The column *av red vars* averages over GDP, inflation and the federal funds rate only. $h = 1, ..., 8$ gives the forecasting horizon. The GB forecasts are available up to 2015. As there is no FFR data from the SPF available, it is averaged over GDP and Inflation forecasts only. The maximum forecast horizon of the SPF is 4.

realized data is given in Figure 8 in the Appendix. It is obvious that neither model nor professional forecasts mimic the actual GDP growth path precisely, especially prior to 2010. Opposite to the other models, the ANN prescribes more and stronger fluctuations leading to closer paths to actual (see $h = 4$ and $h = 8$). Additionally, the inflation forecasts of the network extends its lead relative to BVAR and DSGE, and this superiority increases with the forecasting horizon. In addition, the gap towards the officials' inflation forecast can be reduced by 10 pp. It should be mentioned though, that the Greenbook has missing datapoints as for several projections, the long-term forecasts are missing (especially in crisis times) and in addition the sample's last values are from 2015. Nevertheless, showing the GB as a comparison is interesting, however due to these reasons, the results should not be seen as a proof against the forecasting performance of the ANN. Forecast paths for inflation are given in Figure 9 in the Appendix. The Officials' predictions are quite precise before 2002 and after 2010, while prescribing too low paths in the meantime, during which the other models perform well. The subfigure referring to $h = 8$ underlines the lack of data of the SPF and the incomplete dataset by the Greenbook. The improvement of the ANN with respect to the other models in this section is remarkable due to high inflation volatility[22]. The relative RMSFE

[22]Similar superiority of a ML-model, a random forest in their case, during the 2001:2015 period is provided by Medeiros et al. (2021).

Figure 4: Relative RMSFE Section 2



Note: This figure shows RMSFEs of individual variables as percentage deviation from the ANN over all forecast horizons.

of FFR forecasts in Figure 4 show a switch from inferiority in the short-term to distinct superiority in the long-term (paths are given in Figure 10 in the Appendix). It is obvious that the actual FFR path differs substantially from the other variables constituting less volatility especially during the zero lower bound period after the financial crisis. This, in the short-term, seems to be easily detectable by the BVAR and the officials, while the DSGE models has difficulties in determining the FFR level. In the long run we see larger deviations in general. However, the ANN-based forecasts manage to predict some trends very well as during the onset of the financial crisis. Taking a closer look at the remaining variables (Figure 12) consumption and investment show medium-term superiority of the ANN which matches well the results for GDP. RMSFEs for wage is mixed over the forecast horizons whereas the spread and the hours worked are precisely predicted by the ANN in the long run.

5.1.3. 1987Q3 to 2017Q4 (Section 3)

Analyzing the results of the third section, which includes forecasts from all 123 vintages, the RMSFEs indicate that the BVAR provides the most exact forecasts on average (see Table 3), closely followed by the Model Average (ModAv) and the ANN, while the DSGE performs worse.

Table 3: RMSFE Section 3

Model	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=7$	$h=8$
av. all Vs.								
ANN	0.62	0.71	0.72	0.75	0.77	0.79	0.83	0.85
ModAv	0.60	0.65	0.68	0.70	**0.72**	**0.74**	0.78	**0.79**
BVAR	**0.55**	**0.61**	**0.65**	**0.69**	0.73	0.75	**0.78**	0.80
DSGE	0.80	0.83	0.85	0.87	0.88	0.90	0.91	0.91
av. red. Vs.								
ANN	0.33	0.41	0.42	0.44	0.47	0.49	0.53	0.54
ModAv	0.36	0.38	0.41	**0.43**	**0.46**	**0.49**	**0.52**	**0.53**
BVAR	**0.31**	**0.37**	**0.40**	0.43	0.47	0.50	0.53	0.56
DSGE	0.59	0.55	0.56	0.58	0.59	0.62	0.63	0.64
GB	0.30	0.34	0.38	0.42	0.46	0.48	0.50	0.54
SPF	0.37	0.39	0.41	0.42	NaN	NaN	NaN	NaN

Note: The column *av all vars* shows the RMSFE, of the respective subsample, averaged over all 8 variables. It thus provides a performance measure regarding the multivariate forecasting power of each model. The column *av red vars* averages over GDP, inflation and the federal funds rate only. $h = 1, ..., 8$ gives the forecasting horizon which reaches from 1-step ahead up to 8-steps ahead. The GB forecasts are available up to 2015 only. As there is no FFR data from the SPF available, it is averaged over GDP and Inflation forecasts only. The maximum forecast horizon of the SPF is 4.

Going more into detail focussing on all eight variables, BVAR forecast errors are the smallest for horizons $h = 1:4$ and $h = 7$ (increasing with h), the ModAv provides the best average forecast for $h = 5, 6$ and $h = 8$. The ANN itself slightly lacks behind which might be due to the initially relatively small training sample containing large fluctuations, against which the BVAR is robust due to initial parameter restrictions. Contrary, the bad performance of the DSGE can possibly be explained by a different policy regime prior Greenspan, which does not match the parameter assumptions of the model and hinders the estimation. Hansen's test for superior predictive accuracy supports these results, yielding large p-values for the mentioned models at respective forecast horizons (see Table 12). Focussing on GDP, inflation and FFR forecasts (*av. red. Vs.*), these results roughly remain the same, however the driving force behind the superiority of ModAv's long-term predictions is now the ANN forecast. This can be deduced from larger p-

values for the ANN compared to the BVAR (see Table 12). Including official forecasts in this analysis, GB forecasts are on average better than SPF forecasts with slightly smaller RMSFEs than the best models. However, for $h = 5, 7, 8$ the ModAv is found to have *superior predictive accuracy* compared to all models (including official forecasts).

The analysis of individual variables within section 3 is shifted to the Appendix H.1, as is the closer look at crises periods (Appendix H.2).

6. Discussion

There are several dimensions, along which this paper's results compare to other research. There is the time dimension first, which is tackled by using different subsamples when evaluating the results. This also includes the crises exploration. Second, individual variables' predictions can be contrasted against other traditional and machine-learning models (variable and model dimension). The model dimension may then also be considered for multi-variable predictions.

Time Dimension. First, it becomes clear that when conducting a forecasting comparison, it is recommendable to consider diverse prediction subsets and/or to take the characteristics of the respective period into account when evaluating forecast errors. Differences in results between sections prove that the superiority of models can vary significantly over time. In section 3, the statement by Karlsson (2013) can be supported, that BVARs tend to be good multivariate forecasting tools, while the DSGE produces unexpectedly large RMSFEs. This finding is in line with research by Edge and Gürkaynak (2010) who show bad DSGE-based forecasting performance since the onset of the great moderation which reflects the changed nature of macroeconomic fluctuations. The ANN plays a minor role being superior only for the medium- to long-term forecasts of the reduced set of variables, which can possibly be improved by using a deeper network. However, while the performance of ANN and DSGE improve along time (section 3 to 1), the predicting power of the BVAR decreases. Finally, all models provide even better forecasts for section 1, with the ANN becoming more superior. This result is at contrast to findings by Stock and Watson (2007) and Tulip (2009), who say that for inflation (and output), since the beginning of the Great Moderation the forecastable component has decreased which hindered precise predictions. The results at hand provide evidence, that only the one-step ahead prediction of the Greenbook improves from section 1 to 2 (averaged over the reduced variable set), while at all remaining horizons, prediction accuracies are decreased. Hence, the time dimension reveals an improvement of forecasts in general and the ANN in particular over subsamples, pointing to nonlinearities and hidden interrelations that can be captured by the network.

Individual Variable Model Dimension. The Smets and Wouters (2007) DSGE model is contrasted to Greenbook forecasts by Del Negro and Schorfheide (2013). The authors find a short-run disadvantage for the model which decreases such that forecasts (for output and inflation) become competitive for the medium-and long run. With the paper at

hand, this finding can only be confirmed for medium to long-term GDP growth predictions (however, the ANN-based predictions are even more precise). Averaged over the (reduced) set of variables, the gap between Greenbook and DSGE forecasts increases, which is why the result by Rapach et al. (2013) that both predictions converge over the medium and long-term can also not be acknowledged. Furthermore, the BVAR seems to be the best short-term predictor of FFR in section 1 and 2 which opposes the finding by Karlsson (2013) that VARs typically do not have enough structure to generate predictions about anticipated changes in interest rates.

How do this paper's results relate to other machine-learning-based forecasts of individual variables? There are several papers focussing on the construction of machine-learning methods to predict individual variables. One such project is done by (Cook and Smalter Hall, 2017) who use four different neural network architectures to make univariate unemployment forecasts and compare these to SPF predictions. Their test set lasts from 1997 to 2014 and since unemployment relates to GDP, GDP growth predictions from section 2 can roughly be compared to the authors' results. All of the created neural networks by (Cook and Smalter Hall, 2017) improve upon short-term SPF forecasts, while the most advanced model (a so-called encoder-decoder network) beats the SPF at all horizons. Through this paper's results, their findings can be confirmed that (with the exception of the two step ahead prediction), the ANN-based forecasts are around 10% better than those by the SPF for $h = 2, 3$, and the one step ahead prediction is approximately equally precise. Multivariate inflation forecasting is for example tackled by (Paranhos, 2021) who finds that long-short-term-memory (LSTM) networks generate better predictions than benchmarks at long-term (8 quarters) horizon. A very rich comparative analysis is conducted by (Medeiros et al., 2021), including a large set of variables and models (also BVAR and diverse networks). Random Forests turn out to be the best tool to predict inflation, with this superiority being even more pronounced during 2001 to 2015. Since the authors also split their forecasts into different subsamples, their findings can be compared to this paper's first and second section very well and can attest this improvement among the models' inflation predictions over all horizons and especially the long-term precision. Another paper producing FFR forecasts in comparison to multiple other linear and nonlinear models is done by (Hinterlang, 2020), who finds neural networks to be the best prediction tool at all forecasting horizons, while the superiority increases with further timely distance. This finding can partly be confirmed by this analyses, as the ANN-based FFR forecasts only become superior for $h = 4 : 8$.

Multi-variable Model Dimension. The comparison between average BVAR and DSGE performance in section 1 and 2 is still in favor of the BVAR, although the gap between both models' forecasts is reduced over time. This finding is in line with Binder et al. (2021) but opposite to Edge and Gürkaynak (2010) who claim the Smets and Wouters (2007) model to be mostly superior to a (non-GLP) BVAR. This might of course be driven by different modeling and BVAR specifications and should not discriminate one model class per se.

Furthermore, there is research by (Marcellino, 2004) and Verstyuk (2020) which - sim-

ilar to this paper - deals with multivariate machine-learning-based forecasting of a set of macroeconomic variables instead of a single variable. Based on 15 EMU macroeconomic series from 1970 to 1997, Marcellino (2004) finds heterogenous results with respect to a variety of forecasting methods including neural networks. The authors conclude, that complex models work particularly well for some variables (e.g. GDP and the exchange rate), while simple linear models outperform them for other series. The heterogeneity between the models' performance with respect to the 8 data series analyzed in this paper can be affirmed. However, neither of them has the characteristics of a simple linear model. Using five data series (GDP growth, inflation, commodity prices, FFR and bank reserves), Verstyuk (2020) finds evidence for the LSTM network to be on average superior to the benchmark VAR with the test dataset defined as from 2015 to 2019. Similar conclusions can be drawn from the RMSFEs from section 3 which underline superiority of the ANN for $h = 3 : 8$ over all variables and for $h = 1 : 8$ for the reduced set of variables.

7. Conclusion

This paper contributes a macroeconomic forecasting comparison between a structural DSGE model, a data-driven linear BVAR and an Artificial Neural Network within a multivariate framework. A fully connected feed forward nonlinear autoregressive neural network is contrasted to the DSGE model by Del Negro et al. (2015) and a BVAR using optimized priors as in Giannone et al. (2015). Using real-time data for 8 macroeconomic time series (GDP, inflation, federal funds rate, spread, consumption, investment, wage, hours worked), a forecasting comparison based on expanding window estimations is conducted and analyzed during various subsamples. Moreover, a distinct view at crisis performance delivers further insights. The k-means approach to identify these recessive times is another novel method and a contribution.

The results show, that when focusing on post-financial-crisis times, the ANN yields the lowest RMSFE, averaged over all 8 variables, with forecasting horizons 3 to 8. The core variables' forecasts by the ANN even outperform conventional methods over all horizons and official forecasts over nearly all horizons. This is amongst others driven by up to 50% more precise inflation predictions. Comparing forecasts from 1999 to 2017, the ANN is superior for the medium and long-run forecasts. Considering all variables, this drives the ModAv, focussing on core variables (GDP, inflation and FFR), the ANN itself is the optimal predictor. Whereas in the short-term, the BVAR succeeds, medium-term GDP predictions by the ANN are about 10% and inflation forecasts up to 15% better than benchmarks. FFR forecasts can be improved by up to 60% when using the ANN. Averaging over forecasts between 1987 and 2017, the long-term predictions (5 to 8 quarters ahead) by ANN for GDP, inflation and the federal funds rate are more precise than conventional models' and drive the ModAv[23] as best forecasting tool. During

[23] As explained in Section 2.3, an average of all models' predictions is calculated as a fourth forecasting

the same time, BVAR-based short-term predictions are better than DSGE-based ones. Focusing on forecasting performance during recessions, the ANN proves to be a robust tool for long-term predictions during crises while the BVAR should be considered for the near horizons.

Concluding over the mentioned results, this paper provides evidence for DSGE models to produce appropriate short-term predictions after 1999. The BVAR is suited well when producing average forecasts including the disruptive times prior 1999, especially in the short-run. Furthermore, in several cases, a weighted average of all forecasts constitutes a robust alternative. The ANN however, which is presented as the novel forecasting method in this paper, delivers an overall gain. While its superiority varies with the out-of-sample periods to be forecasted, predictions improve the closer the forecasted periods move towards present times.

The literature on macroeconomic forecasting with neural networks has experienced increasing attention in the recent years. As the results show, this is justifiable and desirable. The specifications of neural networks seem to provide an unlimited amount of variations to forecasting projects and depending on the data and question to solve, the optimal specification might be difficult to find. It would thus be desirable to compare more networks in a comparative analysis, e.g. the mentioned LSTM networks. Further, one can expect improved forecasts by using nowcasts and by enlarging the dataset. It would be of interest to investigate, whether these modifications affect the models' performance differently. These extensions are left for future research. Yet, it is clear, that neural networks deliver a substantial advantage compared to conventional methods. As these findings indicate, nonlinear data-driven ANNs are a useful method when it comes to macroeconomic modeling and forecasting and should be added to the macroeconomists' toolkit.

Acknowledgments

The paper represents the author's personal opinion and does not necessarily reflect the views of the Institute for Monetary and Financial Stability or the Goethe University. Any errors are mine. I would like to thank Natascha Hinterlang, Volker Wieland, Alexander Meyer-Gohde and the 41st International Symposium on Forecasting and its participants for fruitful comments and discussions. Conflicts of interest: none.

References

AKAIKE, H. (1970): "Statistical Predictor Identification," *Annals of the Institute of Statistical Mathematics*, 22, 203–217.

AN, S. AND F. SCHORFHEIDE (2007): "Bayesian Analysis of DSGE Models," *Econometric reviews*, 26, 113–172.

tool.

BERG, T. O. (2016): "Multivariate forecasting with BVARs and DSGE models," *Journal of Forecasting*, 35, 718–740.

BERNANKE, B. S., M. GERTLER, AND S. GILCHRIST (1999): "The Financial Accelerator in a Quantitative Business Cycle Framework," *Handbook of Macroeconomics*, 1, 1341–1393.

BINDER, M., M. WIELAND, V.AND WOLTERS, J. TAYLOR, Z. SUN, AND M. FARKAS (2021): "Forecasting the Great Recession in the United States: First Results from a Model Comparison Excercise," *Mimeo*.

BROOKS, S. P. AND A. GELMAN (1998): "General Methods for Monitoring Convergence of Iterative Simulations," *Journal of Computational and Graphical Statistics*, 7, 434–455.

CANOVA, F. (2011): *Methods for Applied Macroeconomic Research*, Princeton University Press.

CARRIERO, A., T. E. CLARK, AND M. MARCELLINO (2013): "Bayesian VARs: Specification Choices and Forecast Accuracy," *Journal of Applied Econometrics*, 28.

CHANG, Y.-Y., F.-Y. SUN, Y.-H. WU, AND S.-D. LIN (2018): "A Memory-Network Based Solution for Multivariate Time-Series Forecasting," *arXiv preprint arXiv:1809.02105*.

CHRISTIANO, L. J., M. TRABANDT, AND K. WALENTIN (2010): "DSGE Models for Monetary Policy Analysis," in *Handbook of Monetary Economics*, Elsevier, vol. 3, 285–367.

COOK, T. AND A. SMALTER HALL (2017): "Macroeconomic Indicator Forecasting with Deep Neural Networks," *The Federal Reserve Bank of Kansas City Research Working Papers*.

CROUSHORE, D. AND T. STARK (2001): "A Real-Time Data Set for Macroeconomists," *Journal of Econometrics*, 105, 111–130.

DEL NEGRO, M., M. P. GIANNONI, AND F. SCHORFHEIDE (2015): "Inflation in the Great Recession and New Keynesian Models," *American Economic Journal: Macroeconomics*, 7, 168–196.

DEL NEGRO, M. AND F. SCHORFHEIDE (2011): "Bayesian Macroeconometrics," in *The Oxford Handbookof Bayesian Econometrics*, ed. by K. G. v. D. H. Geweke, J., Oxford University Press, 293–389.

——— (2013): "DSGE Model-Based Forecasting," in *Handbook of Economic Forecasting*, Elsevier, vol. 2, 57–140.

DEL NEGRO, M., F. SCHORFHEIDE, F. SMETS, AND R. WOUTERS (2007): "On the Fit of New Keynesian Models," *Journal of Business & Economic Statistics*, 25, 123–143.

DUTTA, G., P. JHA, A. K. LAHA, AND N. MOHAN (2006): "Artificial Neural Network Models for Forecasting Stock Price Index in the Bombay Stock Exchange," *Journal of Emerging Market Finance*, 5, 283–295.

EDGE, R. M. AND R. S. GÜRKAYNAK (2010): "How Useful Are Estimated DSGE Model Forecasts for Central Bankers?" *Brookings Papers on Economic Activity*, 2010, 209–244.

FADLALLA, A. AND C.-H. LIN (2001): "An Analysis of the Applications of Neural Networks in Finance," *Neural Networks in Finance*, 11.

FERNANDEZ-VILLAVERDE, J. AND J. F. RUBIO-RAMIREZ (2004): "Comparing Dynamic Equilibrium Models to Data: A Bayesian Approach," *Journal of Econometrics*, 123, 153–187.

FORESEE, F. D. AND M. T. HAGAN (1997): "Gauss-Newton Approximation to Bayesian Learning," in *Proceedings of International Conference on Neural Networks (ICNN'97)*, IEEE, vol. 3, 1930–1935.

GIANNONE, D. (2016): "Exploiting the Monthly Data Flow in Structural Forecasting," *Journal of Monetary Economics*, 15.

GIANNONE, D., M. LENZA, AND G. E. PRIMICERI (2014): "Replication Data for: Prior Selection for Vector Autoregressions," .

——— (2015): "Prior Selection for Vector Autoregressions," *Review of Economics and Statistics*, 97, 436–451.

GLOROT, X., A. BORDES, AND Y. BENGIO (2011): "Deep Sparse Rectifier Neural Networks," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, JMLR Workshop and Conference Proceedings, 315–323.

HAGAN, M. T. AND M. B. MENHAJ (1994): "Training Feedforward Networks with the Marquardt Algorithm," *IEEE Transactions on Neural Networks*, 5, 989–993.

HANIN, B. AND M. SELLKE (2018): "Approximating Continuous Functions by ReLU Nets of Minimal Width," *arXiv:1710.11278 [cs, math, stat]*, arXiv: 1710.11278.

HANSEN, P. R. (2005): "A Test for Superior Predictive Ability," *Journal of Business & Economic Statistics*, 23, 365–380.

HINTERLANG, N. (2020): "Predicting Monetary Policy using Artificial Neural Networks," *Deutsche Bundesbank Discussion Paper*.

HORNIK, K., M. STINCHCOMBE, AND H. WHITE (1989): "Multilayer Feedforward Networks are Universal Approximators," *Neural networks*, 2, 359–366.

KAASTRA, I. AND M. BOYD (1996): "Designing a Neural Network for Forecasting Financial and Economic Time Series," *Neurocomputing*, 10, 215–236.

KARLSSON, S. (2013): "Forecasting with Bayesian Vector Autoregression," in *Handbook of Economic Forecasting*, Elsevier, vol. 2, 791–897.

KOLASA, M., M. RUBASZEK, AND P. SKRZYPCZYŃSKI (2012): "Putting the New Keynesian DSGE Model to the Real-Time Forecasting Test," *Journal of Money, Credit and Banking*, 44, 1301–1324.

LLOYD, S. (1982): "Least Squares Quantization in PCM," *IEEE Transactions on Information Theory*, 28, 129–137.

MACKAY, D. J. (1992): "Bayesian Interpolation," *Neural Computation*, 4, 415–447.

MARCELLINO, M. (2004): "Forecasting EMU Macroeconomic Variables," *International Journal of Forecasting*, 20, 359–372.

MEDEIROS, M. C., G. F. VASCONCELOS, Á. VEIGA, AND E. ZILBERMAN (2021): "Forecasting Inflation in a Data-Rich Environment: The Benefits of Machine Learning Methods," *Journal of Business & Economic Statistics*, 39, 98–119.

MOHAMMADI, S. (2020): "Neural Network for Univariate and Multivariate Nonlinearity Tests," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 13, 50–70.

PARANHOS, L. (2021): "Predicting Inflation with Neural Networks," *arXiv preprint arXiv:2104.03757*.

PEÑA, D. AND I. SÁNCHEZ (2007): "Measuring the advantages of multivariate vs. univariate forecasts," *Journal of Time Series Analysis*, 28, 886–909.

PFEIFER, J. (2014): "A Guide to Specifying Observation Equations for the Estimation of DSGE Models," *Research Series*, 1–150.

POLITIS, D. N. AND J. P. ROMANO (1994): "The Stationary Bootstrap," *Journal of the American Statistical Association*, 89, 1303–1313.

RAPACH, D., G. ZHOU, G. ELLIOTT, AND A. TIMMERMANN (2013): "Handbook of Economic Forecasting," *G. Elliott and A. Timmermann, Eds., Handbook of Economic Forecasting*.

RUBASZEK, M. AND P. SKRZYPCZYŃSKI (2008): "On the Forecasting Performance of a Small-Scale DSGE Model," *International Journal of Forecasting*, 24, 498–512.

SHEPPARD, K. (2009): "MFE MATLAB Function Reference Financial Econometrics," *Unpublished paper, Oxford University, Oxford. Available at: http://www. kevinsheppard. com/images/9/95/MFE_Toolbox_Documentation. pdf*.

SIMS, C. ET AL. (1999): "Matlab Optimization Software," *QM&RBC Codes*.

SIMS, C. A. (2002): "Solving linear rational expectations models," *Computational economics*, 20, 1.

SMALTER HALL, A. AND T. R. COOK (2017): "Macroeconomic Indicator Forecasting with Deep Neural Networks," *Federal Reserve Bank of Kansas City Working Paper*.

SMETS, F. AND R. WOUTERS (2007): "Shocks and Frictions in US Business Cycles: A Bayesian DSGE Approach," *American Economic Review*, 97, 586–606.

STOCK, J. H. AND M. W. WATSON (2007): "Why has US Inflation Become Harder to Forecast?" *Journal of Money, Credit and Banking*, 39, 3–33.

SWANSON, N. R. AND H. WHITE (1997): "A Model Selection Approach to Real-Time Macroeconomic Forecasting using Linear Models and Artificial Neural Networks," *Review of Economics and Statistics*, 79, 540–550.

TERÄSVIRTA, T., C.-F. LIN, AND C. W. GRANGER (1993): "Power of the Neural Network Linearity Test," *Journal of Time Series Analysis*, 14, 209–220.

TICKNOR, J. L. (2013): "A Bayesian regularized artificial neural network for stock market forecasting," *Expert Systems with Applications*, 40, 5501–5506.

TSAY, R. S. (1986): "Nonlinearity Tests for Time Series," *Biometrika*, 73, 461–466.

TULIP, P. (2009): "Has the Economy Become More Predictable? Changes in Greenbook Forecast Accuracy," *Journal of Money, Credit and Banking*, 41, 1217–1231.

VERSTYUK, S. (2020): "Modeling Multivariate Time Series in Economics: From Auto-Regressions to Recurrent Neural Networks," *Available at SSRN 3589337*.

WHITE, H. (2000): "A Reality Check for Data Snooping," *Econometrica*, 68, 1097–1126.

WIELAND, V., E. AFANASYEVA, M. KUETE, AND J. YOO (2016): "New Methods for Macro-Financial Model Comparison and Policy Analysis," *Handbook of Macroeconomics*, 2, 1241–1319.

WIELAND, V., T. CWIK, G. J. MÜLLER, S. SCHMIDT, AND M. WOLTERS (2012): "A New Comparative Approach to Macroeconomic Modeling and Policy Analysis," *Journal of Economic Behavior & Organization*, 83, 523–541.

WIELAND, V. AND M. H. WOLTERS (2011): "The Diversity of Forecasts from Macroeconomic Models of the US Economy," *Economic Theory*, 47, 247–292.

ZHANG, G., B. E. PATUWO, AND M. Y. HU (1998): "Forecasting with Artificial Neural Networks:: The State of the Art," *International Journal of Forecasting*, 14, 35–62.

Appendix

A. ANN

A.1. Bayesian Regularization Backpropagation

Bayesian regularization means a Bayesian estimation of the regularization parameters, in this case with a Gauss-Newton approximation to the Hessian matrix. This method can conveniently be combined with the Levenberg-Marquardt algorithm (Hagan and Menhaj, 1994). Here, backpropagation is used to calculate the Jacobian of the performance function with respect to weights and biases. According to Foresee and Hagan (1997), page 3, the approach follows these steps:

1. Initialize α, β and the weights.
2. Take one step of the Levenberg-Marquardt algorithm to minimize the objective function
$$F(w) = \beta E_D + \alpha E_W.$$
3. Compute the effective number of parameters
$$\gamma = N - 2\alpha tr(H)^{-1}$$
making use of the Gauss-Newton approximation of the Hessian available in the Levenberg-Marquardt training algorithm:
$$H = \nabla^2 F(w) \approx 2\beta J^T J + 2\alpha I_N,$$
where J is the Jacobian matrix of the training set errors.
4. Compute new estimates for the objective function parameters
$$\begin{aligned} \alpha &= \frac{\gamma}{2E_W(w)} \\ \beta &= \frac{n-\gamma}{2E_D(w)}. \end{aligned}$$
5. Iterate through steps 1 to 3 until convergence.

B. K-Means Clustering

The k-means clustering approach, used to identify crisis periods, follows Lloyd (1982) and assigns n observations to exactly one of k clusters which are defined by centroids. The iterative algorithm proceeds as follows:

1. Randomly choose k initial cluster centers.
2. Compute point to cluster centroid distances of all observations to each centroid.
3. Assign each observation to the cluster with the closest centroid.
4. Compute average of observations in every cluster and obtain k new centroid locations.
5. Loop through steps 2 to 4 until cluster assignments converge.

C. BVAR

C.1. The GLP Prior

In order to choose the tightness (the informativeness) of the prior optimally, Giannone et al. (2015) suggest to use Bayesian techniques as it is conceptually identical to the inference of all other unknown model parameters. Let $p(y|\theta)$ be the likelihood function of the model with a prior distribution $p_\gamma(\theta)$ with θ being the VAR model parameters and γ the hyperparameters. Based on assuming a hierarchical model and rewriting $p_\gamma(\theta)$ with $p(\theta|\gamma)$, the posterior can be obtained applying Bayes' Law:

$$p(\gamma|y) \propto p(y|\gamma)p(\gamma)$$

with $p(\gamma)$ being the prior density of the hyperparameters (the hyperprior) and $p(y|\gamma)$ constituting the marginal likelihood. Hereby, the authors take the most frequently used conjugate priors into account: Minnesota, sum of coefficients and dummy initial observation priors. Key concept of their procedure is the automatic choice of the appropriate amount of shrinkage, i.e. the selection of tighter priors in case that many unknown coefficients are involved relative to the available data, and looser priors vice versa. The prior distributions are assumed to be of normal-inverse-Wishart form:

$$\begin{aligned} \Sigma &\sim IW(\Psi; d) \\ \beta|\Sigma &\sim N(b, \Sigma \otimes \Omega) \end{aligned}$$

where Ψ, d, b and Ω are functions of hyperparameters γ ($d = n + 2$ and $\Sigma = \Psi/(d - n - 1)$).

Three prior densities are combined for the unconditional prior: The Minnesota prior hinges on the assumption that every variable follows a random walk process (with drift). The moments of this prior are characterized by

$$E[(B_s)_{ij}|\Sigma] = \begin{cases} 1 & \text{if } i = j \text{ and } s = 1 \\ 0 & \text{otherwise} \end{cases}$$

$$cov((B_s)_{ij}, (B_r)_{hm}|\Sigma) = \begin{cases} \lambda^2 \dfrac{1}{s^2} \dfrac{\Sigma_{ih}}{\Psi_j/(d-n-1)} & \text{if } m = j \text{ and } r = s \\ 0 & \text{otherwise} \end{cases}$$

where the most important hyperparameter, which drives the overall tightness of this prior, is λ. Next, the sum-of-coefficients prior is considered and implemented using a Theil mixed estimation based an n artificial observations (with n = number of vari-

ables). Giannone et al. (2015) create a set of dummy observations

$$\underset{n\times n}{y^{+}} = diag\left(\frac{\bar{y_0}}{\mu}\right)$$
$$\underset{n\times(1+np)}{x^{+}} = \left[\underset{n\times 1}{0}, y^{+}, ..., y^{+}\right]$$

where $diag(v)$ is the diagonal matrix with the vector v on the main diagonal and where $\bar{y_0}$ is this vector containing the average of the first p observations of each variable. The variance of these prior beliefs is controlled by μ, which is the hyperparameter of interest here. For $\mu \to$ inf, one gets an uninformative prior and for $\mu \to 0$, there is a unit root in every equation which rules out cointegration. The third prior (dummy-initial-observation) is implemented by

$$\underset{1\times n}{y^{++}} = \frac{\bar{y_0}'}{\delta}$$
$$\underset{1\times(1+np)}{x^{++}} = \left[\frac{1}{\delta}, y^{++}, ..., y^{++}\right]$$

where the tightness of the prior is controlled by δ. It becomes uninformative when $\delta \to$ inf and for $\delta \to 0$, all variables are forced to their unconditional mean.

The resulting set of hyperparameters λ, μ, δ and Ψ are treated as additional model parameters, with their hyperpriors being gamma densities, with mode equal to 0.2, 1 and 1 and standard deviations equal to 0.4, 1 and 1. Lastly for Σ, the hyperprior for $\Psi/(d-n-1)$, the authors select an Inverse-Gamma distribution with scale and shape equal to 0.02^2. The resulting marginal likelihood weighs the in-sample fit against the model-complexity. As to Giannone et al. (2015), this procedure produces precise out-of-sample predictions using point and density forecasts.

According to Giannone et al. (2015), the joint posterior density is not available in closed form and is thus simulated from a Gaussian proposal distribution based on a Markov chain Monte Carlo (MCMC) with Metropolis Hastings algorithm. In this paper, I use the original codes offered by the authors. Similar to a DSGE model setting, this BVAR algorithm also requires to define an appropriate jump size which is calibrated to yield an acceptance rate around 25%.

D. DSGE

D.1. The DSGE Model Solution

To solve the DSGE model, one can rewrite intertemporal optimization problems of the agents using Bellmann equations. The equilibrium law of motion is then written as

$$s_t = \Phi(s_{t-1}, \epsilon_t, \theta).$$

Here, s_t is a vector of state variables and ϵ_t is a vector including the innovations of structural shocks. The equilibrium conditions by the log-linearized DSGE model form a system of linear rational expectations difference equations which can be written as

$$\Gamma_0(\theta)s_t = \Gamma_1(\theta)s_{t-1} + \Psi(\theta)\epsilon_t + \Pi(\theta)\eta_t$$

with η is a vector of rational expectations forecast errors. This system can be solved using the technique by Sims (2002) which allows to express η_t as a function of ϵ_t subject to the constraint of a non-explosive law of motion of s_t. The solution is then given by

$$s_t = \Phi_1(\theta)s_{t-1} + \Phi_\epsilon(\theta)\epsilon_t$$

where Φ_1 and Φ_ϵ are functions of model parameters θ. A simpler representation of the measurement equation is given by

$$y_t = \Psi_1(\theta)s_t + \Psi_0(\theta).$$

The state-space representation of the DSGE model is given by the last two equations.

D.2. Measurement Equations

In order to estimate the model, a set of measurement equations is defined that relate elements of s_t to the set of observable variables (see Pfeifer (2014) for a reference):

Output growth	$=\gamma + (y_t - y_{t-1} + z_t)$
Consumption growth	$=\gamma + (c_t - c_{t-1} + z_t)$
Investment growth	$=\gamma + (i_t - i_{t-1} + z_t)$
Real wage growth	$=\gamma + (w_t - w_{t-1} + z_t)$
Hours worked	$=\bar{l} + l_t$
Inflation	$=\pi^* + \pi_t$
FFR	$=R^* + R_t$
Spread	$=SP^* + E_t(\tilde{R}^k_{t+1} - R_t).$

D.3. DSGE Model Estimation

The estimation of the DSGE model now builds on the introduced state-space representation. Assuming that the exogenous shocks are Gaussian, the likelihood function $p(y|\theta)$ for the observable data given the model parameters θ can be evaluated with the Kalman filter. The posterior parameter distribution $p(\theta|y)$ can then be generated using

$$p(\theta|y) \propto p(y|\theta)p(\theta),$$

where $p(\theta)$ is the likelihood function of the prior parameter distribution. A Metropolis-Hastings algorithm is then employed since no analytic expression exists for the pos-

terior distribution. Choosing a starting point for the parameters θ^0, several steps are repeated in a loop (see e.g. Berg (2016) for a detailed description which I summarize here). Draw θ^* from a jumping distribution $J(\theta^*|\theta^{j-1}) = N(\theta^{j-1}, c, \Sigma_m)$, where the inverse of the Hessian Σ_m is computed at the posterior mode and the scaling constant c is chosen to produce an acceptance rate of 25 % to 33%. Then, the acceptance ratio is calculated by $r = p(\theta^*|y)/p(\theta^{j-1}|y)$. After randomly drawing ν from $U(0,1)$, the proposal θ^* is either accepted or discarded and the jumping distribution is updated. Further, a draw of the state variables $s^j|\theta^j$ is obtained from $N(\hat{s}^j, P^j)$, where both are computed with the Kalman filter. Now, generating $\epsilon^j_{T+1}, ..., \epsilon^j_{T+H}$ from $\epsilon_t \sim N(0, I_q)$, a path for the model variables may be established using the transition equation:

$$s_t = \Phi^m_1(\theta^j)s^j_{t-1} + \Phi^m_\epsilon(\theta^j)\epsilon^j_t.$$

Having simulated the paths of the state variables, a path for the observational variables may be discarded relying on the measurement equations:

$$\hat{y}^j_t = \Psi_1(\theta^j)s_t + \Psi_0(\theta^j).$$

Hence, a sample of forecasts are generated from the posterior predictive distribution.

D.4. Settings Estimation

In the following, some important settings for the Bayesian estimation are described in more detail. All unmentioned features are left on default values.

Table 4: Settings for Bayesian Estimation

Feature	**Setting**	**Description**
presample	4	Number of values to be skipped
lik_init	1	Kalman filter initialization
mh_replic	20.000	Number of replications
mh_nblocks	2	Number of parallel chains
mh_drop	0.5	Fraction of parameter vectors to be dropped
mh_jscale	0.275	Scale parameter of the jumping distribution's covariance matrix
mode_compute	4 (6)	Specification of optimizer for mode computation

Setting the *presample* to 4 makes the algorithm skip the first four observations before evaluating the likelihood. However, the values are used as a training sample for starting the Kalman filter iterations. Setting this value to 4 makes the DSGE estimation comparable to a BVAR(4) and ANN(4).

The initialization of the Kalman filter is steered with *lik_init*. A value of 1 can be used for stationary models (as DNGS15) and means that the initial matrix of variance of the error forecast is set equal to the unconditional variance of the state variables.

Several settings can be adjusted for the Metropolis-Hastings algorithm. First, the number of replications is defined by *mh_repplic* and set equal to 20000. The applied number of parallel chains (*mh_nblocks*) is set to 2. Based on these settings, the Monte Carlo Markov Chain (MCMC) diagnostics are generated with the convergence diagnostics according to Brooks and Gelman (1998)[24]. These are based on comparing pooled and within MCMC moments and the length of the highest probability density interval covering 80% of the posterior distribution. Another parameter, *mh_drop* is set to 0.5,which defines that 50% of the initially generates parameter vectors are dropped as a burn-in before using posterior simulations. Last, the *mh_jscale* which is the scale parameter of the jumping distribution's covariance matrix is set. It must be tuned to obtain an acceptance ratio of 25% to 33%. The idea behind this is to increase the variance of the jumping distribution if the acceptance ratio is too high and vice versa. For this project, a value of 0.275 leads to the desired acceptance ratio and is not changed during the expanding window procedure.

Further, the optimizer for the mode computation *mode_compute* is specified. Most of the vintages in the expanding window procedure are estimated by the so-called *csminwel* (around 85%, *mode_compute*=4), for the remaining vintages this setting leads to errors during the estimation and is thus replaced by *mode_compute*=6. This applies a Monte-Carlo based optimization routine first, the goal of which is to identify a region to start the Metropolis-Hastings algorithm and an initial estimate of the posterior covariance matrix of the parameters to be estimated. Advantageous is the fact that the MH algorithm may start from a point with a high posterior density value and not from the posterior mode, to estimate the covariance matrix of the jumping distribution. The favored algorithm, however, is the *csminwel*, developed by Sims et al. (1999), which is often used for Bayesian estimation because of its properties. The algorithm minimizes using a quasi-Newton method with BFGS update of the estimated inverse hessian. It is known to be very robust against certain pathologies common to likelihood function as it tries random search directions and avoids getting stuck at flat spots for example.

[24]For multivariate problems, the procedure does not strictly follow Brooks and Gelman (1998), but transfers their approach to the range of posterior likelihood function instead of the individual parameter. The posterior kernel is the used to combine the parameters into a scalar statistic and check its convergence using the authors univariate convergence diagnostic.

E. Data

E.1. Data Sources

Table 5: Data Sources and Access URL

FRED	https://fred.stlouisfed.org
ALFRED	https://alfred.stlouisfed.org
RTDSM	Real-Time Data Set for Macroeconomists https://www.philadelphiafed.org
Greenbook Predictions	https://www.philadelphiafed.org
SPF Predictions	https://www.philadelphiafed.org

E.2. Data Transformation

Per capita growth rates are computed taking advantage of the population level measure. As Pfeifer (2014) and Edge and Gürkaynak (2010) recommend, a smoothed value of this series should be used to adjust data from other sources[25]. The smoothed population growth rate is given by

$$PG_t = HP(CNP16OV_t/CNP16OV_{t-1}).$$

Leaning on Del Negro et al. (2015), the 8 desired variables are generated as follows:

$$\begin{aligned}
\text{Output growth}_t &= 100 * ln\left(\frac{ROUTPUT_t}{ROUTPUT_{t-1}} * PG_t^{-1}\right) \\
\text{Consumption growth}_t &= 100 * ln\left(\frac{CONS_t}{CONS_{t-1}} * \frac{GDPDEF_{t-1}}{GDPDEF_t} * PG_t^{-1}\right) \\
\text{Investment growth}_t &= 100 * ln\left(\frac{FPI_t}{FPI_{t-1}} * \frac{GDPDEF_{t-1}}{GDPDEF_t} * PG_t^{-1}\right) \\
\text{Real wage growth}_t &= 100 * ln\left(\frac{COMPNFB_t}{COMPNFB_{t-1}} * \frac{GDPDEF_{t-1}}{GDPDEF_t}\right) \\
\text{Hours worked}_t &= 100 * ln\left(\frac{AWHNONAG_t * CE16OV_t/100}{CNP16OV_t}\right)^{26} \\
\text{Inflation}_t &= 100 * ln\left(\frac{GDPDEF_t}{GDPDEF_{t-1}}\right) \\
\text{FFR} &= \frac{1}{4} * EFF \\
\text{Spread} &= \frac{1}{4} * (BAA10YM_t)
\end{aligned}$$

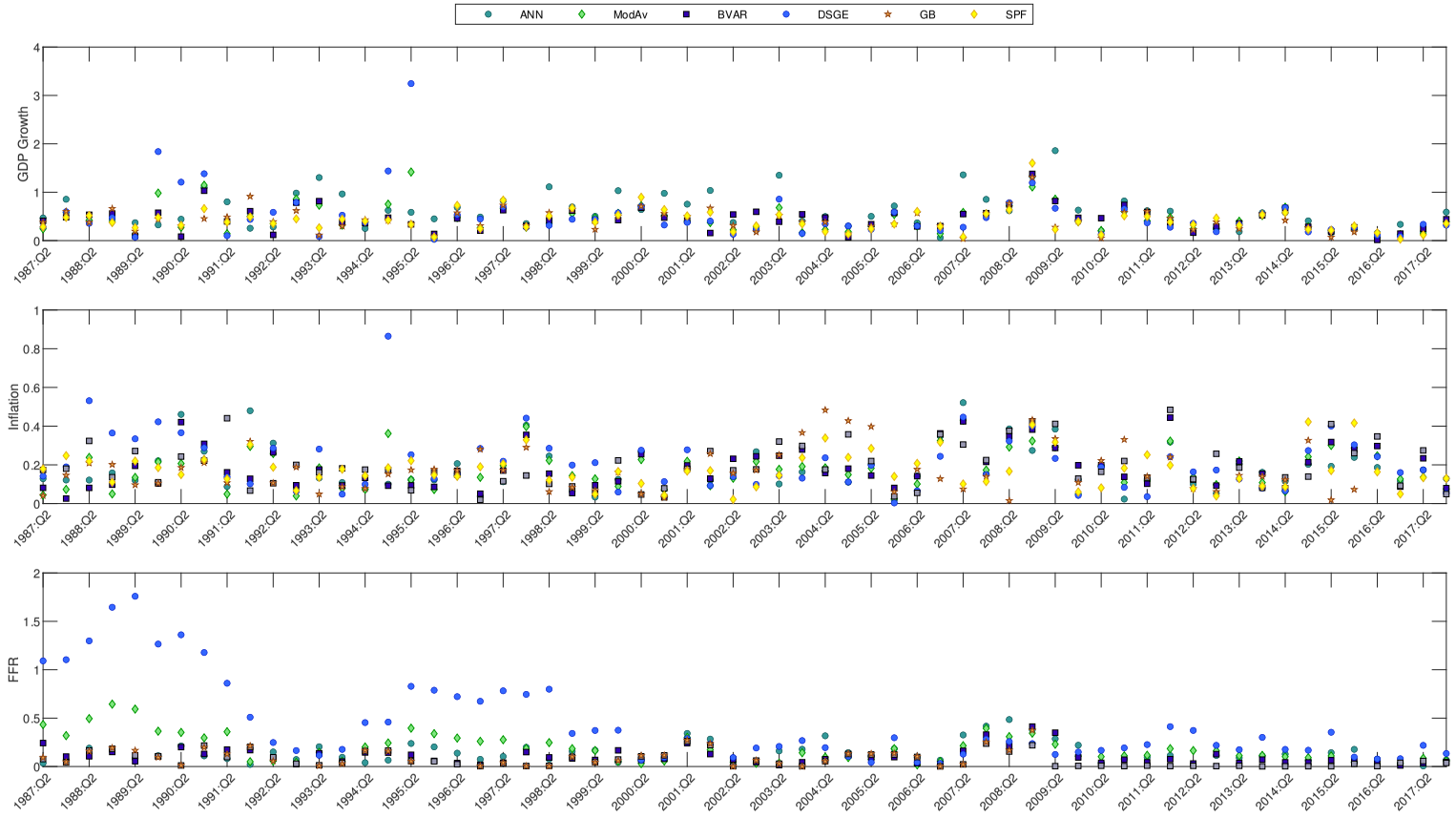
[25]Following the example in Pfeifer (2014), the HP-filter with a smoothing parameter of 10,000 is applied.

While this data specification is required by the DSGE model at hand, the BVAR and the ANN are more flexible with respect to the data format. Following the argumentation by Karlsson (2013), saying that in terms of root mean squared error-measured forecasting performance, a variable specification in differences succeeds one in levels, and for the sake of comparability of the resulting forecasts, the BVAR and also the ANN are estimated with the identical transformed data as the DSGE[27]. The official forecasters' data on real GDP growth is also transformed to obtain population growth adjusted per capita values.

[27]While a specification in levels can make use of any co-integration between the variables, a specification in differences offers some robustness in the presence of structural breaks. The RMSE of the difference-specification is on average 11% smaller.

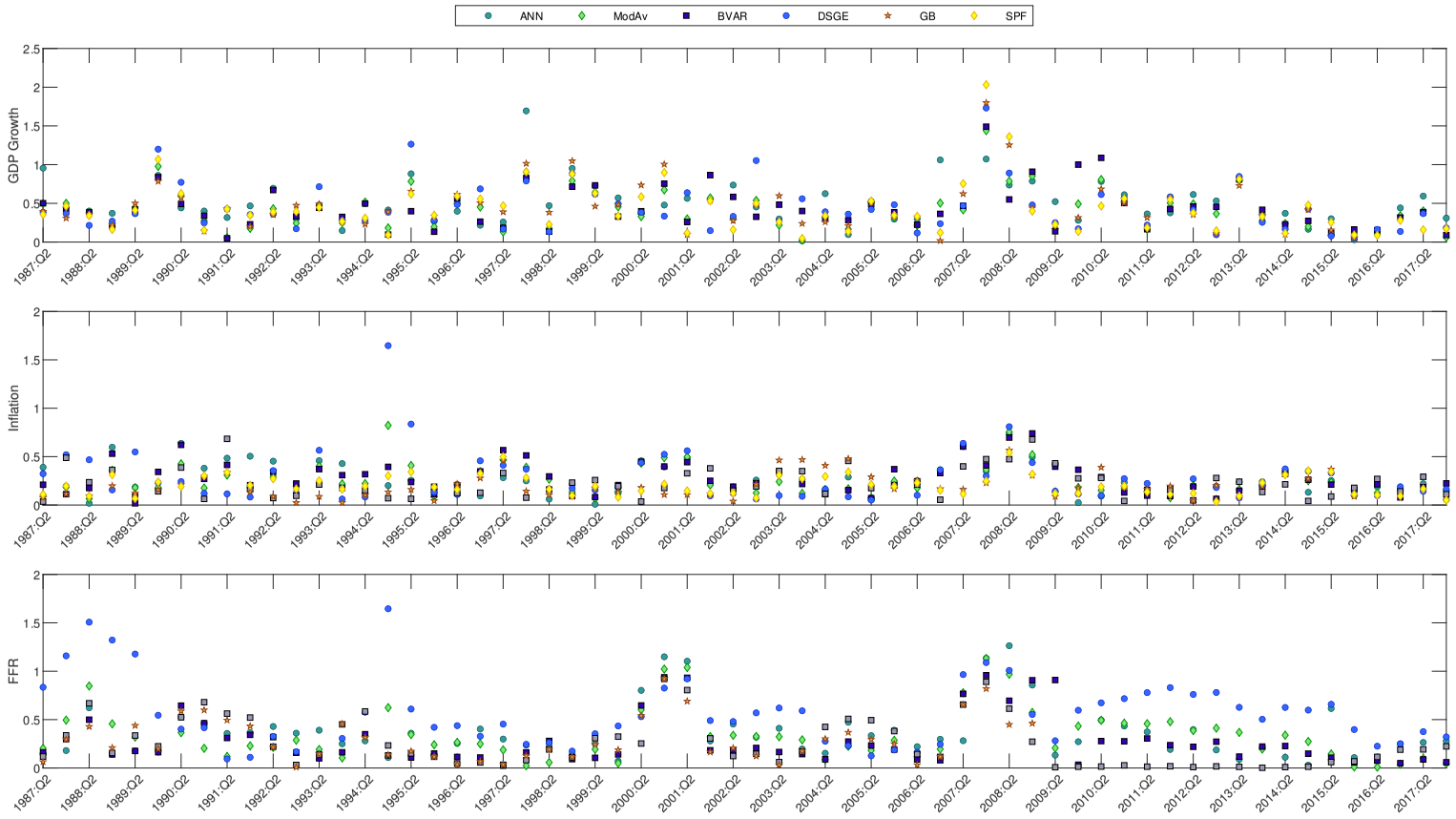
F. Forecast Evaluation

Figure 5: Absolute Forecast Errors over Time (h=1)



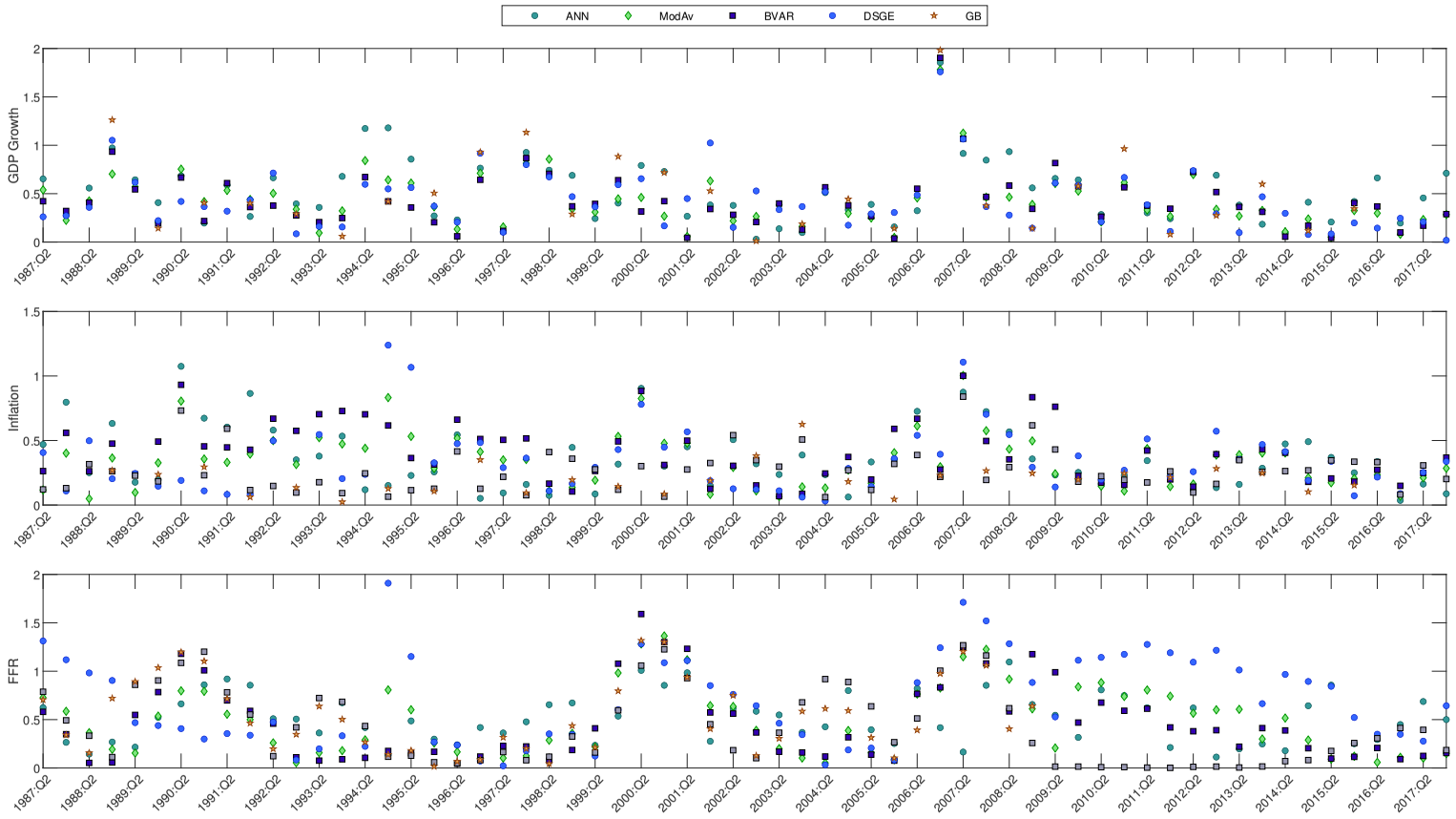
Note: The absolute forecast errors are averaged over every 2 quarters. The time labeled on the x-asis indicates the last quarter of each average.

Figure 6: Absolute Forecast Errors over Time (h=4)



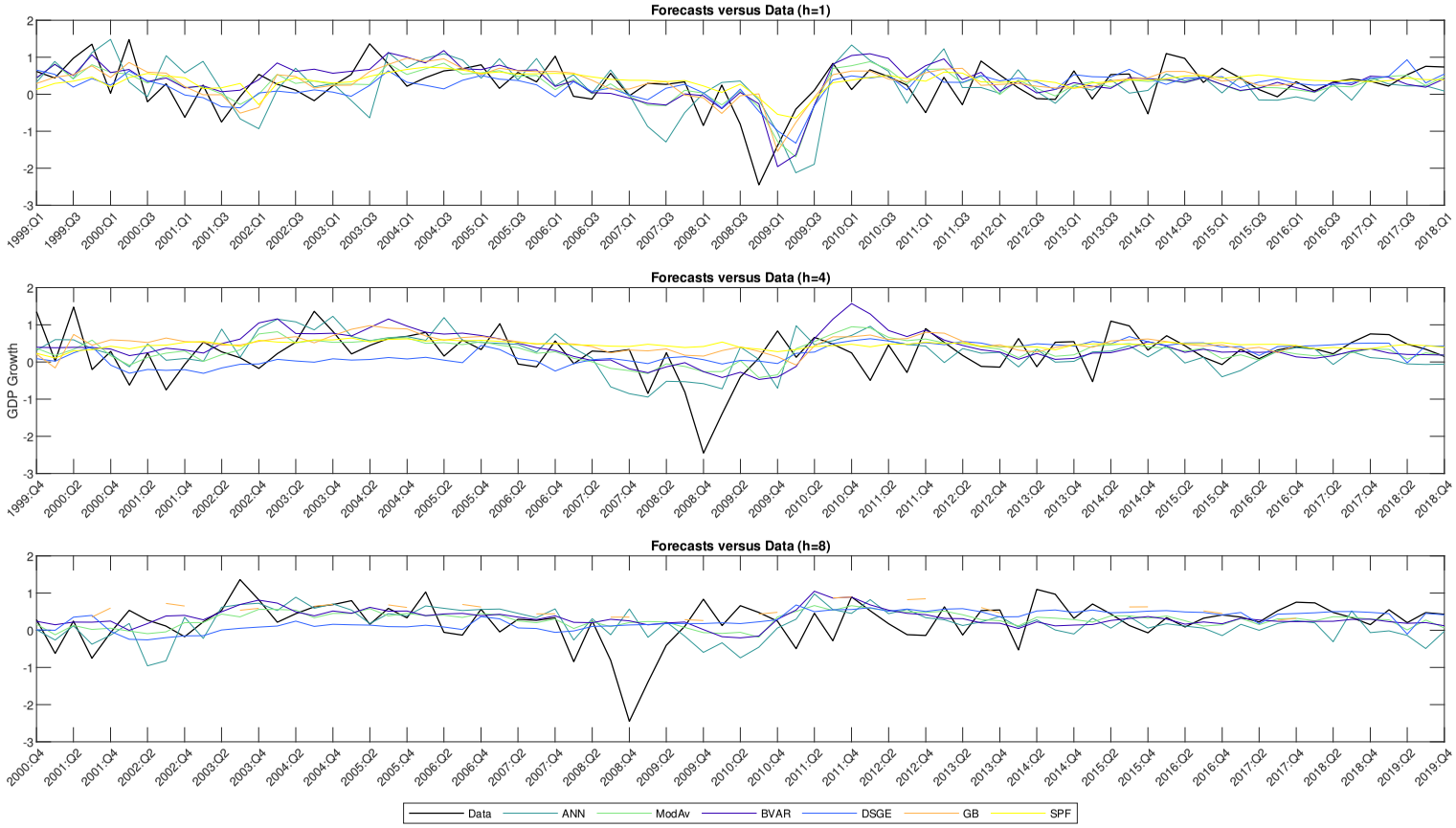
Note: The absolute forecast errors are averaged over every 2 quarters. The time labeled on the x-asis indicates the last quarter of each average.

Figure 7: Absolute Forecast Errors over Time (h=8)



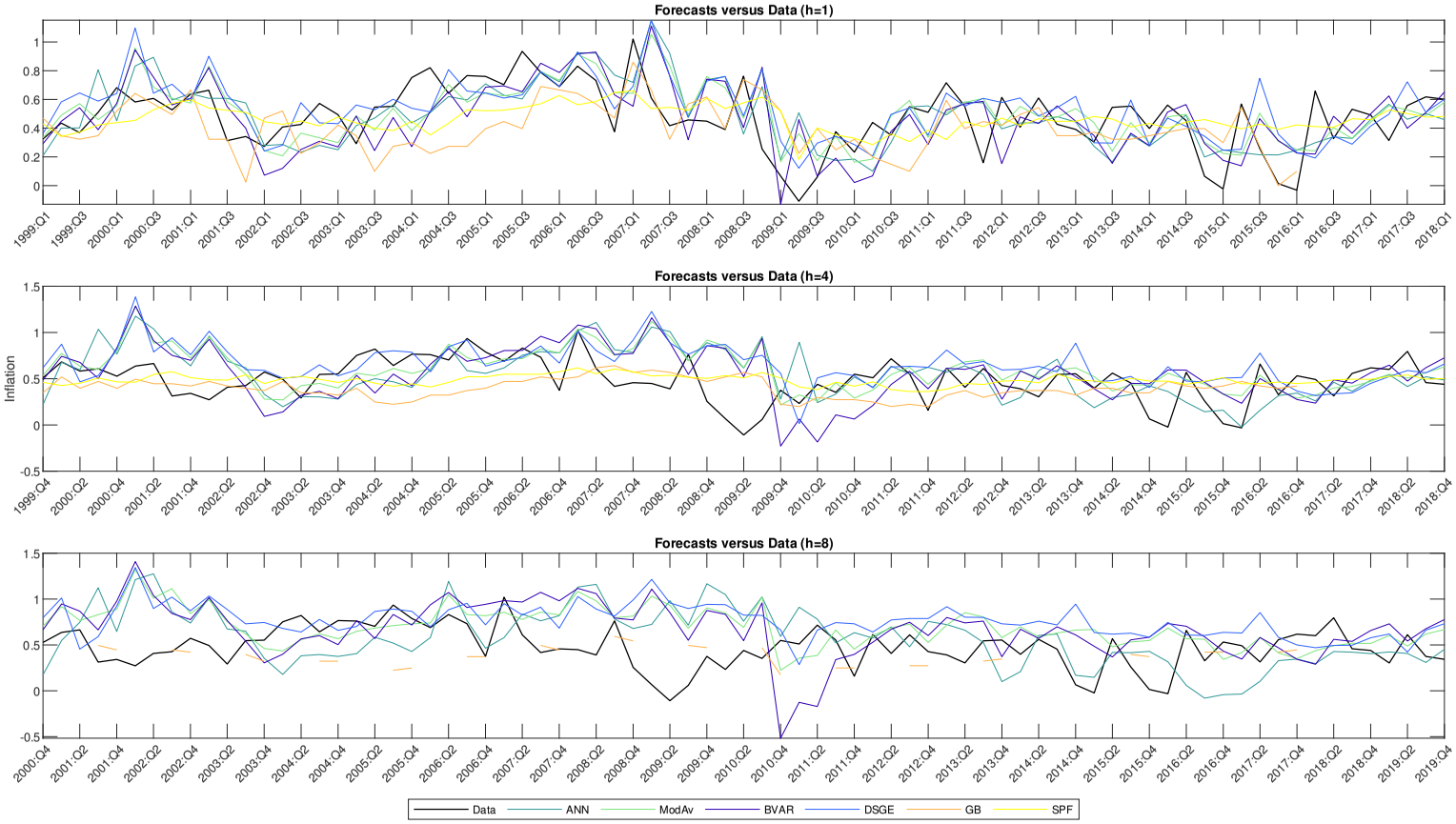
Note: The absolute forecast errors are averaged over every 2 quarters. The time labeled on the x-asis indicates the last quarter of each average.

Figure 8: Forecast GDP 1999Q1:2017Q4



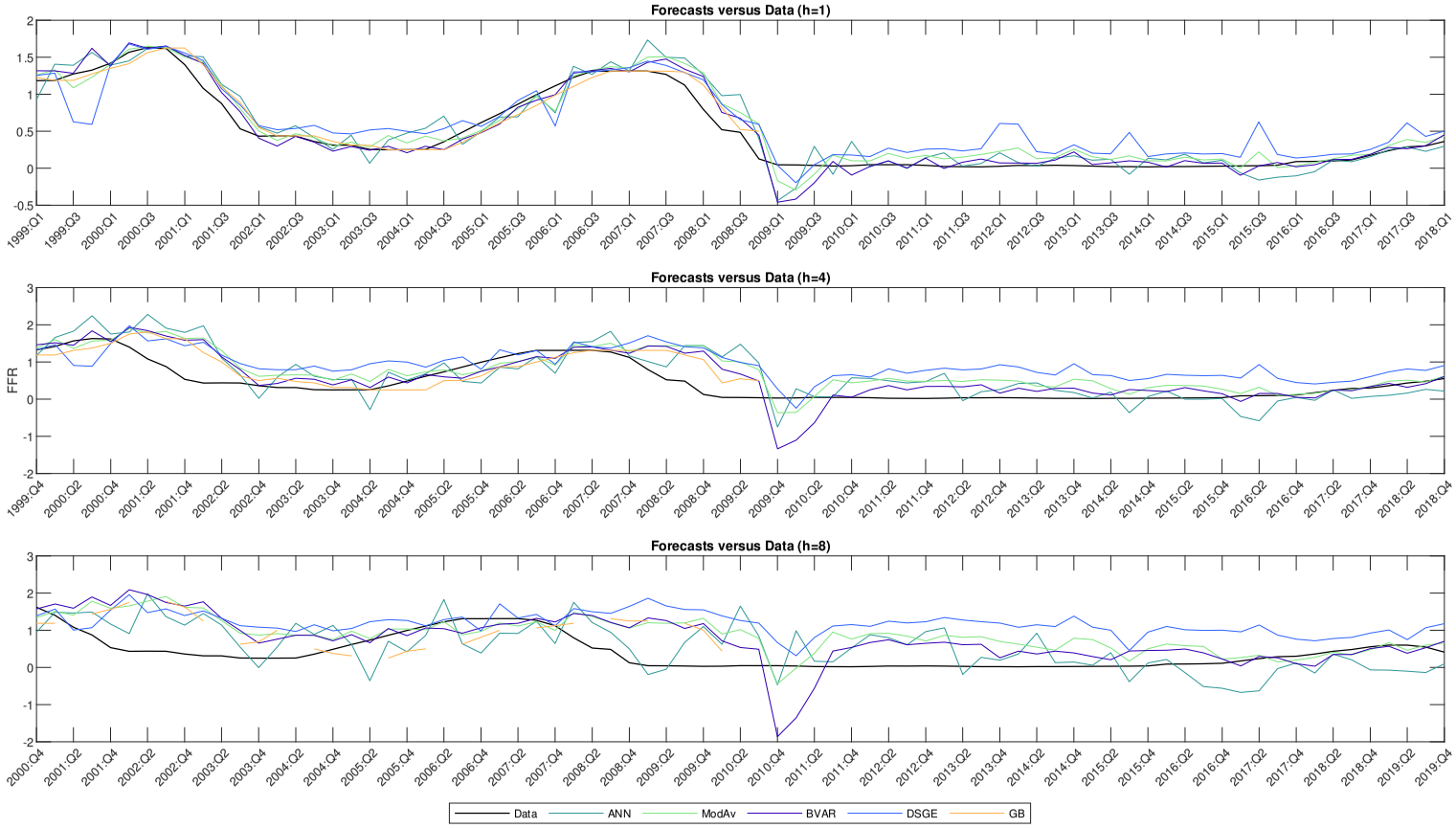
Note: Forecasts of all models for horizons $h = 1, 4, 8$ are plotted against the actual data.

Figure 9: Forecast Inflation 1999Q1:2017Q4



Note: Forecasts of all models for horizons $h = 1, 4, 8$ are plotted against the actual data.

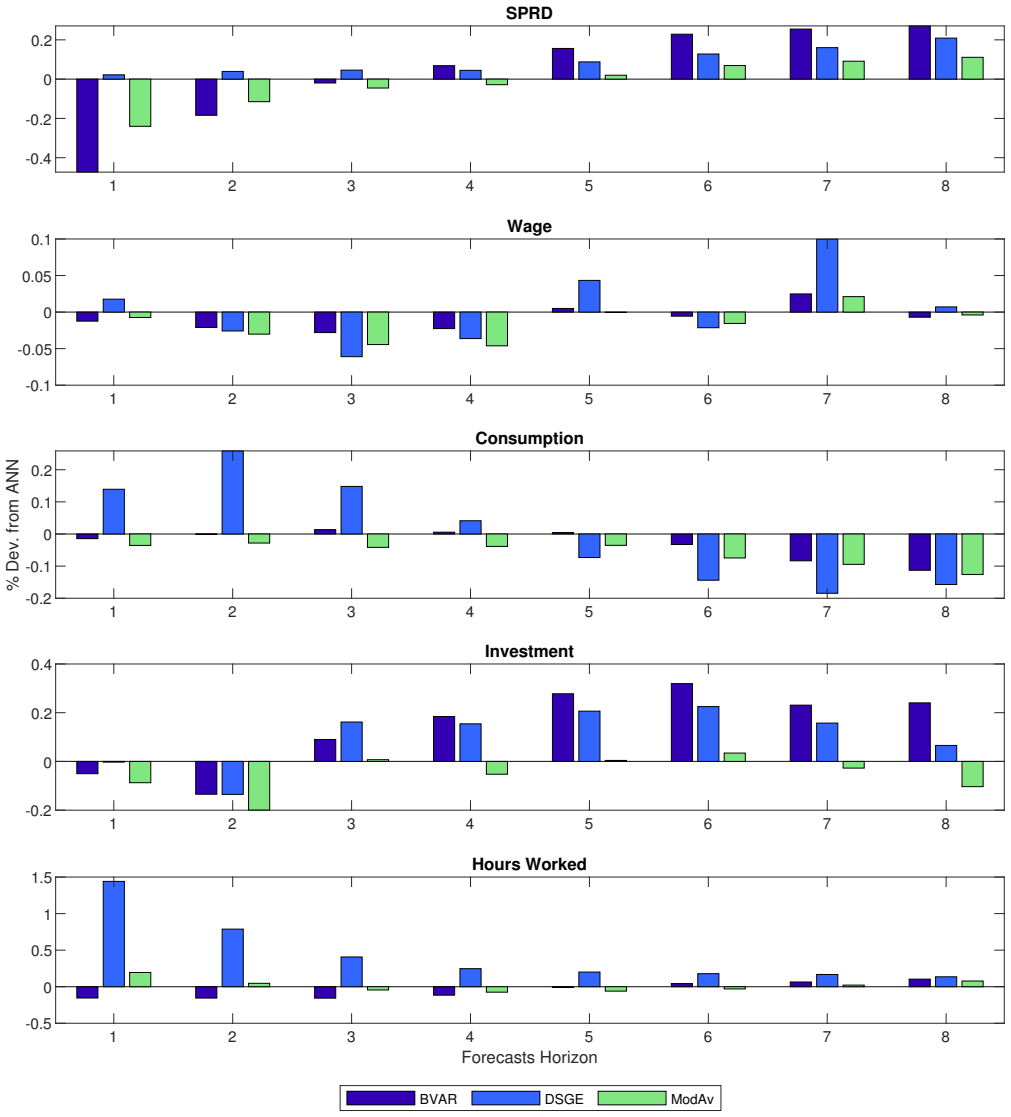
Figure 10: Forecast FFR 1999Q1:2017Q4



Note: Forecasts of all models for horizons $h = 1, 4, 8$ are plotted against the actual data.

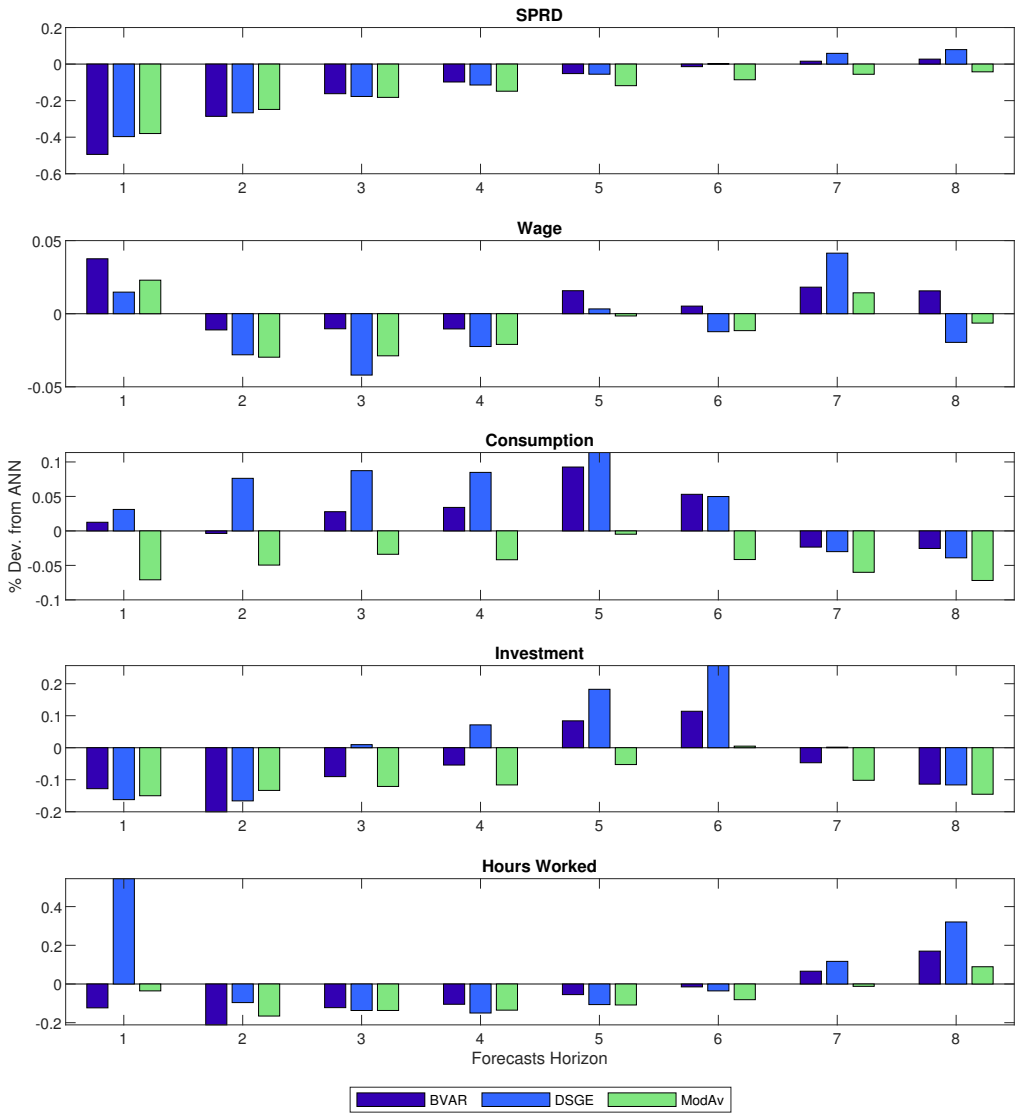
G. *Additional Results*

Figure 11: Relative RMSFE Section 1



Note: This figure shows RMSFEs of individual variables as percentage deviation from the ANN over all forecast horizons.

Figure 12: Relative RMSFE Section 2



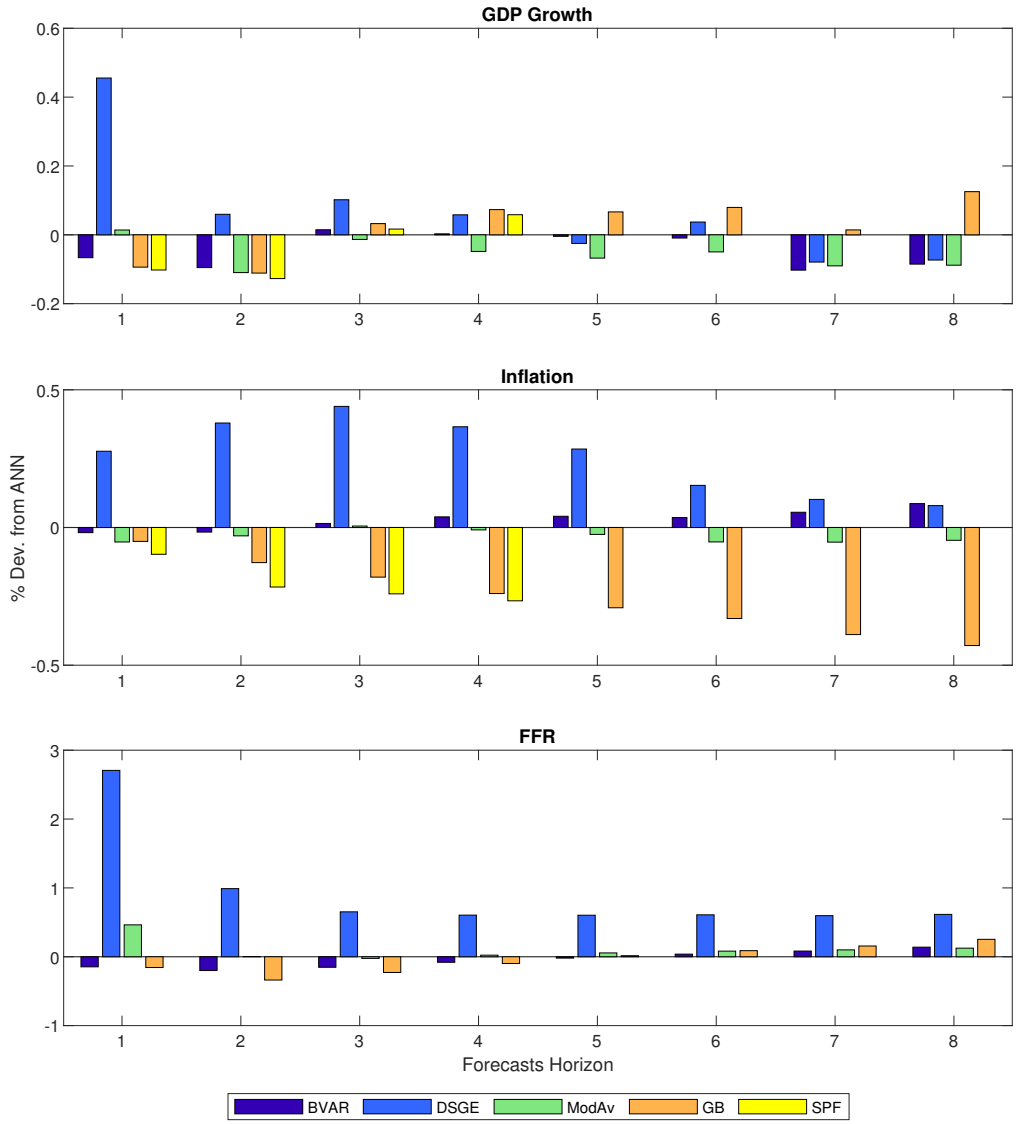
Note: This figure shows RMSFEs of individual variables as percentage deviation from the ANN over all forecast horizons.

H. Additional Analyses

In this section, more detailed analyses are provided. The overarching forecasting section, *Section 3* is shown at first. Further, a closer look is taken at the individual crises periods around the 2001 and 2008 crises. Further, in order to investigate the advantage of multivariate over univariate methods on the one hand, and a source of the superiority of the ANN on the other hand, statistical tests are provided.

H.1. 1987Q3 to 2017Q4 (Section 3)

Figure 13: Relative RMSFE Section 3



Note: This figure shows relative RMSFEs, taking the ANN as benchmark, over all forecast horizons.

Figure 13 gives the RMSFEs relative to the ANN for each individual variable. It becomes clear that with respect to GDP, the medium-term predictions by the ANN are superior to other models and the official forecasts, BVAR forecasts are better but inferior to officials' in the short run, while in the long-run both DSGE and BVAR provide the best predictions. The picture changes looking at inflation: although other research

claims that starting inflation forecasts in the great moderation is difficult (Stock and Watson (2007) and Tulip (2009)), the ANN can improve upon the other models, however it does not approach the official forecasts. With respect to the federal funds rate, short-term predictions are dominated by the BVAR and the GB, while from horizon 5 onwards, the ANN succeeds. All of these findings are supported by the test results for superior predictive accuracy (SPA), shown in Table 12. Relative RMSFEs of the remaining variables are shown in Figure 14. It proves overall inferiority of the ANN with respect to investment forecasts, the spread is forecasted well in the long-term, while for wage and consumption the results fluctuate over forecast horizons.

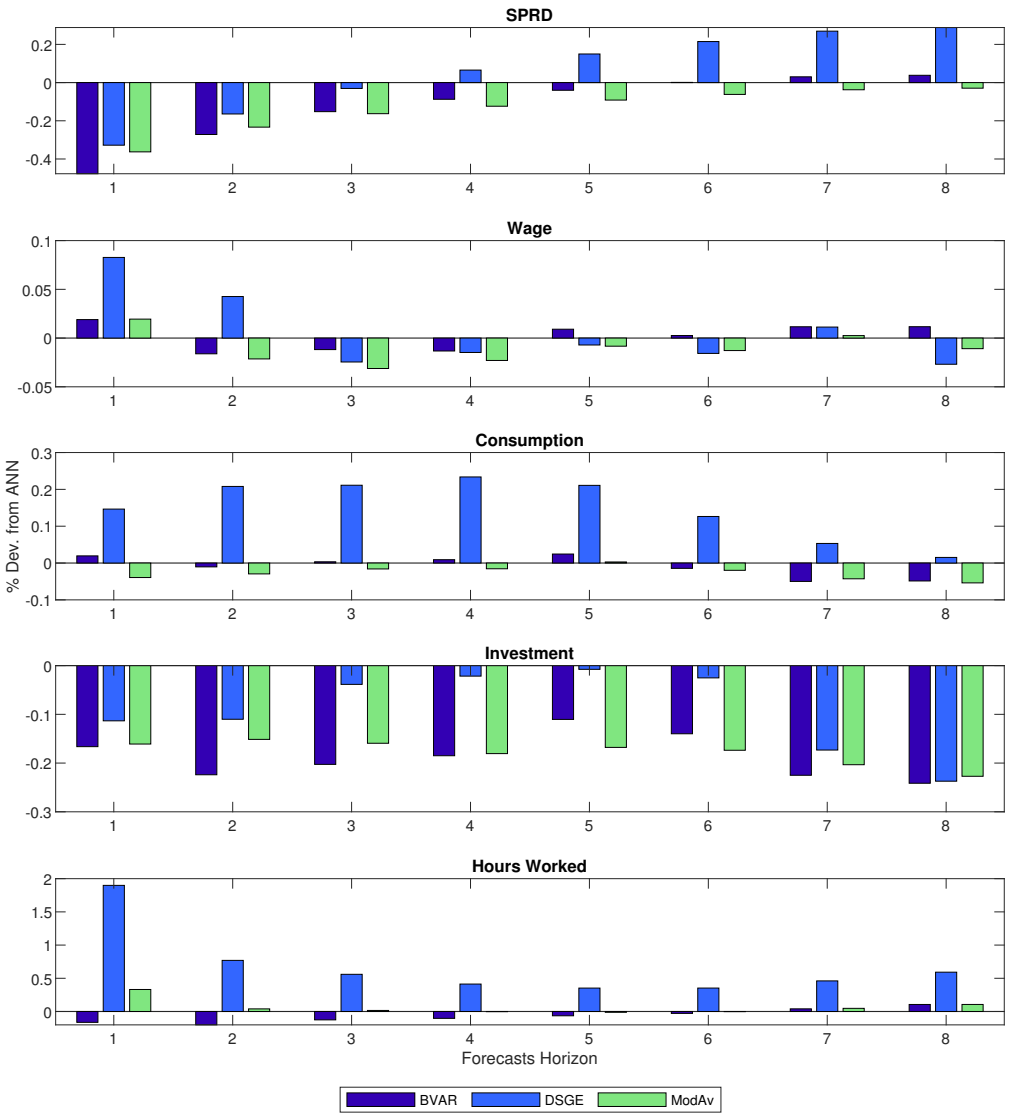
H.2. Forecasting Results by Crisis Periods

Besides the average forecasting performance within certain time segments, an important aspect is a model's ability to provide exact forecasts during disruptive periods. Therefore these predictions are extracted, making use of the identification by the *k-means* clustering of Section 4.2. The main crises are the second golf war around 1990, the 2001 recession which followed the burst of the dot com bubble and the terror attacks on September 11th, and the financial crisis which lasted roughly from 2007 to 2009. For the evaluation, all forecasts for the identified crises periods are taken into account. The results are given in Table 6.

First, as expected, the RMSFEs are larger than those in Table 3 which takes crisis and *normal* times into account. Considering the performance over all variables, the BVAR maintains the leading position in the short-term ($h = 1, 2$) while it drives the even lower RMSFE of the ModAv for $h = 3$ and 4. For the following horizons $h = 5 : 8$, however, the ANN produces lower RMSFEs which is captured by the ModAv (see Table 14 and 15 for SPA test results supporting these interpretations). This is remarkable, as the comparison of results from section 3 and the crisis evaluation discloses the ANN as a robust (over variables) medium to long-term crisis prediction tool. A similar observation can be made for the reduced set of variables as the ModAv produces the lowest RMSFEs for $h = 2 : 8$, which together with the large SPA-test p-values underlines the importance of ANN-based predictions already for the nearer periods ($h = 3 : 8$) in crisis times.

Crisis 2001. Based on Figure 15, one can take a closer look at forecasts during the 2001 recession. The figure shows six plots, each belonging to forecasts made in the vintage stated in the title (2000Q3 to 2001Q4), with one- up to eight-quarter ahead forecasts. The actual (revised) data is plotted in black. Starting with the GDP forecast in 2000Q3, one can see that neither model is able to capture the path of the data (which is especially difficult for the onset of a crisis), but the DSGE short-term prediction is closest to the actual, as is the ANN-based medium-term prediction. Going a step further (2001Q1), the DSGE can somehow trace the drop in GDP growth around 2001Q3, but the recovery thereafter is better forecasted by BVAR and ANN. Based on vintage 2001Q3, the recovery path is predicted more closely by the ANN. In these graphs, the nonlinear (and thereby also more volatile character) of the ANN as a model becomes clear. The inflation forecasts for the one-step ahead prediction are very precise based on 2000Q3 and

Figure 14: Relative RMSFE Section 3



Note: This figure shows RMSFEs of individual variables as percentage deviation from the ANN over all forecast horizons.

Q4, the long-term however is not accurately predicted. Also in 2001Q1, the ANN-based one-step forecast performs very well. Only during the onset of the recovery (2001Q3), the models provide medium- and long-term forecasts mapping the actual inflation path better. Figure 15 shows that also for this variable all models have difficulty predicting the onset of the crisis accompanied by a sharp drop in interest rates. The ANN is the only model including a drop in FFR in its medium to long-term forecasts which,

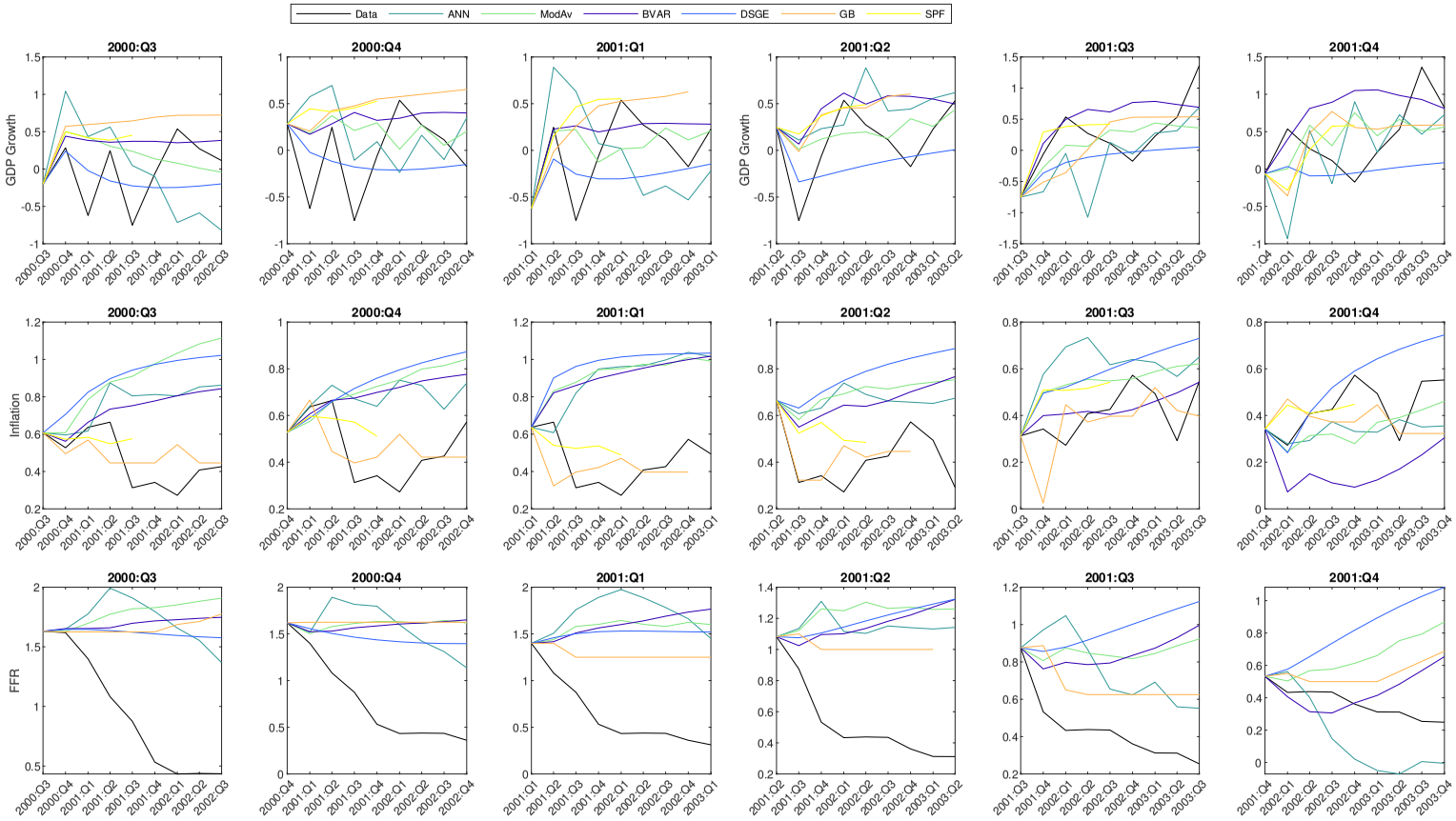
Table 6: RMSFE Crises Periods

Model	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=7$	$h=8$
av. all Vs.								
ANN	0.75	0.92	0.90	0.93	0.95	0.95	1.03	1.06
ModAv	0.75	0.84	**0.86**	**0.88**	**0.92**	**0.93**	**1.00**	**1.03**
BVAR	**0.69**	**0.79**	0.86	0.89	0.97	0.98	1.01	1.06
DSGE	0.99	1.00	1.07	1.07	1.11	1.13	1.16	1.18
av. red. Vs.								
ANN	0.40	0.54	0.54	0.58	0.62	0.63	0.70	0.69
ModAv	0.44	**0.47**	**0.50**	**0.54**	**0.59**	**0.60**	**0.67**	**0.68**
BVAR	**0.39**	0.48	0.53	0.57	0.64	0.65	0.71	0.73
DSGE	0.75	0.62	0.61	0.63	0.66	0.68	0.75	0.76
GB	0.42	0.48	0.49	0.49	0.50	0.49	NaN	NaN
SPF	0.44	0.48	0.51	0.50	NaN	NaN	NaN	NaN

Note: The column *av all vars* shows the RMSFE, of the respective subsample, averaged over all 8 variables. The column *av red vars* averages over GDP, inflation and the federal funds rate only. $h = 1, ..., 8$ gives the forecasting horizon. The GB forecasts are available up to 2015. As there is no FFR data from the SPF available, it is averaged over GDP and Inflation forecasts only. The maximum forecast horizon of the SPF is 4.

however, appears later than actual. Only in 2001Q3, the ANN and Greenbook forecasts capture the tendency of actual FFR paths again.

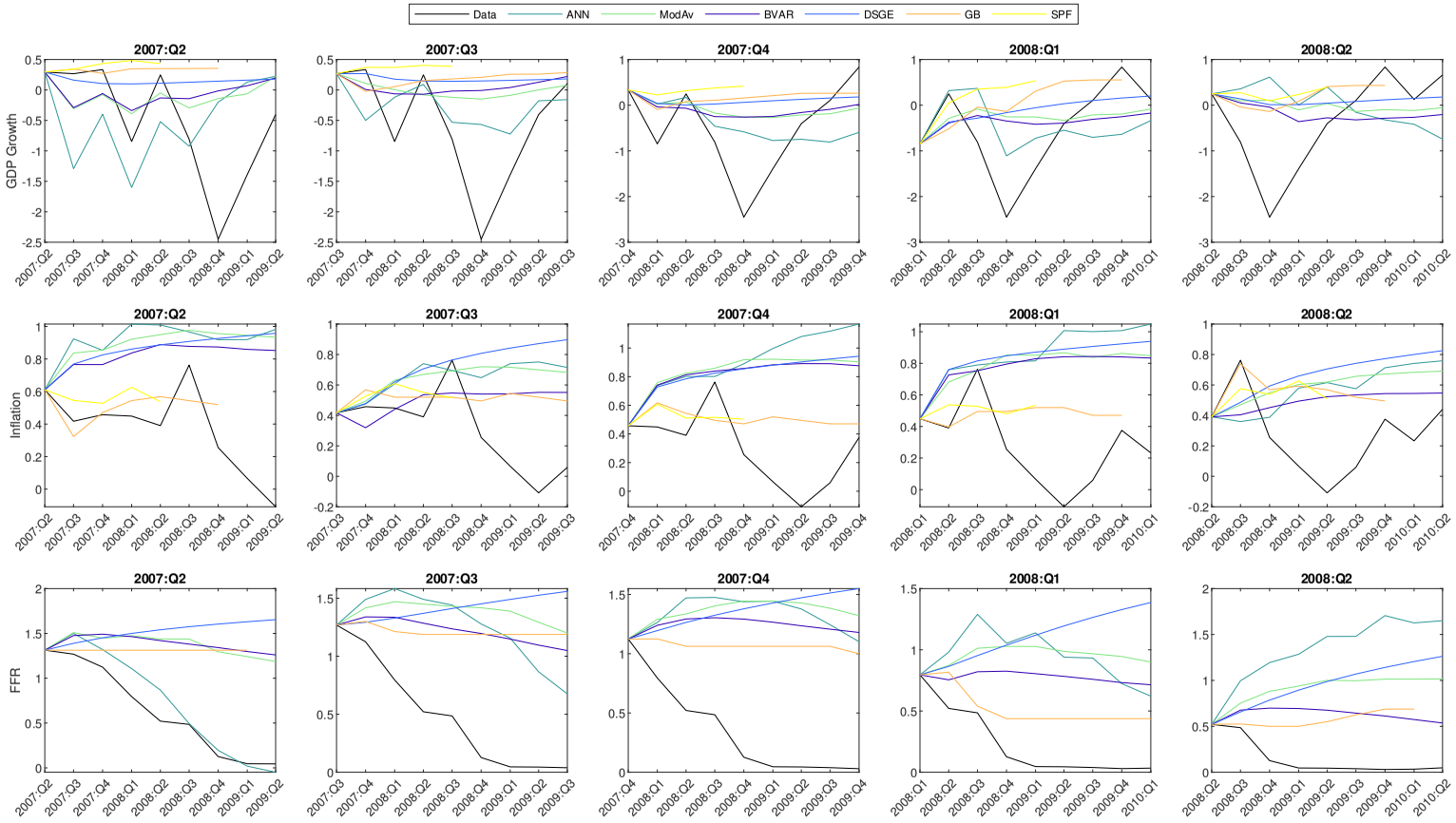
Figure 15: Forecasts 2001 Crisis



Note: This figure plots the forecasted values of each individual model as well as official forecasts against the actual data. The headlines refer to the point of time of the forecast, for which the current data plus forecasts for $h = 1 : 8$ are plotted.

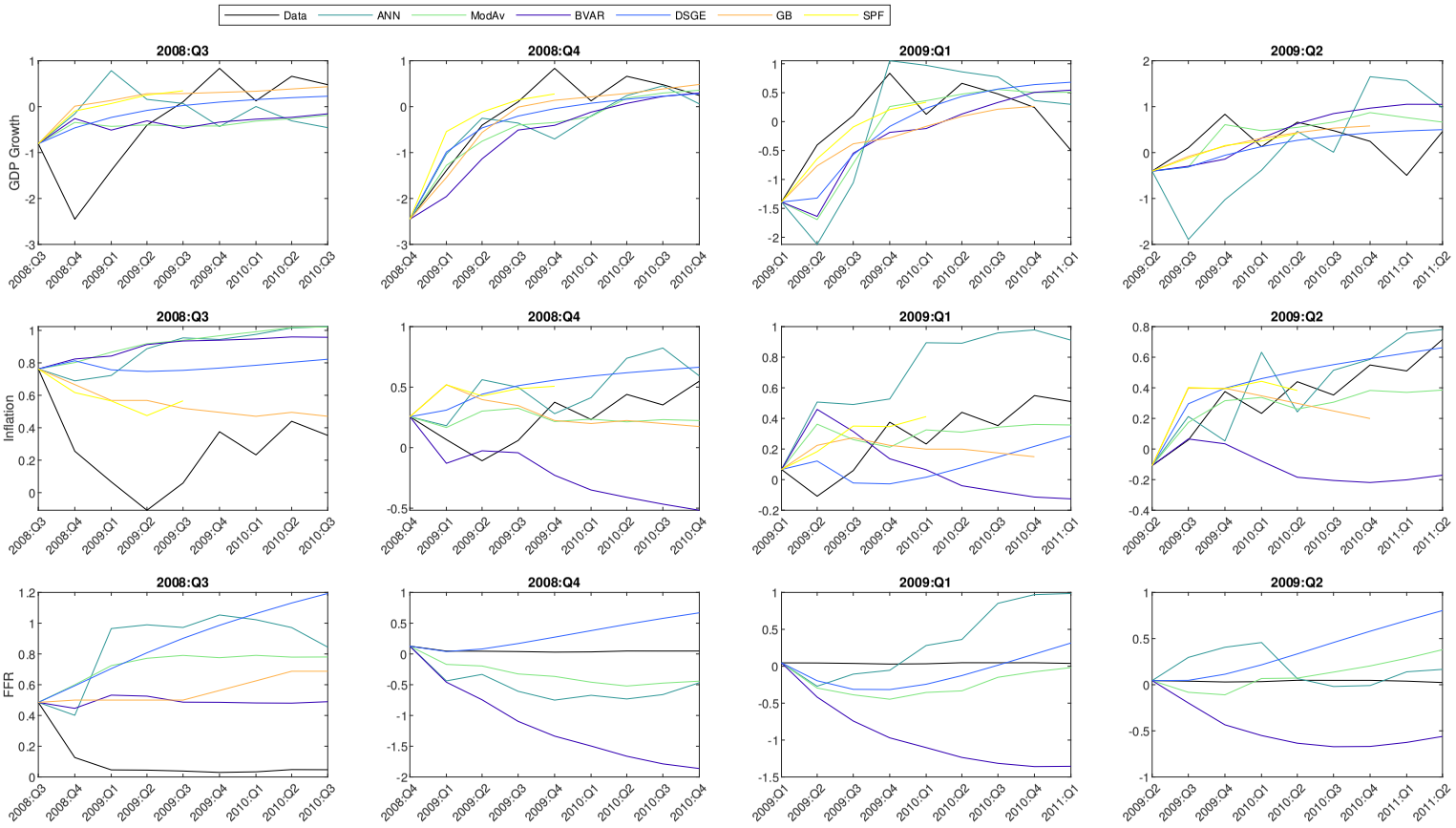
Financial Crisis. During the financial crisis, some more differences between the models become visible. It seems that again, the recovery can be more precisely forecasted than the downturn. Figures 16 and 17 show GDP forecasts based on vintages 2007Q2 to 2009Q2. While no model can forecast the downturn in 2007Q2, the ANN-predicted GDP growth path mimics the actual path remarkably well. One quarter ahead, in 2007Q3, it is the only model (and far from the official forecasts) capturing the drop in 2008Q3/Q4. Forecasts in 2008Q4 can finally capture the downturn in their short-term predictions while also the recovery path is predicted pretty accurately by all models. There is a similar picture regarding inflation forecasts during the onset of the crisis. It improves with vintage 2008Q4, where the DSGE and ANN are able to capture the upward trend, while the BVAR forecasts negative values in the long run. Based on 2009Q2, the predictions by the DSGE and ANN become precise again over all horizons, while the BVAR continues with too low inflation forecasts. Inflation forecasts fail to capture the true path in 2007Q4 and get better one quarter after. The official forecasts are, however, much more precise until 2008Q1. As with GDP growth, the recovery is more precisely predicted starting with forecasts from 2008Q4. The FFR predictions by the ANN from 2007Q2 are extraordinarily close to the actual path. Subsequently, neither model is able to capture well the decrease to the zero lower bound. It takes until 2008Q4 that the models sense the downturn and now undershot with their medium- to long-term predictions. Based on the information set in 2009Q2, at least the ANN shows a fluctuation around the ZLB.

Figure 16: Forecasts during Financial Crisis



Note: This figure plots the forecasted values of each individual model as well as official forecasts against the actual data. The headlines refer to the point of time of the forecast, for which the current data plus forecasts for $h = 1 : 8$ are plotted.

Figure 17: Forecasts during Financial Crisis (continued)



Note: This figure plots the forecasted values of each individual model as well as official forecasts against the actual data. The headlines refer to the point of time of the forecast, for which the current data plus forecasts for $h = 1 : 8$ are plotted.

The crises analysis provides novel insights and evidence that the ANN-based forecasts during recessions improve upon the other models during the medium and long run (for all variables) and even in the nearer term for the reduced variable set. Hence, ANN can be categorized as a robust tool especially for forecasts during disruptive times. Similar results are found by (Medeiros et al., 2021) who use NBER classifications of recessions versus expansions to test whether the superiority of the random forest model varies. The authors find gains of the random forest, which are particularly large during and after the great recession. These results can be confirmed for the ANN through the crisis analysis on the one hand, and the increased superiority of the ANN during section 1 on the other hand. Another crisis analysis is conducted by Wieland and Wolters (2011) who compare various DSGE models with professional forecasters with respect to their predictive power during recessions. First, the authors' finding that DSGE models compare better to professional forecasts during medium-term crises predictions ($h = 2:3$) can be supported, however BVAR and ANN have more predictive power. Furthermore, the detailed analyses of the 2001 and the financial crisis reveal particularly good predictions of the recovery paths by all models, which matches the finding by Wieland and Wolters (2011) and by Del Negro et al. (2015). However, following Binder et al. (2021), this might be a characteristic of the expanding window analysis in comparison to a rolling window framework, which produces better downturn predictions.

I. Statistical Tests

I.1. Uni- versus Multivariate Forecasts

Since in the forecasting literature, there exists a variety of univariate forecasting methods opposed to multivariate approaches, the question arises which information set provides a good basis for exact predictions. Peña and Sánchez (2007) develop a simple test to measure the advantages of the multivariate setup in advance of building the model itself[28]. The authors apply their method to linear models (ARMA versus VARMA); improvements by using the joint dynamics can be expected in case that the timeseries are related.

$$P_{M|U}(h) = 1 - \frac{\sigma_M^2(h)}{\sigma_U^2(h)} \tag{13}$$

is the measure for the expected decrease in the mean squared forecast error of the multivariate model compared to the univariate one ($\sigma_U^2(h) = \beta_h' V \beta_h$ and $V = E(\epsilon_t \epsilon_t')$)[29]. Given a set of observations, the forecast errors are estimated by OLS of an AR(k) model and $P_{M|U}(h)$ can be calculated. A measure of the predictability is given by

$$\hat{F}_{M|U}(h) = 1 - \frac{FPE_M(h)}{FPE_U(h)} \tag{14}$$

where FPE is the final prediction error criterion (Akaike, 1970). The authors suggest, that an analyst can use $\hat{P}_{M|U}(h)$ as a potential benchmark which, in case of inefficient modeling, has a lower limit of $\hat{F}_{M|U}(h)$. As additional information, $\hat{P}_U(h)$ and $\hat{P}_M(h)$ measure the predictability of the series for different forecast horizons, i.e. the decrease in MSFE by using the univariate (U)/multivariate (M) model with respect to using the unconditional mean.

Test results for the analyzed dataset are given in Table 16 in the Appendix. Following the division into several sections, three subsamples are used which are equivalent to the first sample of each analyzed section (1964:1987, 1964:1999, 1964:2010). Evidence is provided, that most variables profit from a multivariate forecasting setup, irrespective of the subsample. Going into detail, GDP benefits the most in the short and medium term, where a multivariate setup improves predictions by up to 33%. For inflation, the advantages increase with the forecasting horizon (38% for h=8 in subsample 1). With respect to FFR, the benefits of a multivariate setup increase over time and horizons (largest in 2010, h=8 with up to 24%). Also spread, investment and the hours worked profit a lot from multivariate approaches, while wage and consumption show less improvements.

These findings justify the focus on multivariate models, nevertheless, also easy univariate benchmarks as a simple random walk process (RW) and an autoregressive model (AR(1:4)) are computed. While these models did not show bad results for some vari-

[28]These codes are publicly available on the authors' website.

[29]This equation shows that when the model is known, multivariate forecasts cannot be less precise than univariate ones. However, when the parameters are estimated this may not be the case.

ables, they could not keep up with the variable robustness of the other multivariate models. Similar results are found by (Medeiros et al., 2021), where RW and AR models are consistently beaten. As this project's goal is to compare models with respect to their forecasts from a *modeling* perspective (i.e. being robust across multiple variables), the simple univariate benchmarks are neglected. Furthermore, while the test deals with linear methods only, it is supposed that nonlinear models can profit even more from the enlarged information set. Thus, the multivariate contemplation can be interpreted as one driver of the superiority of the ANN as the joint dynamics can be exploited best by the network structure (Chang et al., 2018).

I.2. Linearity Tests

To further investigate the interrelation of the analyzed timeseries, multivariate and univariate linearity tests are conducted, again for three subsamples (1964:1987, 1964:1999, 1964:2010). The tests taken into account are Tsay (1986), as well as Teräsvirta et al. (1993)[30]. The H_0 hypothesis of linearity can be rejected at the 1% significance level for the multivariate inspection in any subsample (see Table 7). Examining every variable individually as univariate processes, the null hypothesis is mostly rejected for inflation, FFR and the spread. GDP, wage, investment and hours contain only some degree of nonlinearity, while consumption appears to be linear in its univariate form. This information sheds light on the heterogeneity of the forecasting results in between variables, and why simple linear models perform well for some variables and complex ones better for others, as in Marcellino (2004). Since especially the multivariate test proves nonlinearity of the joint dynamics of the dataset, this analysis provides further evidence that allowing for nonlinearities is key to improving macroeconomic forecasts (a similar conclusion is drawn by Medeiros et al. (2021)).

It should be mentioned, however, that besides multiple nonlinear data-driven approaches, there are also nonlinear theoretical models which might yield better forecasts, specifically of crises situations (Del Negro and Schorfheide, 2013). Nevertheless, these approaches are quite complicated and require even more processing power than the linearized versions[31]. Due to this fact, this research project concentrates on a linearized DSGE model.

[30]Codes are used and made publicly available by Mohammadi (2020).

[31]The required processing power within this project was the largest for the DSGE model which took approximately 45 minutes for the estimation of one vintage. The BVAR was faster with about one minute processing time and the neural network was the fastest with less than one minute per vintage.

Table 7: Multivariate and Univariate Nonlinearity Tests

	1987		1999		2010	
	Teräsvirta	Tsay	Teräsvirta	Tsay	Teräsvirta	Tsay
All Variables	0.000***	0.000***	0.000***	0.000***	0.000***	0.000***
GDP	0.135	0.717	0.000***	0.324	0.623	0.827
Inflation	0.083*	0.763	0.017**	0.003***	0.000***	0.000***
FFR	0.000***	0.000***	0.000***	0.000***	0.000***	0.000***
Spread	0.001***	0.023**	0.001***	0.023**	0.001***	0.023**
Wage	0.200	0.454	0.053*	0.192	0.003***	0.019**
Consumption	0.373	0.488	0.271	0.540	0.184	0.584
Investment	0.012**	0.159	0.013**	0.114	0.013**	0.182
Hours	0.086*	0.190	0.086*	0.190	0.086*	0.190

Note: P-Values for linearity tests by Teräsvirta et al. (1993) and Tsay (1986) are shown. Stars indicate significance levels (*** = 1%, ** = 5%, * = 10%). H_0: The time series is/are linear. *All variables* shows the p-values for the multivariate nonlinearity test.

I.3. Test for Superior Predictive Accuracy

Table 8: SPA with ANN (Section 1)

		h=1	h=2	h=3	h=4	h=5	h=6	h=7	h=8
Av. all V.	**red. Models**								
	ANN	0.165	0.007	1.000	1.000	1.000	1.000	1.000	1.000
	BVAR	1.000	1.000	0.140	0.088	0.004	0.011	0.003	0.002
	DSGE	0.002	0.007	0.001	0.002	0.000	0.001	0.000	0.010
Av. red. V.	**all Models**								
	ANN	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	BVAR	0.020	0.444	0.043	0.172	0.107	0.227	0.407	0.264
	DSGE	0.004	0.070	0.000	0.000	0.000	0.000	0.000	0.000
	GB	0.000	0.000	0.000	0.005	0.001	0.071	0.155	0.008
	SPF	0.004	0.167	0.280	0.349	NaN	NaN	NaN	NaN
	red. Models								
	ANN	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	BVAR	0.018	0.422	0.030	0.119	0.100	0.145	0.275	0.181
	DSGE	0.002	0.067	0.000	0.000	0.000	0.000	0.000	0.000
GDP	**all Models**								
	ANN	0.614	0.155	0.049	0.031	0.381	0.384	0.003	0.003
	BVAR	0.115	0.210	0.085	0.100	0.219	0.151	0.377	0.108
	DSGE	0.519	1.000	0.293	0.281	1.000	1.000	1.000	1.000
	GB	0.151	0.164	0.006	0.188	0.026	0.168	0.022	0.000
	SPF	1.000	0.645	1.000	1.000	NaN	NaN	NaN	NaN
	red. Models								
	ANN	1.000	0.254	0.028	0.013	0.344	0.298	0.004	0.003
	BVAR	0.080	0.221	0.066	0.064	0.169	0.097	0.311	0.112
	DSGE	0.523	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Infl	**all Models**								
	ANN	0.599	1.000	1.000	1.000	1.000	1.000	0.049	0.241
	BVAR	0.001	0.007	0.000	0.023	0.001	0.051	0.007	0.000
	DSGE	0.271	0.043	0.003	0.002	0.000	0.004	0.001	0.002
	GB	1.000	0.002	0.001	0.001	0.040	0.227	1.000	1.000
	SPF	0.511	0.400	0.209	0.210	NaN	NaN	NaN	NaN
	red. Models								
	ANN	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	BVAR	0.001	0.007	0.000	0.017	0.001	0.037	0.004	0.000
	DSGE	0.186	0.070	0.010	0.013	0.000	0.002	0.002	0.001
FFR	**red. Models**								
	ANN	0.002	0.001	0.013	0.116	0.275	0.390	1.000	1.000
	BVAR	1.000	1.000	1.000	1.000	1.000	1.000	0.388	0.279
	DSGE	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Note: This table shows results for the test for superior predictive accuracy by Hansen (2005). Either ANN or ModAv is included in the set of models. Forecasts for *Av. red. V.* (GDP, inflation, FFR) are tested; *Av. all V.* is the test for the average over all 8 variables. The **red. Models** setup contains ANN/ModAv, BVAR and DSGE, the **all Models** specification also contains Greenbook and SPF forecasts. The model stated per line is treated as the Benchmark, which is tested against the alternatives (red. Models or all Models). Small p-values mean that one can reject the null hypothesis of SPA; large p-values are in favor of the respective benchmark model.

Table 9: SPA with ModAv (Section 1)

		h=1	h=2	h=3	h=4	h=5	h=6	h=7	h=8
Av. all V.	**red. Models**								
	ModAv	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	BVAR	0.188	0.096	0.079	0.001	0.025	0.031	0.056	0.028
	DSGE	0.000	0.000	0.000	0.000	0.002	0.002	0.002	0.020
Av. red. V.	**all Models**								
	ModAv	1.000	1.000	0.504	0.442	0.074	0.054	0.023	0.026
	BVAR	0.070	0.255	0.290	0.371	1.000	1.000	1.000	1.000
	DSGE	0.000	0.019	0.000	0.000	0.000	0.000	0.000	0.000
	GB	0.004	0.000	0.003	0.030	0.007	0.199	0.241	0.129
	SPF	0.001	0.022	1.000	1.000	NaN	NaN	NaN	NaN
	red. Models								
	ModAv	1.000	1.000	1.000	0.389	0.048	0.032	0.011	0.017
	BVAR	0.058	0.244	0.278	1.000	1.000	1.000	1.000	1.000
	DSGE	0.000	0.011	0.000	0.000	0.000	0.000	0.000	0.000
GDP	**all Models**								
	ModAv	0.532	0.705	0.119	0.249	0.356	0.458	1.000	0.376
	BVAR	0.121	0.136	0.091	0.043	0.051	0.041	0.336	0.052
	DSGE	0.547	1.000	0.299	0.367	1.000	1.000	0.530	1.000
	GB	0.190	0.158	0.005	0.188	0.007	0.156	0.024	0.000
	SPF	1.000	0.670	1.000	1.000	NaN	NaN	NaN	NaN
	red. Models								
	ModAv	1.000	0.612	0.078	0.126	0.361	0.257	1.000	0.370
	BVAR	0.070	0.104	0.066	0.026	0.039	0.023	0.290	0.052
	DSGE	0.544	1.000	1.000	1.000	1.000	1.000	0.514	1.000
Infl	**all Models**								
	ModAv	0.557	0.085	0.010	1.000	1.000	0.222	0.011	0.032
	BVAR	0.001	0.002	0.000	0.023	0.026	0.218	0.005	0.004
	DSGE	0.268	0.045	0.004	0.002	0.000	0.000	0.000	0.000
	GB	1.000	0.001	0.001	0.021	0.480	1.000	1.000	1.000
	SPF	0.510	1.000	1.000	0.298	NaN	NaN	NaN	NaN
	red. Models								
	ModAv	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	BVAR	0.000	0.002	0.004	0.021	0.019	0.366	0.331	0.262
	DSGE	0.172	0.044	0.009	0.001	0.000	0.000	0.000	0.000
FFR	**red. Models**								
	NAR	0.001	0.000	0.000	0.000	0.001	0.001	0.001	0.001
	BVAR	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	DSGE	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Note: This table shows results for the test for superior predictive accuracy by Hansen (2005). Either ANN or ModAv is included in the set of models. Forecasts for *Av. red. V.* (GDP, inflation, FFR) are tested; *Av. all V.* is the test for the average over all 8 variables. The **red. Models** setup contains ANN/ModAv, BVAR and DSGE, the **all Models** specification also contains Greenbook and SPF forecasts. The model stated per line is treated as the Benchmark, which is tested against the alternatives (red. Models or all Models). Small p-values mean that one can reject the null hypothesis of SPA; large p-values are in favor of the respective benchmark model.

Table 10: SPA with ANN (Section 2)

		h=1	h=2	h=3	h=4	h=5	h=6	h=7	h=8
Av. all V.	**red. Models**								
	ANN	0.096	0.031	0.105	0.134	1.000	1.000	1.000	0.538
	BVAR	1.000	1.000	1.000	1.000	0.508	0.186	0.641	1.000
	DSGE	0.678	0.185	0.273	0.245	0.250	0.101	0.083	0.144
Av. red. V.	**all Models**								
	ANN	0.576	0.069	1.000	1.000	1.000	1.000	1.000	1.000
	BVAR	0.559	0.724	0.666	0.540	0.231	0.195	0.258	0.033
	DSGE	1.000	0.268	0.061	0.019	0.008	0.003	0.014	0.002
	GB	0.225	1.000	0.707	0.731	0.363	0.212	0.079	0.000
	SPF	0.011	0.353	0.329	0.628	NaN	NaN	NaN	NaN
	red. Models								
	ANN	0.539	0.045	1.000	1.000	1.000	1.000	1.000	1.000
	BVAR	0.494	1.000	0.510	0.319	0.177	0.146	0.196	0.023
	DSGE	1.000	0.300	0.088	0.016	0.006	0.003	0.011	0.002
GDP	**all Models**								
	ANN	0.075	0.002	1.000	1.000	1.000	1.000	0.651	0.464
	BVAR	0.119	0.344	0.073	0.138	0.119	0.200	0.649	0.754
	DSGE	1.000	0.693	0.626	0.529	0.275	0.123	1.000	1.000
	GB	0.224	0.534	0.070	0.122	0.053	0.026	0.150	0.000
	SPF	0.266	1.000	0.279	0.302	NaN	NaN	NaN	NaN
	red. Models								
	ANN	0.041	0.026	1.000	1.000	1.000	1.000	0.546	0.472
	BVAR	0.086	0.325	0.055	0.103	0.096	0.152	0.591	0.760
	DSGE	1.000	1.000	0.420	0.300	0.267	0.084	1.000	1.000
Infl	**all Models**								
	ANN	0.252	0.086	0.069	0.079	0.259	0.177	0.072	0.044
	BVAR	0.022	0.001	0.001	0.031	0.119	0.086	0.028	0.014
	DSGE	0.325	0.007	0.001	0.021	0.116	0.078	0.031	0.011
	GB	0.191	0.013	0.004	0.005	1.000	1.000	1.000	1.000
	SPF	1.000	1.000	1.000	1.000	NaN	NaN	NaN	NaN
	red. Models								
	ANN	0.378	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	BVAR	0.008	0.367	0.275	0.024	0.157	0.138	0.066	0.045
	DSGE	1.000	0.343	0.120	0.432	0.172	0.349	0.154	0.075
FFR	**all Models**								
	ANN	0.045	0.016	0.041	0.161	0.478	1.000	1.000	1.000
	BVAR	1.000	0.193	0.283	0.766	1.000	0.450	0.198	0.035
	DSGE	0.000	0.000	0.000	0.000	0.004	0.007	0.016	0.007
	GB	0.489	1.000	1.000	1.000	0.422	0.192	0.039	0.000
	red. Models								
	ANN	0.031	0.014	0.034	0.133	0.324	1.000	1.000	1.000
	BVAR	1.000	1.000	1.000	1.000	1.000	0.285	0.119	0.025
	DSGE	0.000	0.001	0.002	0.003	0.003	0.006	0.009	0.003

Note: This table shows results for the test for superior predictive accuracy by Hansen (2005). Either ANN or ModAv is included in the set of models. Forecasts for *Av. red. V.* (GDP, inflation, FFR) are tested; *Av. all V.* is the test for the average over all 8 variables. The **red. Models** setup contains ANN/ModAv, BVAR and DSGE, the **all Models** specification also contains Greenbook and SPF forecasts. The model stated per line is treated as the Benchmark, which is tested against the alternatives (red. Models or all Models). Small p-values mean that one can reject the null hypothesis of SPA; large p-values are in favor of the respective benchmark model.

Table 11: SPA with ModAv (Section 2)

		h=1	h=2	h=3	h=4	h=5	h=6	h=7	h=8
Av. all V.	**red. Models**								
	ModAv	1.000	0.318	1.000	1.000	1.000	1.000	1.000	1.000
	BVAR	0.396	1.000	0.213	0.039	0.002	0.001	0.014	0.047
	DSGE	0.263	0.322	0.045	0.046	0.036	0.017	0.003	0.000
Av. red. V.	**all Models**								
	ModAv	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	BVAR	0.053	0.409	0.223	0.211	0.163	0.194	0.263	0.147
	DSGE	0.374	0.091	0.000	0.000	0.000	0.000	0.000	0.000
	GB	0.060	0.589	0.415	0.478	0.354	0.372	0.292	0.024
	SPF	0.001	0.071	0.220	0.458	NaN	NaN	NaN	NaN
	red. Models								
	ModAv	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	BVAR	0.037	0.267	0.138	0.122	0.126	0.144	0.196	0.095
	DSGE	0.263	0.060	0.000	0.000	0.000	0.000	0.000	0.000
GDP	**all Models**								
	ModAv	0.321	0.919	1.000	1.000	1.000	1.000	1.000	1.000
	BVAR	0.079	0.174	0.040	0.020	0.058	0.047	0.192	0.005
	DSGE	1.000	0.717	0.626	0.280	0.172	0.075	0.433	0.204
	GB	0.199	0.545	0.010	0.031	0.006	0.003	0.056	0.000
	SPF	0.232	1.000	0.243	0.221	NaN	NaN	NaN	NaN
	red. Models								
	ModAv	0.150	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	BVAR	0.052	0.130	0.030	0.019	0.045	0.037	0.166	0.005
	DSGE	1.000	0.551	0.435	0.144	0.170	0.047	0.285	0.210
Infl	**all Models**								
	ModAv	0.495	0.026	0.012	0.069	0.124	0.109	0.027	0.010
	BVAR	0.000	0.001	0.001	0.029	0.060	0.066	0.025	0.013
	DSGE	0.320	0.008	0.001	0.021	0.000	0.037	0.030	0.012
	GB	0.183	0.014	0.004	0.005	1.000	1.000	1.000	1.000
	SPF	1.000	1.000	1.000	1.000	NaN	NaN	NaN	NaN
	red. Models								
	ModAv	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	BVAR	0.000	0.009	0.065	0.028	0.036	0.086	0.058	0.065
	DSGE	0.311	0.066	0.001	0.205	0.000	0.022	0.023	0.034
FFR	**red. Models**								
	ModAv	0.056	0.037	0.075	0.142	0.306	0.594	0.772	1.000
	BVAR	1.000	0.185	0.413	0.654	1.000	1.000	1.000	0.471
	DSGE	0.000	0.000	0.000	0.000	0.001	0.002	0.004	0.003
	GB	0.490	1.000	1.000	1.000	0.404	0.275	0.172	0.111
	red. Models								
	NAR	0.028	0.016	0.033	0.063	0.152	0.328	0.480	1.000
	BVAR	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.368
	DSGE	0.001	0.002	0.002	0.002	0.001	0.001	0.001	0.001

Note: This table shows results for the test for superior predictive accuracy by Hansen (2005). Either ANN or ModAv is included in the set of models. Forecasts for *Av. red. V.* (GDP, inflation, FFR) are tested; *Av. all V.* is the test for the average over all 8 variables. The **red. Models** setup contains ANN/ModAv, BVAR and DSGE, the **all Models** specification also contains Greenbook and SPF forecasts. The model stated per line is treated as the Benchmark, which is tested against the alternatives (red. Models or all Models). Small p-values mean that one can reject the null hypothesis of SPA; large p-values are in favor of the respective benchmark model.

Table 12: SPA with ANN (Section 3)

		h=1	h=2	h=3	h=4	h=5	h=6	h=7	h=8
Av. all V.	**red. Models**								
	ANN	0.005	0.002	0.015	0.018	0.300	0.302	0.133	0.114
	BVAR	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	DSGE	0.053	0.033	0.031	0.025	0.035	0.021	0.017	0.012
Av. red. V.	**all Models**								
	ANN	0.202	0.020	0.402	0.474	0.469	0.555	0.646	1.000
	BVAR	1.000	0.319	0.621	0.578	0.439	0.361	1.000	0.373
	DSGE	0.053	0.031	0.018	0.012	0.008	0.010	0.019	0.025
	GB	0.614	1.000	1.000	1.000	1.000	1.000	0.599	0.016
	SPF	0.004	0.071	0.203	0.274	NaN	NaN	NaN	NaN
	red. Models								
	ANN	0.135	0.008	0.270	0.385	0.521	1.000	0.479	1.000
	BVAR	1.000	1.000	1.000	1.000	1.000	0.371	1.000	0.273
	DSGE	0.052	0.031	0.021	0.014	0.010	0.012	0.018	0.021
GDP	**all Models**								
	ANN	0.245	0.021	1.000	1.000	0.544	0.649	0.085	0.111
	BVAR	0.506	0.442	0.698	0.722	0.577	1.000	1.000	1.000
	DSGE	0.101	0.155	0.176	0.269	1.000	0.482	0.358	0.555
	GB	0.806	0.645	0.706	0.247	0.048	0.108	0.028	0.000
	SPF	1.000	1.000	0.677	0.393	NaN	NaN	NaN	NaN
	red. Models								
	ANN	0.354	0.018	1.000	1.000	0.493	0.598	0.054	0.108
	BVAR	1.000	1.000	0.544	0.577	0.493	1.000	1.000	1.000
	DSGE	0.107	0.180	0.139	0.178	1.000	0.399	0.355	0.551
Infl	**all Models**								
	ANN	0.230	0.021	0.013	0.013	0.014	0.009	0.005	0.005
	BVAR	0.333	0.001	0.000	0.000	0.000	0.000	0.000	0.000
	DSGE	0.054	0.061	0.055	0.059	0.055	0.065	0.041	0.021
	GB	0.522	0.257	0.414	0.463	1.000	1.000	1.000	1.000
	SPF	1.000	1.000	1.000	1.000	NaN	NaN	NaN	NaN
	red. Models								
	ANN	0.662	0.594	1.000	1.000	1.000	1.000	1.000	1.000
	BVAR	1.000	1.000	0.621	0.524	0.528	0.601	0.496	0.330
	DSGE	0.108	0.117	0.111	0.195	0.192	0.264	0.331	0.369
FFR	**all Models**								
	ANN	0.086	0.058	0.092	0.207	0.522	1.000	1.000	1.000
	BVAR	0.757	0.115	0.493	0.721	1.000	0.401	0.225	0.111
	DSGE	0.029	0.014	0.002	0.000	0.001	0.002	0.002	0.004
	GB	1.000	1.000	1.000	1.000	0.449	0.238	0.071	0.001
	red. Models								
	ANN	0.058	0.040	0.063	0.138	0.373	1.000	1.000	1.000
	BVAR	1.000	1.000	1.000	1.000	1.000	0.248	0.141	0.070
	DSGE	0.031	0.021	0.005	0.001	0.001	0.002	0.002	0.003

Note: This table shows results for the test for superior predictive accuracy by Hansen (2005). Either ANN or ModAv is included in the set of models. Forecasts for *Av. red. V.* (GDP, inflation, FFR) are tested; *Av. all V.* is the test for the average over all 8 variables. The **red. Models** setup contains ANN/ModAv, BVAR and DSGE, the **all Models** specification also contains Greenbook and SPF forecasts. The model stated per line is treated as the Benchmark, which is tested against the alternatives (red. Models or all Models). Small p-values mean that one can reject the null hypothesis of SPA; large p-values are in favor of the respective benchmark model.

Table 13: SPA with ModAv (Section 3)

		h=1	h=2	h=3	h=4	h=5	h=6	h=7	h=8
Av. all V.	**red. Models**								
	ModAv	0.229	0.045	0.156	0.296	1.000	1.000	0.381	0.491
	BVAR	1.000	1.000	1.000	1.000	0.365	0.372	1.000	1.000
	DSGE	0.033	0.031	0.016	0.004	0.002	0.001	0.001	0.001
Av. red. V.	**all Models**								
	ModAv	0.179	0.411	0.470	0.753	1.000	0.657	1.000	1.000
	BVAR	1.000	0.498	0.602	0.594	0.401	0.370	0.516	0.260
	DSGE	0.039	0.026	0.012	0.005	0.004	0.003	0.002	0.003
	GB	0.598	1.000	1.000	1.000	0.646	1.000	0.416	0.069
	SPF	0.004	0.064	0.197	0.283	NaN	NaN	NaN	NaN
	red. Models								
	ModAv	0.131	0.294	0.394	1.000	1.000	1.000	1.000	1.000
	BVAR	1.000	1.000	1.000	0.437	0.309	0.292	0.377	0.158
	DSGE	0.039	0.019	0.010	0.005	0.003	0.002	0.002	0.001
GDP	**all Models**								
	ModAv	0.298	0.795	1.000	1.000	1.000	1.000	0.651	1.000
	BVAR	0.468	0.461	0.526	0.208	0.169	0.344	1.000	0.601
	DSGE	0.093	0.116	0.102	0.032	0.229	0.091	0.419	0.427
	GB	0.779	0.649	0.302	0.012	0.005	0.021	0.018	0.000
	SPF	1.000	1.000	0.539	0.133	NaN	NaN	NaN	NaN
	red. Models								
	ModAv	0.437	1.000	1.000	1.000	1.000	1.000	0.649	1.000
	BVAR	1.000	0.409	0.382	0.162	0.141	0.269	1.000	0.607
	DSGE	0.088	0.100	0.087	0.020	0.225	0.061	0.437	0.426
Infl	**all Models**								
	ModAv	0.410	0.022	0.014	0.016	0.039	0.013	0.002	0.000
	BVAR	0.292	0.001	0.000	0.001	0.000	0.000	0.000	0.000
	DSGE	0.051	0.057	0.053	0.052	0.054	0.056	0.038	0.024
	GB	0.512	0.252	0.359	0.519	1.000	1.000	1.000	1.000
	SPF	1.000	1.000	1.000	1.000	NaN	NaN	NaN	NaN
	red. Models								
	ModAv	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	BVAR	0.383	0.472	0.502	0.353	0.148	0.031	0.004	0.002
	DSGE	0.047	0.062	0.058	0.069	0.064	0.130	0.138	0.150
FFR	**red. Models**								
	ModAv	0.021	0.001	0.013	0.046	0.160	0.392	0.608	1.000
	BVAR	0.426	0.124	0.252	0.416	1.000	1.000	1.000	0.597
	DSGE	0.029	0.015	0.002	0.000	0.000	0.001	0.001	0.000
	GB	1.000	1.000	1.000	1.000	0.449	0.392	0.315	0.163
	red. Models								
	NAR	0.024	0.007	0.005	0.022	0.083	0.227	0.383	1.000
	BVAR	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.408
	DSGE	0.028	0.020	0.007	0.001	0.000	0.001	0.000	0.000

Note: This table shows results for the test for superior predictive accuracy by Hansen (2005). Either ANN or ModAv is included in the set of models. Forecasts for *Av. red. V.* (GDP, inflation, FFR) are tested; *Av. all V.* is the test for the average over all 8 variables. The **red. Models** setup contains ANN/ModAv, BVAR and DSGE, the **all Models** specification also contains Greenbook and SPF forecasts. The model stated per line is treated as the Benchmark, which is tested against the alternatives (red. Models or all Models). Small p-values mean that one can reject the null hypothesis of SPA; large p-values are in favor of the respective benchmark model.

Table 14: SPA with ANN: Crisis Periods

		h=1	h=2	h=3	h=4	h=5	h=6	h=7	h=8
Av. all V.	**red. Models**								
	ANN	0.171	0.028	0.090	0.044	1.000	1.000	0.171	0.161
	BVAR	1.000	1.000	1.000	1.000	0.326	0.155	1.000	1.000
	DSGE	0.043	0.050	0.021	0.013	0.013	0.009	0.014	0.020
Av. red. V.	**all Models**								
	ANN	0.071	0.007	0.101	0.000	0.071	0.196	0.356	1.000
	BVAR	0.026	0.001	0.011	0.083	0.037	0.053	0.360	0.189
	DSGE	0.044	0.026	0.004	0.000	0.000	0.001	0.119	0.113
	GB	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.361
	SPF	0.002	0.033	0.148	0.252	NaN	NaN	NaN	NaN
	red. Models								
	ANN	1.000	0.046	1.000	0.327	1.000	1.000	1.000	1.000
	BVAR	0.417	1.000	0.397	1.000	0.336	0.335	0.519	0.149
	DSGE	0.053	0.071	0.060	0.026	0.156	0.122	0.219	0.095

Note: This table shows results for the test for superior predictive accuracy by Hansen (2005). Either ANN or ModAv is included in the set of models. Forecasts for *Av. red. V.* (GDP, inflation, FFR) are tested; *Av. all V.* is the test for the average over all 8 variables. The **red. Models** setup contains ANN/ModAv, BVAR and DSGE, the **all Models** specification also contains Greenbook and SPF forecasts. The model stated per line is treated as the Benchmark, which is tested against the alternatives (red. Models or all Models). Small p-values mean that one can reject the null hypothesis of SPA; large p-values are in favor of the respective benchmark model.

Table 15: SPA with ModAv: Crisis Periods

		h=1	h=2	h=3	h=4	h=5	h=6	h=7	h=8
Av. all V.	**red. Models**								
	ModAv	0.124	0.149	0.415	0.450	1.000	1.000	1.000	1.000
	BVAR	1.000	1.000	1.000	1.000	0.093	0.110	0.332	0.342
	DSGE	0.039	0.042	0.005	0.002	0.001	0.000	0.001	0.001
Av. red. V.	**all Models**								
	ModAv	0.068	0.032	0.345	0.158	0.083	0.097	1.000	1.000
	BVAR	0.013	0.003	0.013	0.084	0.033	0.041	0.213	0.109
	DSGE	0.051	0.022	0.003	0.000	0.000	0.000	0.004	0.016
	GB	1.000	1.000	1.000	1.000	1.000	1.000	0.561	0.371
	SPF	0.002	0.030	0.147	0.247	NaN	NaN	NaN	NaN
	red. Models								
	ModAv	0.161	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	BVAR	1.000	0.332	0.059	0.109	0.053	0.059	0.170	0.072
	DSGE	0.048	0.053	0.013	0.000	0.001	0.001	0.003	0.010

Note: This table shows results for the test for superior predictive accuracy by Hansen (2005). Either ANN or ModAv is included in the set of models. Forecasts for *Av. red. V.* (GDP, inflation, FFR) are tested; *Av. all V.* is the test for the average over all 8 variables. The **red. Models** setup contains ANN/ModAv, BVAR and DSGE, the **all Models** specification also contains Greenbook and SPF forecasts. The model stated per line is treated as the Benchmark, which is tested against the alternatives (red. Models or all Models). Small p-values mean that one can reject the null hypothesis of SPA; large p-values are in favor of the respective benchmark model.

Table 16: Testing the Advantage of Multivariate versus Univariate Forecasts

	1987				1999				2010			
	$\hat{P}_U(h)$	$\hat{P}_M(h)$	$\hat{P}_{M\|U}(h)$	$\hat{F}_{M\|U}(h)$	$\hat{P}_U(h)$	$\hat{P}_M(h)$	$\hat{P}_{M\|U}(h)$	$\hat{F}_{M\|U}(h)$	$\hat{P}_U(h)$	$\hat{P}_M(h)$	$\hat{P}_{M\|U}(h)$	$\hat{F}_{M\|U}(h)$
GDP												
h=1	0.062	0.324	0.280	0.188	0.074	0.326	0.273	0.210	0.096	0.395	0.331	0.288
h=4	-0.028	0.131	0.155	0.090	0.001	0.154	0.154	0.092	0.007	0.185	0.180	0.135
h=8	0.040	0.217	0.184	0.091	0.084	0.143	0.064	0.015	0.062	0.148	0.092	0.051
Inflation												
h=1	0.563	0.576	0.029	-0.094	0.810	0.845	0.184	0.109	0.788	0.805	0.076	0.042
h=4	0.362	0.471	0.172	0.071	0.571	0.656	0.197	0.156	0.583	0.623	0.097	0.057
h=8	0.139	0.472	0.386	0.330	0.285	0.503	0.305	0.268	0.331	0.456	0.187	0.155
FFR												
h=1	0.858	0.886	0.199	0.106	0.881	0.907	0.220	0.170	0.921	0.936	0.188	0.153
h=4	0.455	0.495	0.073	0.002	0.496	0.581	0.168	0.125	0.592	0.655	0.154	0.122
h=8	0.113	0.160	0.053	-0.023	0.125	0.264	0.160	0.115	0.200	0.396	0.245	0.215
Spread												
h=1	0.803	0.812	0.045	-0.209	0.783	0.804	0.095	-0.067	0.816	0.838	0.120	0.002
h=4	0.337	0.682	0.521	0.374	0.332	0.617	0.427	0.311	0.306	0.590	0.409	0.328
h=8	0.196	0.640	0.552	0.511	0.197	0.568	0.462	0.417	0.123	0.487	0.415	0.386
Wage												
h=1	-0.018	0.200	0.214	0.113	0.083	0.084	0.001	-0.056	-0.008	0.012	0.019	-0.018
h=4	0.049	0.057	0.008	-0.068	-0.019	0.026	0.044	-0.005	0.011	0.011	0.000	-0.039
h=8	-0.016	0.000	0.016	-0.063	-0.012	0.016	0.028	-0.023	-0.032	-0.011	0.020	-0.018
Cons												
h=1	0.025	0.215	0.195	0.127	0.057	0.231	0.184	0.126	0.101	0.249	0.165	0.115
h=4	-0.021	0.101	0.119	0.023	0.008	0.189	0.182	0.122	0.010	0.193	0.185	0.140
h=8	0.041	0.192	0.157	0.060	0.059	0.147	0.094	0.046	0.042	0.150	0.113	0.078
Inv												
h=1	0.200	0.366	0.207	0.123	0.244	0.378	0.178	0.113	0.322	0.509	0.276	0.225
h=4	-0.029	0.207	0.229	0.144	-0.017	0.262	0.274	0.221	0.012	0.353	0.345	0.320
h=8	0.033	0.274	0.249	0.128	0.037	0.190	0.159	0.070	0.013	0.167	0.156	0.108
Hours												
h=1	0.800	0.846	0.228	0.050	0.813	0.859	0.243	0.108	0.852	0.881	0.196	0.092
h=4	0.105	0.432	0.365	0.227	0.164	0.431	0.319	0.221	0.228	0.371	0.185	0.097
h=8	0.090	0.308	0.239	0.178	0.094	0.261	0.184	0.141	0.160	0.229	0.083	0.048

Note: This table shows results for the a-priori advantage of using a multivariate approach to forecasting (Peña and Sánchez, 2007).

IMFS WORKING PAPER SERIES

Recent Issues

204 / 2024	Alina Tänzer	The Effectiveness of Central Bank Purchases of long-term Treasury Securities:A Neural Network Approach
203 / 2024	Gerhard Rösl	A present value concept for measuring welfare
202 / 2024	Reimund Mink Karl-Heinz Tödter	Staatsverschuldung und Schuldenbremse
201 / 2024	Balint Tatar Volker Wieland	Taylor Rules and the Inflation Surge: The Case of the Fed
200 / 2024	Athanasios Orphanides	Enhancing resilience with natural growth targeting
199 / 2024	Thomas Jost Reimund Mink	Central Bank Losses and Commercial Bank Profits – Unexpected and Unfair?
198 / 2024	Lion Fischer Marc Steffen Rapp Johannes Zahner	Central banks sowing the seeds for a green financial sector? NGFS membership and market reactions
197 / 2023	Tiziana Assenza Alberto Cardaci Michael Haliassos	Consumption and Account Balances in Crises: Have We Neglected Cognitive Load?
196 / 2023	Tobias Berg Rainer Haselmann Thomas Kick Sebastian Schreiber	Unintended Consequences of QE: Real Estate Prices and Financial Stability
195 / 2023	Johannes Huber Alexander Meyer-Gohde Johanna Saecker	Solving Linear DSGE Models With Structure Preserving Doubling Methods
194 / 2023	Martin Baumgärtner Johannes Zahner	Whatever it takes to understand a central banker – Embedding their words using neural networks
193 / 2023	Alexander Meyer-Gohde	Numerical Stability Analysis of Linear DSGE Models – Backward Errors, Forward Errors and Condition Numbers
192 / 2023	Otmar Issing	On the importance of Central Bank Watchers
191 / 2023	Anh H. Le	Climate Change and Carbon Policy: A Story of Optimal Green Macroprudential and Capital Flow Management

190 / 2023	Athanasios Orphanides	The Forward Guidance Trap
189 / 2023	Alexander Meyer-Gohde Mary Tzaawa-Krenzler	Sticky information and the Taylor principle
188 / 2023	Daniel Stempel Johannes Zahner	Whose Inflation Rates Matter Most? A DSGE Model and Machine Learning Approach to Monetary Policy in the Euro Area
187 / 2023	Alexander Dück Anh H. Le	Transition Risk Uncertainty and Robust Optimal Monetary Policy
186 / 2023	Gerhard Rösl Franz Seitz	Uncertainty, Politics, and Crises: The Case for Cash
185 / 2023	Andrea Gubitz Karl-Heinz Tödter Gerhard Ziebarth	Zum Problem inflationsbedingter Liquiditätsrestriktionen bei der Immobilienfinanzierung
184 / 2023	Moritz Grebe Sinem Kandemir Peter Tillmann	Uncertainty about the War in Ukraine: Measurement and Effects on the German Business Cycle
183 / 2023	Balint Tatar	Has the Reaction Function of the European Central Bank Changed Over Time?
182 / 2023	Alexander Meyer-Gohde	Solving Linear DSGE Models with Bernoulli Iterations
181 / 2023	Brian Fabo Martina Jančoková Elisabeth Kempf Luboš Pástor	Fifty Shades of QE: Robust Evidence
180 / 2023	Alexander Dück Fabio Verona	Monetary policy rules: model uncertainty meets design limits
179 / 2023	Josefine Quast Maik Wolters	The Federal Reserve’s Output Gap: The Unreliability of Real-Time Reliability Tests
178 / 2023	David Finck Peter Tillmann	The Macroeconomic Effects of Global Supply Chain Disruptions
177 / 2022	Gregor Boehl	Ensemble MCMC Sampling for Robust Bayesian Inference
176 / 2022	Michael D. Bauer Carolin Pflueger Adi Sunderam	Perceptions about Monetary Policy
175 / 2022	Alexander Meyer-Gohde Ekaterina Shabalina	Estimation and Forecasting Using Mixed-Frequency DSGE Models
174 / 2022	Alexander Meyer-Gohde Johanna Saecker	Solving linear DSGE models with Newton methods