CIRPÉE

Centre Interuniversitaire sur le Risque,
les Politiques Économiques et l'Emploi

Cahier de recherche/Working Paper **11-04**

A Comparison of Forecasting Procedures for Macroeconomic Series: the Contribution of Structural Break Models

Luc Bauwens
Gary Koop
Dimitris Korobilis
Jeroen V.K. Rombouts

Janvier/January 2011

Bauwens: CORE
Koop: University of Strathclyde
Korobilis: CORE
Rombouts: HEC Montreal, CIRANO, CIRPEE and CORE

This research was supported by the ESRC under grant RES-062-23-2646, and by the contract "Projet d'Actions de Recherche Concertées" 07/12-002 of the "Communauté française de Belgique", granted by the "Académie universitaire Louvain". This text presents research results of the Belgian Program on Interuniversity Poles of Attraction initiated by the Belgian State, Prime Minister's Office, Science Policy Programming.

Abstract:

This paper compares the forecasting performance of different models which have been proposed for forecasting in the presence of structural breaks. These models differ in their treatment of the break process, the parameters defining the model which applies in each regime and the out-of-sample probability of a break occurring. In an extensive empirical evaluation involving many important macroeconomic time series, we demonstrate the presence of structural breaks and their importance for forecasting in the vast majority of cases. However, we find no single forecasting model consistently works best in the presence of structural breaks. In many cases, the formal modeling of the break process is important in achieving good forecast performance. However, there are also many cases where simple, rolling OLS forecasts perform well.

Keywords: Forecasting, change-points, Markov switching, Bayesian inference

JEL Classification: C11, C22, C53

1 Introduction

Structural breaks are commonly found to be present in many macroeconomic and financial time series (e.g. Stock and Watson (1996) and Ang and Bekaert (2002)) and to be one of the major reasons of poor forecasting performance (e.g. Clements and Hendry (1998)). This has led to several papers which work with forecasting methods which are robust to breaks (e.g. Pesaran and Timmermann (2007), Eklund, Kapetanios, and Price (2009) or Clark and McCracken (2009)) or formally model the break process (e.g. Pesaran, Pettenuzzo, and Timmermann (2006), Koop and Potter (2007), Giordani and Kohn (2008), Maheu and Gordon (2008) and D'Agostino, Gambetti, and Giannone (2009)). It is an open empirical question as to which types of methods or models will work best when dealing with the sort of structural change present in many macroeconomic data sets. The purpose of this paper is to shed light on this question. We compare empirically the forecasting performance of existing models that explicitly allow for structural breaks both in the sample period and in the forecast period. Two such models are given in Pesaran, Pettenuzzo, and Timmermann (2006), hereafter PPT, and Koop and Potter (2007), hereafter KP, and these form the main focus of our forecasting evaluation.[1] Conventional time-varying parameter (TVP) models such as that used by D'Agostino, Gambetti, and Giannone (2009) also allow explicitly for structural breaks in-sample and out-of-sample and are also included in our forecasting evaluation. In addition, we include some benchmark forecasting procedures such as recursive and rolling OLS.

Our study is in the spirit of Meese and Geweke (1984), Stock and Watson (1996) and Marcellino, Stock, and Watson (2006) in the sense that we investigate the performance of various forecasting approaches at different forecast horizons in a set of macroeconomic time series using relatively simple forecasting models (i.e. extensions of autoregressive, AR, models). We evaluate forecast performance using a variety of metrics. In addition to a conventional measure based on point forecasts (i.e. root mean squared forecast error, RMSE), we compare the approaches using average predictive likelihoods (APL) which are based on the entire predictive density.

[1] The mixture innovation model of Giordani and Kohn (2008) can also be used to forecast in the presence of structural breaks. The mixture innovation approach can nest several popular models of structural change, including some variants of the models presented in KP.

In this paper we focus on PPT and KP as two representative examples of models which address the issues which arise when forecasting subject to structural breaks. Such forecasting models can differ in three important aspects. First, they can differ in the priors they use for the parameters which define the conditional mean (and possibly the conditional variance) of the dependent variable. PPT uses a hierarchical prior of the sort commonly used in the panel data literature where conditional mean coefficients are all assumed to be drawn from some common distribution. KP uses a hierarchical prior motivated by the state space literature where the conditional mean coefficients in the most recent regime are most relevant when a break occurs. Second, they can differ in the hierarchical prior used for the regime durations. For instance, PPT assume a Geometric distribution for regime duration whereas KP assume a Poisson distribution. Third, they can differ in whether they impose the restriction that a precise number of breaks occurs in a sample of size T or whether the number of in-sample breaks is treated as unknown. The former approach is adopted by PPT, involves an (arguably, see Koop and Potter (2007) and Koop and Potter (2009)) unattractive prior at the end of the sample and requires the calculation of marginal likelihoods. The latter approach is adopted by KP and does not involve these drawbacks.

Of course, it is an empirical matter which of these approaches works well in practice and it is possible that each approach works well in some cases but not others. KP and PPT each illustrate the performance of their approaches with a single time series (and with modeling details calibrated to that particular series). The purpose of this paper is to investigate these and related approaches for a wide variety of macroeconomic series. We select twenty-three of the most important quarterly US macroeconomic time series and compare PPT and KP to a variety of forecasting methods. We find that structural breaks are an important feature of most of the time series we consider. Handling such breaks is shown to be an important issue for forecasting. However, we find that there is no one single method which can be recommended universally. That is, for some series PPT forecasts best, for others KP does, for others simpler methods such as rolling OLS forecasts performs best. We argue that this is empirically-sensible and stress the importance of tailoring forecasting models to the empirical application at hand (as opposed to recommending a single approach as being universally best for all macroeconomic time series).

In Section 2, we compare in a non-technical manner the specifications of the PPT and

KP models that we use in our empirical evaluation. Technical details are provided in appendices. In Section 3, we present the estimation results of applying PPT and KP to the series we analyze, focussing on what breaks we find in the series. In Section 4 we discuss the implementation of our forecasting evaluation and in Section 5 we present our main results. Section 6 contains the results of sensitivity analyses and the last section our conclusions.

2 Models with Structural Breaks

In this section, we present and compare the PPT and KP models. After providing a framework for structural break models (sub-section 2.1), we discuss how the parameters of different regimes are linked (sub-section 2.2), how the break process is modelled (sub-section 2.3), and how the number of breaks is determined (sub-section 2.4).

2.1 A Framework for Structural Break Modelling

A linear regression model framework for discussing structural break models is:

$$y_t = Z_t \beta_{s_t} + \sigma_{s_t} \varepsilon_t, \tag{1}$$

where y_t is the dependent variable, Z_t (with m elements in total) contains lagged dependent variables or lagged exogenous variables available for forecasting y_t, and ε_t is i.i.d. $N(0,1)$.

Equation (1) allows for β_{s_t} and σ_{s_t} to vary over time with $s_t \in \{1, .., K\}$ a random variable indicating which regime applies at time t. The vector β_{s_t} determines the conditional mean of y_t and, thus, we will refer to them as conditional mean coefficients with σ_{s_t} being the volatilities.

Different structural break models vary in the way they model the break process. To simplify the exposition, we will focus here on β_{s_t} and assume $\sigma_{s_t} = \sigma$. But we stress that breaks in volatilities can be modelled in exactly the same manner as breaks in the conditional mean coefficients and in our empirical work we allow for breaks in volatility.[2]

Suppose we are working with a model with $K-1$ breaks which occur at unknown times

[2] Furthermore, we could allow for breaks in volatility to occur independently of breaks in the conditional mean. In this case, s_t is a bivariate discrete random variable with the first element controlling breaks in conditional mean and the second element controlling breaks in volatility.

$\tau_1, .., \tau_{K-1}$. Thus, we can write:

$$y_t = \begin{cases} Z_t\beta_1 + \sigma\varepsilon_t \text{ if } s_t = 1 \text{ (i.e. } t \leq \tau_1), \\ Z_t\beta_2 + \sigma\varepsilon_t \text{ if } s_t = 2 \text{ (i.e. } \tau_1 < t \leq \tau_2), \\ ... \\ Z_t\beta_{K-1} + \sigma\varepsilon_t \text{ if } s_t = K-1 \text{ (i.e. } \tau_{K-2} < t \leq \tau_{K-1}), \\ Z_t\beta_K + \sigma\varepsilon_t \text{ if } s_t = K \text{ (i.e. } \tau_{K-1} < t \leq T). \end{cases} \tag{2}$$

Different structural break models arise through different formulations for β_{s_t} and s_t. From a Bayesian point of view these can be interpreted as hierarchical priors. In the next two sections, we discuss modelling of β_{s_t} and s_t respectively.

2.2 Linking the Conditional Mean Coefficients in Different Regimes

It is possible to allow for β_j for $j = 1, .., K$ to be completely independent of one another (i.e. after a break occurs, pre-break information provides absolutely no information about what likely values for the new conditional mean coefficients are). But, in practice, it is typically desirable to avoid such independence. Even when simply doing an in-sample analysis, structural break models can be over-parameterized and placing more structure on the model can help avoid this problem. That is, if β_j is completely independent of all other regimes, one must estimate it using data only from regime j. With relatively short macroeconomic data sets, possibly high dimensional β_j and possibly multiple structural breaks, it may be hard to obtain precise estimates of β_j. When forecasting subject to structural breaks, an even more serious problem occurs. Suppose a break occurs during the forecast period, and the conditional mean coefficient switches from β_j to β_{j+1}. Forecasting must be done using β_{j+1}. If we assume complete independence of conditional mean coefficients across regimes, then immediately after the break we have no data-based information to estimate β_{j+1}. In a Bayesian forecasting exercise, this means the prior for β_{j+1} will be used to produce forecasts. Given a common desire to use relatively noninformative priors, this could lead to extreme and unreasonable forecasts when a break occurs. This has motivated various models which link β_j and β_{j+1} in some manner.

In this paper, we consider two main approaches which relate to those in PPT and KP, respectively. Appendices provide precise details (including discussion of relevant posterior and predictive simulation algorithms), but the basic idea in PPT is to adopt a link of the form:

$$\beta_j = \beta_0 + u_j$$

for $j = 1, .., K$, where u_j is i.i.d. $N_m(0, B_0)$ or, equivalently, in Bayesian language, a hierarchical prior of the form:

$$\beta_j \sim N_m(\beta_0, B_0), \tag{3}$$

the parameters β_0 and B_0 are assumed unknown and can be estimated from the data. Thus, the conditional mean coefficients in each regime are drawn from a common distribution. This practice is commonly used in panel data models with random effects or in random coefficient models and results from that literature can easily be adapted to show that β_0 and B_0 reflect average values across all regimes. If a break occurs in a forecast period, this means that the new value of the conditional mean coefficients will be drawn from a distribution which reflects the values of the coefficients from all past regimes. This is an empirically sensible approach in environments where breaks occur, but in a recurrent way. It allows, for instance, for the 1950s, 1970s, 1990s and 2000's to be different regimes, but the regime in the 2000s is just as likely to be similar to the 1950s as to more recent regimes.

In contrast, KP adopt a hierarchical prior motivated by the state space literature on TVP models. They specify random walk evolution of coefficients:

$$\beta_j = \beta_{j-1} + u_j$$

where u_j is specified as above, or equivalently,

$$\beta_j | \beta_{j-1} \sim N_m(\beta_{j-1}, B_0). \tag{4}$$

The KP prior is similar to the PPT prior, except that, when a structural break occurs, the conditional mean coefficients are drawn from a distribution centered at β_{j-1}. Thus, it is the most recent regime which has the most influence on conditional mean coefficients in a new regime. This is a common modelling assumption in macroeconomic models such as TVP-VARs and, indeed, the KP model is equivalent to a TVP regression model if $s_t = t$ and, thus, $K = T$.

It is worth noting that with either the PPT prior or the KP prior, it is possible to introduce exogenous explanatory variables into the hierarchical prior (e.g. in the PPT prior we could have $\beta_0 = W_t b_0$ for some lagged variables W_t) although we do not explore this avenue in the present paper.

2.3 Modeling the Break Process

The break process is modelled through $S_T = (s_1, .., s_T)'$ where $s_t \in \{1, 2, .., K\}$ are the regime identifying (or state) variables defined previously. It is possible to use a noninformative prior which does not restrict the timing of the breaks. This is an approach developed in Koop and Potter (2009). However, unless the number of breaks is small, computation is difficult (or infeasible) due to the large number of possible configurations of K breakpoints. Furthermore, when forecasting under structural breaks, it is necessary to forecast the probability that a break occurs during the forecast period and this cannot be done using a noninformative prior for S_T. This has led to an interest in informative hierarchical priors for the break process. The most popular of these is developed in Chib (1998) and adopted by PPT. This begins by assuming a restricted Markov process for S_T:

$$\begin{aligned} \Pr\left(s_t = i | s_{t-1} = i\right) &= p_i \\ \Pr\left(s_t = i + 1 | s_{t-1} = i\right) &= 1 - p_i. \end{aligned} \tag{5}$$

Thus, if regime i holds at time $t-1$, then at time t the process can either remain in regime i (with probability p_i) or a break occurs and the process moves to regime $i+1$ (with probability $1-p_i$).

Equation (5) can be interpreted as a hierarchical prior. Note that the durations of regimes are defined as:

$$d_i = \tau_i - \tau_{i-1}$$

and it can be shown that (5) implies a Geometric prior distribution for d_i. KP argue that this may be restrictive in some situations. For instance, the geometric distribution is decreasing and, thus, this hierarchical prior imposes $p\left(d_i\right) > p\left(d_i + 1\right)$. They suggest the use of the more flexible Poisson distribution for the durations:

$$d_i - 1 \sim Po(\lambda_i) \tag{6}$$

where $Po(\lambda_i)$ denotes the Poisson distribution with mean λ_i. However, in the present paper, in order to maintain a fair degree of computational simplicity and comparability across our forecasting approaches, we implement the KP approach using the Geometric prior implied by (5).[3]

[3] The KP model of this paper is thus to be understood from here on to differ from the model of Koop and Potter (2007) in this aspect.

Either of these two hierarchical priors can be used for forecasting purposes. However, when forecasting with structural breaks, we need to estimate the probability that a break occurs during the forecasting period. In some cases, it can be desirable to include more information on the break process or further restrict the model to ensure parsimony. Thus, we note a few empirically useful extensions of the previous priors. First, it is possible to assume a hierarchical prior for p_i or λ_i such that they are drawn from some common distribution. An extreme limiting case of such an approach would involve setting $\lambda_1 = ... = \lambda_K$ or $p_1 = ... = p_K$. Second, it is possible to allow for either p_i or λ_i to depend on lags of themselves (e.g. the prior for λ_i can depend on λ_{i-1}) or durations of past regimes. Some of these possibilities are investigated in KP, but are not pursued here.

2.4 Choosing the Number of Breaks

Thus far, we have said nothing about choosing $K-1$, the number of breaks. But this raises an important issue. Note that both the Geometric and Poisson duration distributions which arise using (5) or (6) are unbounded distributions. Thus, it is possible that any regime endures beyond the end of the sample. For instance, if the sample runs from $t = 1, .., T$ and the model has three breaks, it is possible that $s_T = 1$ or 2 and, thus, that the third regime has not begun before T. PPT and KP adopt two different ways of dealing with this issue, which we describe in turn.

PPT, following Chib (1998), impose additional prior information beyond (5). Intuitively, we can impose that exactly K regimes occur in sample by adding prior information of the form:

$$\Pr[s_T = K | s_{T-1} = K] = \Pr[s_T = K | s_{T-1} = K-1] = 1. \tag{7}$$

Thus, if the process reaches the final regime before the end of the sample it stays there. But if it has not reached the final regime by period $T-1$, it must switch to the final regime. If K exceeds 2, additional restrictions are required. To express these restrictions in words, consider the case $K = 3$. If, in period $T-1$, we are not already in the third regime, then it must be the case that a regime switch occurs in period T and this must be imposed on the model. Similarly, if, in period $T-2$, we are still in the first regime, then we must impose that regime switches occur in both periods $T-1$ and T, in order to ensure that $K = 3$. Note that, as discussed in Koop and Potter (2009), this can lead to a pile-up of prior probability near the end of the sample, leading to a prior which is quite informative (and, thus, potentially influential) precisely at the time forecasting is being done.

KP simply recommend working with models which allow for breakpoints to occur out-of-sample. Statistically, working with such models poses no difficulties for a Bayesian using a proper prior. Consider the case where regime j occurs entirely out-of-sample. It appears that there is no data to directly estimate β_j. However, Bayesian inference is still possible. If the prior for β_j were independent of the conditional mean coefficients in the other regimes, then its posterior would simply equal its prior. Such an approach would allow for valid statistical inference but could yield poor forecasting results unless strong prior information existed about β_j. However, using hierarchical priors such as (3) or (4) allows for data information from in-sample regimes to spill over into out-of-sample regimes and, thus, the posterior for β_j will contain data information. More importantly, allowing for regimes to occur out-of-sample allows the researcher to estimate the number of regimes in-sample. For instance, if the researcher allows for two breakpoints, but one of these occurs after time T, then (in-sample) this is equivalent to estimating a model with one breakpoint. This means that the researcher can simply select a value for the maximum number of breakpoints to allow for as opposed to doing a search over all possible numbers. By contrast, with the PPT approach, marginal likelihoods are calculated for $K = 1, ..K^{\max}$ and the value with the highest marginal likelihood is selected. This need for calculation of marginal likelihoods increases the computational burden.

3 Breaks in US Macroeconomic Series

We apply the PPT and KP models to twenty-three quarterly series for the USA (listed in Table 1) which are among the most important macroeconomic variables. The sample period is 1959, first quarter, till 2010, second quarter. As indicated in the table, we have transformed most series to growth rates or first differences, and in this we are proceeding as in the literature, see e.g. Stock and Watson (1996). We use AR(q) models in each regime, hence, Z_t in (2) contains an intercept and the first q lags of y_t.

Our previous explanation of the PPT and KP approaches assumed homoskedastic errors. In our empirical implementation, we relax this assumption and allow the error variances to change when the AR coefficients do using the same hierarchical priors as in PPT and KP.

Details about prior densities and posterior evaluation are provided in Appendix A for PPT models and in Appendix B for KP. Further discussion of the prior is given in Section 6.2.

In Table 2 we report the break dates found in the PPT-AR(1) and AR(4) models (called PPT1 and PPT4 hereafter), and similarly KP-AR(1) and AR(4) models (KP1 and KP4), using the complete sample. In Table 3 we report the posterior means of the AR(1) equations for each regime. The reported break dates are medians of posterior distributions and there is some uncertainty (though not much) about these point estimates.

We do not find any break in six series (6, 14, 15, 17, 22, 23) both with PPT and KP (irrespective of the lag order), and in four other series with PPT (series 2, 5, 13, 19), see Table 2. No series has more than two breaks with KP, while only series 21 has three breaks with PPT1. To a large extent, the break numbers and dates are robust with respect to the lag order (1 or 4), keeping in mind that for dates we report posterior medians. This is much less the case with respect to the type of model (PPT and KP). For example, even if three series (4, 8, 9) have a break in the last three years of the sample according to both models, KP detects more breaks of this type than PPT, see series 2, 12, and 16.

Thus there is evidence that macroeconomic series are subject to breaks since about three quarters of our series have at least one break when modeled by structural break models. The next obvious question is how large are the parameter changes when breaks occur and what parameters are affected. Table 3 contains the posterior means of the parameters of the AR(1) equations of each regime for each series, over the full sample period. Focusing on the series with more than one regime (for PPT and KP), we observe that the most sensitive parameter is the variance of the error term. It decreases substantially for some series in the first half of the eighties, see the break in series 1, 7, 16, 18, and 20 with both models, and in series 5 with KP, corresponding to what has been named the great moderation (the decrease is about seventy-five percent on average for these series). The error variance increases quite a lot in 2007 or 2008 for series 4, 8, 9 with PPT, and 2, 8, 9 and 16 with KP (see the last break). These increases correspond to the great recession triggered by a widespread financial crisis. Other cases are the reduction by half of the variance of series 3 in 1993, and the quadrupling for series 7 in 2000. The interest rate series (10 and 11) witness also large changes: a tenfold increase in 1979 corresponds to the beginning of the Volcker period at the Fed, which is followed by a decrease at the next break in 1985.

Table 1: Variables used in forecast evaluation

	Acronym	T	Definition
1	GDPC96	5	Real Gross Domestic Product, 3 Decimal
2	GDPDEF	5	Gross Domestic Product: Implicit Price Deflator
3	PCECC96	5	Real Personal Consumption Expenditures
4	PCECTPI	5	Personal Consumption Expenditures Chain-type Price Index
5	GPDIC96	5	Real Gross Private Domestic Investment, 3 Decimal
6	OPHPBS	5	Business Sector: Output Per Hour of All Persons
7	ULCNFB	5	Nonfarm Business Sector: Unit Labor Cost
8	CPIAUCSL	6	Consumer Price Index for All Urban Consumers: All Items
9	PPIFCG	6	Producer Price Index: Finished Consumer Goods
10	TB3MS	2	3-Month Treasury Bill: Secondary Market Rate
11	GS10	2	10-Year Treasury Constant Maturity Rate
12	M1SL	6	M1 Money Stock
13	M2SL	6	M2 Money Stock
14	UTL11	1	Capacity Utilization: Manufacturing
15	SP500	5	S&P 500 Index
16	INDPRO	5	Industrial Production Index
17	HOUST	4	Housing Starts: New Privately Owned Housing Units Started
18	AHEMAN	5	Average Hourly Earnings: Manufacturing
19	UNRATE	2	Civilian Unemployment Rate
20	PAYEMS	5	Total Nonfarm Payrolls: All Employees
21	EXUSUK	5	U.S. / U.K Foreign Exchange Rate
22	PMI	1	ISM Manufacturing: PMI Composite Index
23	NAPMNOI	1	ISM Manufacturing: New Orders Index

T (transformation applied to original series): 1 = no transformation, 2 = first difference, 4 = log, 5 = first difference of logged variables, 6 = second difference of logged variables. Sample period (after data transformation): 1959Q1-2010Q2 (206 observations). Data source: St. Louis ALFRED database (http://alfred.stlouisfed.org).

In some series, the constant and the AR(1) coefficients change also, but less spectacularly than the variance. This happens to the two interest rates. Keeping in mind that they are in first differences, the changes of the coefficients (in particular the sign change of the constant) around 1985 correspond to the start of a long period of decrease of interest rates. A change of sign of the constant happens also in series 21 in the last quarter of 1967 (first break with KP, second break with PPT). The British pound sterling came under pressure in the mid-sixties since the exchange rate against the dollar was considered too high and was eventually devalued by 14.3% to 2.40 on 18 November 1967. This suggests that the first break detected with PPT in 1967, second quarter, is spurious.

Table 2: Break dates based on full sample

		q	PPT-AR(q)			KP-AR(q)	
1	GDPC96	1	1983:Q1	-	-	1983:Q4	-
		4	*1982:Q2*	-	-	*1983:Q4*	-
2	GDPDEF	1	-	-	-	*1984:Q3*	*2008:Q4*
		4	-	-	-	1981:Q2	2008:Q4
3	PCECC96	1	-	-	-	1993:Q1	-
		4	*1987:Q2*	-	-	*1992:Q2*	-
4	PCECTPI	1	2008Q1	-	-	1991:Q3	2006:Q4
		4	*2007Q3*	-	-	-	*2008:Q4*
5	GPDIC96	1	-	-	-	1984:Q4	-
		4	-	-	-	-	-
6*	OPHPBS	1	-	-	-	-	-
		4	-	-	-	-	-
7	ULCNFB	1	1983Q2	1999Q2	-	1984:Q1	2000:Q1
		4	-	-	-	-	-
8	CPIAUCSL	1	2008Q1	-	-	2008:Q4	-
		4	*2007Q3*	-	-	*2008:Q4*	-
9	PPIFCG	1	2008Q1	-	-	1972:Q3	2008:Q4
		4	*2007Q3*	-	-	-	*2008:Q4*
10	TB3MS	1	-	1979Q2	1984Q3	1979:Q4	1985:Q1
		4	*1965Q1*	*1978Q3*	*1983Q4*	*1979:Q4*	*1985:Q2*
11	GS10	1	1979Q2	1986Q1	-	1979:Q4	1986:Q4
		4	*1978Q3*	*1985Q2*	-	*1966:Q2*	-
12	M1SL	1	1978Q4	-	-	-	2008:Q3
		4	*1978Q1*	-	-	*1979.Q2*	*2008:Q4*
13	M2SL	1	-	-	-	-	-
		4	-	-	-	*1979.Q2*	-
14*	UTL11	1	-	-	-	-	-
		4	-	-	-	-	-
15*	SP500	1	-	-	-	-	-
		4	-	-	-	-	-
16	INDPRO	1	1982Q4	-	-	1984:Q1	2008:Q2
		4	*1980Q3*	-	-	*1983:Q4*	*2008:Q3*
17*	HOUST	1	-	-	-	-	-
		4	-	-	-	-	-
18	AHEMAN	1	1969Q2	1981Q4	-	1982:Q4	-
		4		*1980Q2*	-	*1983:Q4*	-
19	UNRATE	1	-	-	-	-	-
		4	-	-	-	*1983:Q4*	-
20	PAYEMS	1	1983Q3	-	-	1984Q2	-
		4	*1982Q2*	-	-	*1983:Q3*	-
21	EXUSUK	1	1967Q2	1967Q4	1971Q2	1967:Q4	-
		4	*1966Q3*	-	-	*1983:Q3*	-
22*	PMI	1	-	-	-	-	-
		4	-	-	-	-	-
23*	NAPMNOI	1	-	-	-	-	-
		4	-	-	-	-	-

Break dates are defined as the first observation of the new regime, using the median of the posterior of the states.

Table 3: Posterior means of AR(1) break models

S	R	PPT-AR(1)			KP-AR(1)			AR(1) full sample			AR(1) last 40 data		
		c	ϕ	σ^2	c	ϕ	σ^2	c	ϕ	σ^2	c	ϕ	σ^2
1	1	0.55	0.30	1.12	0.62	0.28	1.10	0.52	0.32	0.69	0.19	0.49	0.36
	2	0.39	0.46	0.32	0.38	0.45	0.29	-	-	-	*0.51*	*0.29*	*0.26*
2	1	0.12	0.87	0.09	0.14	0.87	0.10	0.11	0.87	0.09	0.33	0.41	0.09
	2	-	-	-	0.24	0.35	0.04	-	-	-	*0.29*	*0.58*	*0.03*
	3	-	-	-	0.23	-0.19	0.35	-	-	-	-	-	-
3	1	0.56	0.31	0.46	0.65	0.26	0.51	0.56	0.31	0.45	0.23	0.54	0.20
	2	-	-	-	0.46	0.42	0.22	-	-	-	*0.72*	*0.02*	*0.35*
4	1	0.13	0.86	0.10	0.13	0.87	0.10	0.15	0.83	0.13	0.42	0.24	0.23
	2	0.14	0.51	0.84	-0.09	0.01	0.64	-	-	-	*0.40*	*0.49*	*0.08*
	3	-	-	-	0.20	0.10	0.45	-	-	-	-	-	-
5	1	0.07	0.18	0.21	0.10	0.10	0.30	0.07	0.18	0.21	-0.02	0.54	0.12
	2	-	-	-	0.06	0.30	0.11	-	-	-	*0.08*	*0.02*	*0.11*
6*	1	0.56	-0.01	0.72	0.56	-0.01	0.71	0.57	-0.01	0.71	0.59	0.10	0.53
	-	-	-	-	-	-	-	-	-	-	*0.33*	*0.14*	*0.41*
7	1	0.60	0.40	1.28	0.60	0.43	1.21	0.51	0.32	1.11	0.26	-0.14	1.01
	2	0.47	0.14	0.33	0.47	0.14	0.33	-	-	-	*0.55*	*0.08*	*0.34*
	3	0.39	-0.07	1.33	0.38	-0.07	1.35	-	-	-	-	-	-
8	1	0.00	-0.29	0.18	0.01	-0.30	0.18	-0.00	-0.30	0.27	-0.01	-0.35	0.76
	2	-0.03	-0.30	2.69	-0.17	-0.30	3.62	-	-	-	*0.00*	*-0.35*	*0.18*
9	1	0.02	-0.38	1.08	0.03	-0.45	0.40	0.02	-0.32	1.69	0.05	-0.27	5.13
	2	0.03	-0.30	19.08	0.02	-0.38	1.32	-	-	-	*0.00*	*-0.37*	*1.65*
	3	-	-	-	-0.16	-0.23	21.71	-	-	-	-	-	-
10	1	0.04	0.36	0.32	0.03	0.33	0.31	-0.01	0.23	0.57	-0.06	0.61	0.19
	2	-0.01	0.25	3.80	-0.00	0.19	3.45	-	-	-	*0.01*	*0.67*	*0.13*
	3	-0.03	0.55	0.14	-0.03	0.56	0.14	-	-	-	-	-	-
11	1	0.04	0.22	0.08	0.03	0.21	0.06	-0.00	0.23	0.23	-0.06	0.02	0.13
	2	0.01	0.22	0,93	-0.02	0.24	0.22	-	-	-	*-0.05*	*0.38*	*0.23*
	3	-0.04	0.22	0.16	-0.05	0.20	0.15	-	-	-	-	-	-
12	1	0.01	-0.18	0.36	-0.00	-0.32	0.77	0.00	-0.30	0.96	0.05	-0.28	2.12
	2	-0.00	-0.31	1.35	-0.04	-0.24	7.30	-	-	-	*-0.07*	*-0.11*	*0.94*
13	1	-0.00	-0.15	0.48	-0.01	-0.15	0.46	-0.01	-0.15	0.47	-0.04	-0.16	0.79
	-	-	-	-	-	-	-	-	-	-	*-0.05*	*-0.12*	*0.30*
14*	1	0.25	0.97	0.02	0.26	0.97	0.02	0.25	0.97	0.02	0.42	0.94	0.02
	-	-	-	-	-	-	-	-	-	-	*0.45*	*0.95*	*0.01*
15*	1	0.11	0.24	0.45	0.11	0.23	0.44	0.11	0.23	0.45	-0.04	0.36	0.67
	-	-	-	-	-	-	-	-	-	-	*0.28*	*-0.15*	*0.50*
16	1	0.39	0.44	4.00	0.46	0.45	3.11	0.35	0.51	1.99	-0.00	0.70	1.23
	2	0.21	0.67	0.85	0.25	0.61	0.69	-	-	-	*0.33*	*0.54*	*0.62*
	3	-	-	-	0.05	0.76	2.90	-	-	-	-	-	-
17*	1	0.18	0.97	0.01	0.18	0.97	0.01	0.21	0.97	0.01	-0.27	1.03	0.01
	-	-	-	-	-	-	-	-	-	-	*0.63*	*0.91*	*0.01*
18	1	0.83	0.14	0.40	0.72	0.52	0.40	0.38	0.65	0.26	0.63	0.04	0.07
	2	1.00	0.47	0.26	0.55	0.22	0.07	-	-	-	*0.49*	*0.28*	*0.07*
	3	0.58	0.19	0.08	-	-	-	-	-	-	-	-	-
19	1	0.01	0.64	0.07	0.01	0.64	0.07	0.01	0.65	0.07	0.03	0.74	0.07
	-	-	-	-	-	-	-	-	-	-	*-0.02*	*0.70*	*0.03*
20	1	0.14	0.76	0.17	0.15	0.76	0.17	0.08	0.83	0.10	-0.02	0.86	0.07
	2	0.03	0.89	0.04	0.03	0.90	0.04	-	-	-	*0.06*	*0.88*	*0.03*
21	1	-0.00	0.20	0.00	-0.00	0.19	0.00	-0.02	0.26	0.17	0.00	0.41	0.19
	2	-0.31	0.27	0.40	-0.03	0.26	0.20	-	-	-	*0.05*	*0.14*	*0.35*
	3	0.01	0.12	0.00	-	-	-	-	-	-	-	-	-
	4	-0.03	0.24	0.22	-	-	-	-	-	-	-	-	-
22*	1	0.89	0.83	0.16	0.96	0.82	0.15	1.01	0.81	0.15	0.93	0.82	0.13
	-	-	-	-	-	-	-	-	-	-	*0.95*	*0.82*	*0.07*
23*	1	1.20	0.78	0.27	1.31	0.76	0.26	1.40	0.75	0.26	1.36	0.75	0.33
	-	-	-	-	-	-	-	-	-	-	*1.30*	*0.77*	*0.14*

S = series number (see Table 2); R = regime number. Each AR(1) is written $y_t = c + \phi y_{t-1} + \sigma \epsilon_t$. Two estimations are reported in the block "AR(1) last 40 data": on the first row, the results are for the last 40 points of the full sample, on the second row (in italics), they are for the last 40 points ending at seventy percent of the full sample.

4 Forecasting Implementation

In this section, we explain how we forecast with the PPT and KP models, and in sub-section 4.3 we review briefly the other models with which we generate alternative forecasts to be compared with the forecasts coming from the break models.

The setup is the following: we shall carry out a recursive forecasting exercise for the final α percent of the observations. This means that we first estimate the models with an initial sample consisting of $1 - \alpha$ percent of the data, and we forecast future observations. Then we add one data point, estimate and forecast again, until we have consumed all the data.

4.1 Forecasting with PPT

With the PPT approach, if one were to assume that no breaks occur out-of-sample, forecasting could be done in a straightforward way based on the posterior density of the the parameters of the regime that holds at the end of the estimation sample. Such an approach, of course, does not address the issue of forecasting when breaks can occur out-of-sample. Appendix A provides details about how predictive simulation is implemented for the PPT model.

To choose the number of breaks, we choose a maximum number of regimes, $K^{\max}$, evaluate the marginal likelihood for $K = 1, .., K^{\max}$ and select the optimal number of regimes as the one which maximizes the marginal likelihood. However, in the context of a recursive forecasting exercise, we want $K^{\max}$ to vary over time as the number of regimes can increase as time goes by. Accordingly, we adopt the following strategy.

Using the initial sample of observations, we calculate the optimal number of regimes as described in the preceding paragraph. Then we begin our recursive forecasting exercise. Let K_t be the number of regimes in a model using data through time t. We compute marginal likelihoods for $K_t = \{1, \ldots, K^*_{t-1} + 1\}$ where K^*_{t-1} is the optimal number of regimes at $t - 1$ and select K^*_t as the value that maximizes the marginal likelihood. We do this for $t = T_0 + 1, \ldots, T - h$ where $T_0 = \alpha T$. Marginal likelihoods are calculated as described in Bauwens and Rombouts (2010), based on output from the posterior simulator.

We calculate two predictive densities, one which assumes no future break, and one of which allows for a possible single break in the forecast period. The necessary details are given in Appendix A.

4.2 Forecasting with KP

With the KP approach, dealing with out-of-sample structural breaks is straightforward. Suppose regime j holds at the end of the estimation sample (called t) and, thus, $s_t = j$. The posterior simulation algorithm produces $\Pr(s_{t+1} = j|Y_t)$ and $\Pr(s_{t+1} = j+1|Y_t)$, where $Y_t = (y_1, .., y_t)'$. Furthermore, the posterior simulation algorithm provides us with draws from $p(\beta_j, \sigma_j|Y_t)$ and $p(\beta_{j+1}, \sigma_{j+1}|Y_t)$. These are the components needed to do forecasting with structural breaks. Appendix C provides details about how predictive simulation is implemented for the KP model.

Defining the optimal number of regimes for each sample in our recursive forecasting exercise is done in a way similar to the PPT model described previously, but without the need to compute marginal likelihoods. Using output from the posterior simulator using data through time t, we calculate the optimal number of breaks as $K_t^* = median(\Pr(s_t|data))$, i.e. the median of the posterior of the state variable of the last observation.

In particular, we run the model for $t = T_0$ (where $T_0 = \alpha T$) for a large number of breaks. Then instead of using marginal likelihoods to estimate the optimal number of breaks at time T_0, we just use the estimate $K_{T_0}^* = median(\Pr(s_{T_0}|data))$. In the next period ($t = T_0 + 1$) we estimate the KP model with K_{T_0+1} breaks and forecast, where we define $K_{T_0+1} = K_{T_0}^* + 1$. From the Gibbs sampler output we estimate $K_{T_0+1}^* = median(\Pr(s_{T_0+1}|data))$. Then we increase the observations by one ($t = T_0 + 2$) and set $K_{T_0+2} = K_{T_0+1}^* + 1$ and so on.

In words, with number of observations t we always allow for one more break than the optimal number of breaks estimated in the previous sample $t-1$. However, when we set the number of breaks using the formula $K_t = K_{t-1}^* + 1$, this doesn't necessarily mean that we forecast with exactly $K_{t-1}^* + 1$ breaks at time t. This is the maximum number of breaks. This implies that it might be the case $K_t^* = K_{t-1}^*$ so that the number of regimes we use to forecast hasn't changed. Therefore, as we progress at time $t+1$ we set $K_{t+1} = K_t^* + 1 = K_{t-1}^* + 1$. Nevertheless, if the optimal number of estimated regimes at time t has actually changed to $K_t^* = K_{t-1}^* + 1$ (we discovered an additional break), then we ought to set at time $t+1$ a maximum number of regimes $K_{t+1} = K_t^* + 1 = K_{t-1}^* + 2$.

In the recursive forecasting setting, we repeat this procedure for $t = T_0 + 1, \ldots, T - h$.

4.3 Forecasting with Other Approaches

In addition to the forecasting methods of KP and PPT outlined above, we consider a variety of other "no-break" models.

Our first approach is a standard TVP-AR(1) model. This is a restricted special case of the KP approach. That is, if we adopt the KP framework but set $s_t = t$ for all time periods (or equivalently, $K_t^{\max} = t$ and $\Pr(s_t = t|s_{t-1} = t-1) = 1$ then we obtain the standard TVP model which is of the form

$$\begin{aligned} y_t &= Z_t\beta_t + \sigma_t\varepsilon_t \\ \beta_t &= \beta_{t-1} + u_t \\ log\left(\sigma_t\right) &= log\left(\sigma_{t-1}\right) + v_t \end{aligned} \tag{8}$$

where $\varepsilon_t \sim N(0,1)$, $u_t \sim N(0, B_0)$ and $v_t \sim N(0,\delta)$. Note that for this special case we need extra care in defining our priors, since the autoregressive coefficients evolve as random walks for all t periods and they can easily become explosive. The priors we use for this model are

$$\begin{aligned} \beta_0 &\sim N_m\left(0, 4I_m\right) \\ log\left(\sigma_0\right) &\sim N\left(0,1\right) \\ B_0^{-1} &\sim Wishart\left(m+1, (0.001^2(m+1)R)^{-1}\right) \\ \delta^{-1} &\sim Gamma\left(1, 0.1\right). \end{aligned}$$

where R is a diagonal matrix with elements $R\{1,1\} = 5$ for the intercept, and $R\{i,i\} = 1/i$ for lag length $i = 1, ..., p$. Forecasting in this model requires first to simulate the future paths of the time-varying coefficients β_t and $log\left(\sigma_t\right)$ using their random walk specifications. Then conditional on these simulated out-of-sample coefficients, we forecast y_{T+h} as in a simple regression model.

We also present recursive and rolling AR(q) forecasting results (with q set to one and to four). Bayesian inference is used for these models using the same prior density as in the PPT implementations if we allow for only a single regime. For the rolling forecasts we use a window of ten years (forty observations). We tried a window of five years but the forecast results are much deteriorated by this choice. A window of ten years seems reasonable since we have about thirty-five years available before the forecast period, and we want to make this different enough from the sample used with the recursive approach.[4]

Finally we also use an unobserved component model with stochastic volatility (UC-SV). We follow the formulation of Stock and Watson (2007), who specify a model with only a time-varying trend (no AR dynamics), which takes the form

$$\begin{aligned} y_t &= \mu_t + \sigma_{\epsilon,t}\varepsilon_t \\ \mu_t &= \mu_{t-1} + \sigma_{\eta,t}\eta_t \\ log\left(\sigma_{\epsilon,t}\right) &= log\left(\sigma_{\epsilon,t-1}\right) + v_t \\ log\left(\sigma_{\eta,t}\right) &= log\left(\sigma_{\eta,t-1}\right) + w_t \end{aligned} \tag{9}$$

where in this case, $(\varepsilon_t, \eta_t) \sim N(0, I_2)$, $u_t \sim N(0, \gamma_1)$ and $v_t \sim N(0, \gamma_2)$. For U.S. inflation, Stock and Watson (2007) set $\gamma_1 = \gamma_2 = 0.2$. We estimate these parameters and the priors we use to forecast with this model are

$$\begin{aligned} \mu_0 &\sim N_m(0, 4) \\ log\left(\sigma_{\epsilon,0}\right) &\sim N(0, 1) \\ log\left(\sigma_{\eta,t}\right) &\sim N(0, 1) \\ B_0^{-1} &\sim Gamma\left(1, 0.1\right) \\ \gamma^{-1} &\sim Gamma\left(1, 0.1\right). \end{aligned}$$

Forecasting in the above model is similar in spirit with the TVP and KP models. We first need to simulate the future values of the time-varying parameters, and then plug in these simulated values in the first equation in 9.

Table 4 lists the models used in the forecasting evaluations, with a short definition.

5 Results of Forecasting Evaluations

For each series listed in Table 1, we carry out a recursive forecasting exercise for the final thirty percent of the observations: we first estimate the models with an initial sample consisting of seventy percent of the data, and we forecast at the horizons h equal 1 and 4. Then we add one data point, estimate and forecast again, until the end of the data. Thus we have

[4] Choosing the window size optimally is discussed in Pesaran and Timmermann (2007). Their analytical results do not apply to AR models. Using the cross-validation procedure they propose is left for future research.

Table 4: Models used in the forecasting evaluations

Name	Description
PPT10	PPT, AR(1), 0 break allowed in forecast period
PPT11	PPT, AR(1), 1 break allowed in forecast period
PPT40	PPT, AR(4), 0 break allowed in forecast period
PPT41	PPT, AR(4), 1 break allowed in forecast period
KP1	KP, AR(1)
KP1	KP, AR(4)
TVP1	TVP-AR(1)
TVP4	TVP-AR(4)
ROW1	AR(1) estimated with rolling window of 10 years
ROW4	AR(4) estimated with rolling window of 10 years
REC1	AR(1) estimated on expanding window
REC4	AR(4) estimated on expanding window
UC-SV	Unobserved component model with stochastic volatility

61 one-step and 58 four-step ahead forecasts on which we can base the forecast evaluations. For $h > 1$, our forecasts are all iterated (see, e.g., Marcellino, Stock, and Watson (2006) for a motivation for use of iterated over direct forecasts).

Our forecast metrics are RMSE and the average of log predictive likelihoods (APL). RMSE is based on point forecasts and we use the predictive median as point forecast. The predictive likelihood is the predictive density evaluated at the observed outcome. This is estimated by a nonparametric kernel smoother using draws from the predictive simulator.

For each series in Table 1, we provide in Appendix C the RMSE and APL values from the recursive forecasting exercise. For one-step ahead forecasts, see Tables 13 (RMSE) and 15 (APL) and for four-step ahead forecasts see Tables 14 and 16. We report the relative values, with the model in the last column (UC-SV) serving as reference.

The RMSE/APL values for the reference model are reported to fix their order of magnitude. For example, in Table 13, we see that for the UC-SV model and the first series, the RMSE is equal to 0.608, whereas the relative RMSE for PPT10 is 0.989, implying that

PPT11 has a RMSE 1.1 percent lower than the UC-SV model. For each series, the smallest (for RMSE) or largest (for APL) value across all models is in bold. If this global minimum is in the set of break models, the value in italics is the minimum across the no-break models.[5] If the global minimum is in the latter group, the value in italics is the minimizer across the break models.

We discuss the results based on the RMSE criterion in subsection 5.1, and in subsection 5.2 the results based on the APL criterion. Generally, we are interested in three questions:

> **Question 1:** How does the forecasting performance differ between break models and no-break models?
>
> **Question 2:** How does the forecasting performance differ between PPT, KP, and TVP?
>
> **Question 3:** How does the forecasting performance differ between lag orders?

5.1 RMSE Results

To summarize the contents of Tables 13 and 14, we provide in Table 5 the list of the best model for each series, together with the relative performance of the best break model with respect to the best no-break model. It appears that according to the RMSE criterion, at horizon one, the break models are the best in 83 percent of all series (26 for PPT, 22 for KP1, and 35 for TVP1). At horizon four, the break models forecast better in 70 percent (30 for PPT1, 10 for KP1, and 30 for TVP1). REC is best for four series at horizon one and five at horizon four, ROW is best only for one series at horizon four, and UC-SV as well. These scores do not take account of the magnitude of the differences of the RMSE between the different models (for this see below). Though there are many cases where the best model differs between horizons one and four, a switch between a break model and a no-break one happens in seven series on a total of twenty-three.

With the results in Tables 5–13-14, we can answer to our questions about the forecasting performance of the different models.

Question 1: To answer, we compare the best break model RMSE value to the best no-break model value, see columns "% diff." in Table 5. For example, a value of -3 (+3) means that the

[5] The set of break models consists of PPT, KP and TVP models, and the set of no-break models consists of UC-SV, ROW and REC.

best break (no-break) model has its RMSE three percent smaller (larger) than the RMSE of the best no-break (break) model. Although for a high proportion of the series the differences are negative, they are nevertheless small, by what we mean they are less than five percent (often much less). Exceptions are, at horizon one, series 10 (-6 for KP1), 17 (-18 for TVP1), and 20 (-11 for KP4). At horizon four, one difference is larger than 5 (series 10, +11 for REC4). A test for the nullity of the mean of the differences is significant at the five percent level for horizon one, but not for horizon four. In brief, there is some weak evidence in our results that break models perform a little better than no-break models.

Table 5: Relative performance of best forecasting models
on last thirty percent of sample
(Root mean squared error criterion)

		$h = 1$		$h = 4$	
		best model	% diff.	best model	% diff
1	GDPC96	PPT41	-1.9	TVP4	-2.9
2	GDPDEF	TVP4	-0.6	UC-SV	2.0
3	PCECC96	PPT40	-2.7	KP4	-4.6
4	PCECTPI	TVP1	-1.2	TVP1	-0.2
5	GPDIC96	PPT40	-0.3	PPT10	-0.8
6*	OPHPBS	REC4	0.2	REC4	0.1
7	ULCNFB	TVP1	-1.6	TVP1	-0.3
8	CPIAUCSL	TVP4	2.0	PPT10	-0.3
9	PPIFCG	TVP4	3.1	REC4	-0.6
10	TB3MS	KP1	-5.8	REC4	0.2
11	GS10	PPT40	-0.4	PPT40	-0.3
12	M1SL	TVP1	-0.3	PPT11	-0.2
13	M2SL	REC4	0.4	KP1	0.0
14*	UTL11	PPT40	-3.5	PPT40	-1.6
15*	SP500	PPT11	-0.1	TVP1	-0.5
16	INDPRO	KP4	-1.4	TVP4	-1.3
17*	HOUST	TVP1	-17.8	PPT11	-1.0
18	AHEMAN	TVP1	-0.6	ROW4	0.3
19	UNRATE	KP4	-1.8	PPT41	1.8
20	PAYEMS	KP4	-10.6	TVP4	-3.2
21	EXUSUK	REC4	0.2	TVP1	-0.5
22*	PMI	KP4	-0.1	REC4	0.2
23*	NAPMNOI	REC4	0.2	REC4	10.6
	Mean		-1.93		-0.13
	St. Dev.		4.39		2.75
	t-stat		-2.11		-0.24

Source: results in Tables 13-14. See Table 4 for definitions of models. The "%diff" are computed as [(smallest RMSE across the break models/smallest RMSE across the no-break models)-1]x100.

Question 2: The relative differences (in percent) between the RMSE of the different models are shown in Table 6. For example, the value -0.49 of series 1 for a comparison of PPT10 and KP1 means that PPT10 is performing better than KP1 by almost half a percent. Means and standard deviations are given at the bottom of each column. The results show that for most series the differences are small, and there are a few cases where they are large. On average, at horizon one, PPT performs slightly better than KP, and TVP better than the other two models. At horizon four, PPT performs better on average than the other two models, and TVP dominates KP. Nevertheless given the large standard deviations due to a few large differences, no mean is significant even at the ten percent level.

Table 6: Performance comparison of break models
on last thirty percent of sample
(Root mean squared error criterion)

	Series	$\frac{PPT10}{KP1}$	$\frac{PPT10}{TVP1}$	$\frac{KP1}{TVP1}$	$\frac{PPT40}{KP4}$	$\frac{PPT40}{TVP4}$	$\frac{KP4}{TVP4}$
1	GDPC96	-0.49	-1.01	-0.52	-2.02	-1.88	0.14
2	GDPDEF	-5.53	1.56	7.50	-1.33	2.74	4.12
3	PCECC96	1.12	-1.55	-2.65	-2.20	-6.06	-3.94
4	PCECTPI	-0.19	9.01	9.22	-2.17	4.56	6.87
5	GPDIC96	0.56	-1.23	-1.78	-0.24	-1.26	-1.02
6*	OPHPBS	-1.50	-3.15	-1.67	0.09	-3.89	-3.97
7	ULCNFB	-4.78	12.06	17.69	-0.60	2.49	3.11
8	CPIAUCSL	1.76	0.40	-1.33	-42.13	7.39	85.56
9	PPIFCG	1.32	1.15	-0.16	-45.80	5.15	94.01
10	TB3MS	1.92	-9.36	-11.07	-6.70	-8.13	-1.54
11	GS10	-0.58	-0.54	0.03	-0.48	-3.15	-2.68
12	M1SL	-5.64	0.18	6.17	20.01	21.60	1.32
13	M2SL	-0.11	-0.30	-0.19	-47.93	-2.66	86.94
14*	UTL11	6.44	29.54	21.70	-3.68	-23.08	-20.14
15*	SP500	0.21	-0.22	-0.43	-24.30	-1.03	30.74
16	INDPRO	-1.12	-6.75	-5.70	16.55	10.33	-5.34
17*	HOUST	-0.81	27.36	28.41	-1.33	-20.82	-19.75
18	AHEMAN	-9.09	0.32	10.35	1.49	5.24	3.69
19	UNRATE	-1.03	-13.49	-12.59	2.63	-14.77	-16.95
20	PAYEMS	-2.35	-10.12	-7.95	2.63	-14.93	-17.11
21	EXUSUK	-0.28	-0.95	-0.67	-0.44	-3.01	-2.58
22*	PMI	-1.08	3.03	4.15	1.34	-1.91	-3.21
23*	NAPMNOI	0.28	-1.48	-1.75	2.72	0.40	-2.26
	Mean	-0.91	1.50	2.47	-5.82	-2.03	9.39
	St. Dev.	3.11	10.10	9.89	17.53	9.96	33.16
	t-stat	-1.40	0.71	1.20	-1.59	-0.98	1.36

Source: results in Tables 13-14. See Table 4 for definitions of models. The values for column header $\frac{A}{B}$ are computed as [(RMSE of model A/RMSE of model B)-1]x100.

Question 3: The relative differences (in percent) between the RMSE of the different models are reported in Table 7. These results indicate that the models with four lags perform a little better than those with one lag, maybe not a surprise for quarterly data. However, the differences are significant at the ten percent level on average only for PPT and REC.

Another question of interest is whether allowing for a possible single break (rather than no break) in the forecast period makes a difference in the PPT approach. Pesaran, Pettenuzzo, and Timmermann (2006) found on their example (a single series) that this decreases RMSE at all horizons on their full sample and on several subsamples. We don't find this to be significant on average for our series with one lag (t-stat -0.05 at horizon 1 and 0.81 at horizon four), but with four lags there is some evidence in favor of allowing for a possible break: the performance is improved on average by 0.49 percent at horizon 1 (t-stat 1.91) and by 2 percent at horizon four (t-stat 1.75).

5.2 APL Results

We summarize the contents of Tables 15 and 16 in Table 8 where we list the best model for each series, together with the relative performance of the best break model with respect to the best no-break model. It appears that according to the APL criterion, at horizon one, the break models are the best in 22 percent of all series (9 for PPT, 4 for KP1, and 9 for TVP1). At horizon four, the break models forecast better also in 22 percent (13 for PPT1, 0 for KP1, and 9 for TVP1). ROW is the best at horizon one for fourteen series (61 percent) and seventeen (74 percent) at horizon four. REC is the best for four series at horizon one and one at horizon four, and UC-SV is dominated by all other models. These scores do not take account of the magnitude of the differences of the APL between the different models but suggest that ROW is by far dominating the other models (for this see question 1 below). Though there are many cases where the best model differs between horizons one and four, a switch between a break model and a no-break model happens in six series on a total of twenty-three.

With the results in Tables 8–15-16, we can answer to our questions about the forecasting performance of the different models.

Question 1: To answer, we compare the best break model APL value to the best no-break model value, see columns "% diff." in Table 8. For example, a value of +4 (-4) means that the best break (no-break) model has its APL four percent larger than the APL of the best no-

Table 7: Performance comparison of lag orders
on last thirty percent of sample
(Root mean squared error criterion)

	Series	$\frac{PPT10}{PPT40}$	$\frac{KP1}{KP4}$	$\frac{TVP1}{TVP4}$	$\frac{ROW1}{ROW4}$	$\frac{REC1}{REC4}$
1	GDPC96	4.46	2.85	3.54	2.88	4.54
2	GDPDEF	4.73	9.39	5.95	-3.48	11.67
3	PCECC96	18.43	14.53	13.01	7.35	14.96
4	PCECTPI	1.99	-0.03	-2.18	-1.60	2.87
5	GPDIC96	1.50	0.70	1.47	-3.07	1.87
6*	OPHPBS	3.38	5.05	2.59	-1.20	3.85
7	ULCNFB	8.80	13.58	-0.49	-1.60	12.54
8	CPIAUCSL	-3.25	-44.97	3.49	1.00	3.29
9	PPIFCG	3.59	-44.59	7.68	3.59	7.65
10	TB3MS	0.02	-8.44	1.37	-7.72	-1.58
11	GS10	4.96	5.06	2.22	4.62	4.97
12	M1SL	-17.84	4.49	-0.28	-9.13	-0.82
13	M2SL	13.09	-41.05	10.41	11.03	13.57
14*	UTL11	54.71	39.99	-8.14	43.62	40.93
15*	SP500	-0.95	-25.17	-1.75	-9.70	-2.74
16	INDPRO	-12.93	2.63	3.02	-4.10	5.94
17*	HOUST	6.30	5.74	-33.9	1.46	5.61
18	AHEMAN	-4.42	6.71	0.27	-2.70	17.20
19	UNRATE	-0.11	3.58	-1.59	-5.87	-1.48
20	PAYEMS	8.90	14.47	3.08	4.27	2.60
21	EXUSUK	2.48	2.31	0.35	2.19	2.25
22*	PMI	12.25	14.98	6.86	10.11	13.91
23*	NAPMNOI	5.20	7.76	7.21	1.88	7.92
	Mean	5.01	-0.45	1.05	1.91	7.46
	St. Dev.	13.35	20.40	8.87	10.65	9.25
	t-stat	1.80	-0.11	0.57	0.86	3.87

Source: results in Tables 13-14. See Table 4 for definitions of models. The values for column header $\frac{A}{B}$ are computed as [(RMSE of model A/RMSE of model B)-1]x100.

Table 8: Relative performance of best forecasting models
on last thirty percent of sample
(Average predictive likelihood criterion)

		$h = 1$		$h = 4$	
		best model	% diff.	best model	% diff
1	GDPC96	PPT10	0.4	ROW1	-0.4
2	GDPDEF	ROW1	-7.2	ROW4	-14.3
3	PCECC96	REC1	-12.4	ROW1	-9.9
4	PCECTPI	ROW1	-8.4	ROW4	-22.7
5	GPDIC96	ROW1	-16.0	ROW1	-16.7
6*	OPHPBS	ROW1	-3.1	ROW1	-8.1
7	ULCNFB	ROW1	-5.4	ROW1	-4.5
8	CPIAUCSL	ROW4	-7.2	ROW1	-4.5
9	PPIFCG	REC4	2.9	TVP1	5.9
10	TB3MS	ROW1	-0.8	PPT10	6.9
11	GS10	ROW4	-9.9	ROW1	-7.1
12	M1SL	REC4	-1.7	REC4	-2.6
13	M2SL	TVP4	2.7	ROW1	-1.2
14*	UTL11	ROW4	-17.8	ROW1	-1.6
15*	SP500	KP1	0.7	PPT10	0.1
16	INDPRO	TVP1	7.8	TVP1	9.7
17*	HOUST	ROW1	-19.0	ROW4	-13.6
18	AHEMAN	PPT10	0.1	ROW1	-0.7
19	UNRATE	ROW1	-14.2	ROW1	-15.5
20	PAYEMS	ROW1	1.6	PPT10	6.8
21	EXUSUK	ROW1	-2.5	ROW1	0.5
22*	PMI	ROW4	-13.4	ROW1	-9.1
23*	NAPMNOI	REC4	-0.6	ROW1	-3.9
	Mean		-5.36		-4.64
	St. Dev.		7.40		8.31
	t-stat		-3.48		-2.68

Source: results in Tables 15-16. See Table 4 for definitions of models. The "%diff" are computed as [(largest APL across the break models/largest APL across the no-break models)-1]x100.

break (break) model. At horizon one, the differences are larger than five percent in absolute value for twelve series, and only for one (series 16) the difference is positive. At horizon four, nine differences are smaller than minus five percent and four are larger than five percent. A test for the nullity of the mean of the differences is significant at the one percent level for both horizons. In brief, there is strong evidence in our results that the no-break models (especially ROW) perform much better than break models, though there are a few exceptions (series 16 at both horizons, series 9, 10 and 20 at horizon four).

Question 2: The relative differences (in percent) between the APL of the different models are shown in Table 9. For example, the value 8.28 of series 1 for a comparison of PPT10 and

KP1 means that PPT10 is performing better than KP1 by a little more than 8 percent. The differences vary a lot, and there are a few cases where they are very large. On average, at both horizons, PPT performs slightly better than KP but not significantly even at the ten percent level, and TVP is significantly dominated by the other two models.

Table 9: Performance comparison of break models
on last thirty percent of sample
(Average predictive likelihood criterion)

	Series	$\frac{PPT10}{KP1}$	$\frac{PPT10}{TVP1}$	$\frac{KP1}{TVP1}$	$\frac{PPT40}{KP4}$	$\frac{PPT40}{TVP4}$	$\frac{KP4}{TVP4}$
1	GDPC96	8.28	1.71	-6.07	0.08	1.11	1.03
2	GDPDEF	3.31	48.96	44.18	-3.33	37.08	41.81
3	PCECC96	-1.86	-7.41	-5.65	-7.41	-12.08	-5.04
4	PCECTPI	-4.47	19.62	25.22	-2.76	25.70	29.27
5	GPDIC96	-2.65	6.73	9.64	-3.95	6.32	10.69
6*	OPHPBS	-0.83	-4.18	-3.38	-1.63	1.13	2.80
7	ULCNFB	4.92	-3.50	-8.02	-1.66	-7.39	-5.82
8	CPIAUCSL	-3.80	14.15	18.66	-3.66	12.83	17.11
9	PPIFCG	-1.51	-10.38	-9.00	-4.14	-4.64	-0.52
10	TB3MS	-2.33	0.50	2.90	4.16	4.44	0.27
11	GS10	9.48	20.51	10.08	3.25	13.08	9.52
12	M1SL	8.12	-2.35	-9.68	-2.62	-2.25	0.39
13	M2SL	-0.89	-4.28	-3.42	69.45	-7.92	-45.66
14*	UTL11	29.10	306.29	214.72	10.75	337.30	294.86
15*	SP500	-1.14	6.42	7.64	-0.47	2.88	3.37
16	INDPRO	3.52	-5.72	-8.92	7.66	-15.32	-21.35
17*	HOUST	-1.10	371.17	376.40	-2.92	397.96	412.92
18	AHEMAN	8.09	44.02	33.24	0.49	22.09	21.49
19	UNRATE	-2.10	28.62	31.38	-8.20	20.69	31.47
20	PAYEMS	8.28	57.59	45.54	9.65	52.01	38.63
21	EXUSUK	-5.57	5.73	11.97	-3.31	-1.97	1.39
22*	PMI	-8.71	39.06	52.32	-4.01	44.46	50.50
23*	NAPMNOI	-0.91	30.98	32.19	-4.96	33.54	40.51
	Mean	1.97	17.79*	16.56*	2.19	15.60*	15.29*
	St. Dev.	7.78	26.17*	24.77*	15.47	26.86*	29.09*
	t-stat	1.21	3.12	3.06	0.68	2.66	2.41

Source: results in Tables 15-16. See Table 4 for definitions of models. The values for column header $\frac{A}{B}$ are computed as [(APL of model A/APL of model B)-1]x100. Means and standard deviations with a $*$ superscript are computed excluding the values for series 14 and 17.

Question 3: The relative differences (in percent) between the RMSE of the different models are reported in Table 7. On average models with four lags do not perform better than models with one lag at the ten percent level, except for recursive OLS.

Unlike for the RMSE criterion, the relative performances of PPT with no break and one break allowed in the forecast period are significantly different on average for our series.

Allowing for one break deteriorates the performance on average: with one lag by 0.68 percent at horizon one (t-stat -5.29) and 5.42 percent at horizon four (t-stat -6.65); with four lags by 0.93 percent at horizon one (t-stat -6.05) and 8.40 percent at horizon four (t-stat 14.4). This is is explained by the increase of the predictive variances when one break is allowed, while the predictive means do not change much as witnessed by the RMSE results.

Table 10: Performance comparison of lag orders
on last thirty percent of sample
(Average predictive likelihood criterion)

	Series	$\frac{PPT10}{PPT40}$	$\frac{KP1}{KP4}$	$\frac{TVP1}{TVP4}$	$\frac{ROW1}{ROW4}$	$\frac{REC1}{REC4}$
1	GDPC96	4.36	-3.54	3.75	5.41	-1.57
2	GDPDEF	7.34	0.43	-1.22	0.80	-4.74
3	PCECC96	-4.93	-10.31	-9.73	-7.72	-4.16
4	PCECTPI	-5.97	-4.29	-1.19	4.55	-6.49
5	GPDIC96	3.27	1.89	2.87	4.61	-1.80
6*	OPHPBS	-2.70	-3.48	2.69	2.74	-3.98
7	ULCNFB	2.43	-4.00	-1.70	4.75	-7.28
8	CPIAUCSL	-5.73	-5.58	-6.82	-3.42	-8.66
9	PPIFCG	-2.48	-5.09	3.77	-2.42	-4.42
10	TB3MS	1.73	8.49	5.71	5.43	-3.27
11	GS10	1.03	-4.73	-5.21	-2.64	-2.10
12	M1SL	6.39	-4.17	6.50	6.41	-1.69
13	M2SL	-7.51	58.12	-11.03	-5.85	-8.81
14*	UTL11	-7.43	-20.59	-0.36	-19.11	-21.82
15*	SP500	3.99	4.69	0.54	3.33	2.86
16	INDPRO	11.49	15.95	0.14	7.18	-7.98
17*	HOUST	-1.42	-3.23	4.18	1.49	-2.75
18	AHEMAN	12.61	4.69	-4.54	1.50	-16.72
19	UNRATE	6.24	-0.39	-0.32	6.15	1.41
20	PAYEMS	9.17	10.55	5.30	3.91	-3.60
21	EXUSUK	7.73	10.31	-0.11	6.83	0.57
22*	PMI	-7.62	-2.87	-4.03	-6.47	-10.15
23*	NAPMNOI	-1.09	-5.13	0.84	3.85	-5.43
	Mean	1.34	1.64	-0.43	0.92	-5.33
	St. Dev.	6.26	14.52	4.74	6.25	5.54
	t-stat	1.03	0.54	-0.44	0.71	-4.61

Source: results in Tables 15-16. See Table 4 for definitions of models. The values for column header $\frac{A}{B}$ are computed as [(APL of model A/APL of model B)-1]x100.

5.3 Discussion of Previous Results

For the APL criterion and the last thirty percent of the sample that serves as forecast period, we find that the no-break models, especially rolling AR, perform significantly better than the break models. For the RMSE criterion, we find some weak evidence in favor of break models. Why this difference?

The APL criterion takes into account the whole shape of the predictive density. This is not normal despite the assumption of normality (conditional on the parameters), because it is integrated with respect to a posterior distribution that is not symmetric. However our predictive densities are very moderately skewed since we forecast at short horizons. Therefore, we can summarize the shape of our predictive by their standard deviation. The RMSE results indicate that in terms of the location of the point forecasts in the support of the predictive densities, the two kinds of models (break/no-break) are roughly equivalent on average (of course, individual exceptions occur). Thus logically the differences in the APL results must be (at least partly) due to differences in the standard deviation of the predictive densities. In the results, we find some weak evidence that supports our explanation.

Our rationale uses estimation results reported in Table 3 for the PPT- and KP-AR(1) models and also for the no-break AR(1) models estimated with an expanding window (AR(1) full sample header, named REC hereafter) and a rolling window of forty observations (AR(1) last forty data, named ROW1 hereafter). For the latter, in the last three columns of the table, we report two sets of point estimates: on the first row the estimates are computed with the last forty observations of the full sample, on the second row (in italics), they are computed with the last forty observations of the sample that ends just before the forecast period begins (1995). We call the latter the pre-forecast sample. For example, for series 1, the posterior expectation of the error variance is equal to 0.36 for the last forty observations of the full sample, 0.26 for the pre-forecast sample, and 0.69 for the full sample.

If we compare the pre-forecast ROW1 variance estimates with those of the regime generating the PPT and KP forecasts, we find that for most series the ROW1 estimate is smaller than the PPT, KP, or even REC estimates.[6] This is nothing else but the effect of the great moderation. Since the variance of the error determines to a large extent the predictive variance, we expect that that for the series witnessing this effect, the predictive densities are more concentrated when based on estimates using essentially data in that period than using data covering the period that precedes the great moderation (remember that the great moderation

[6] For series where no break is detected, estimates for the three models should obviously be almost identical, and this is indeed the case. See series identified by a * superscript on their identification number in Table 3.

starts in the mid-eighties and our forecast period starts about ten years later). Thus for an observation that is not far from the mean, the predictive density of ROW1 should be larger than the predictive of PPT, if the predictive densities have similar means. For an observation far in the tails, the reverse is true. We indeed observe this on many graphs of predictive densities. Hence if the observations of the forecast sample are not outliers in the predictive, and the predictive of both models have approximately the same mean at every date, the APL of ROW1 should be larger than the APL of PPT.[7]

To be concrete on this, let us compare the σ^2 estimate that is effective at the beginning of the forecast period from PPT-AR(1) with the σ^2 estimate from the AR(1) model on the pre-forecast period. The error variance estimates of AR(1) models are smaller on average by 17.45 percent (t-stat -2.42). A comparison of the APL values reveals that they increase on average by 8.91 percent (t-stat. 4.25) at horizon one, and by 8.42 percent (t-stat 3.26) at horizon four. The correlation coefficients between the series of percentage changes of the variances and of the APL are, as expected, negative:-0.21 (t-stat -0.98) for horizon one, and -0.29 (t-stat-1.41) for horizon four. These negative correlations support our previous explanation of why ROW1 performs better than PPT in terms of APL, though they are not much significant statistically (the p-values of the t-statistics are 0.33 and 0.17). Similar computations with KP-AR(1) instead of PPT10-AR(1) give similar results, with correlations of -0.21 (t-stat -0.72) at horizon one and -0.14 (t-stat -2.01) at horizon four.

6 Sensitivity Analyses

We perform two sensitivity checks. The first is with respect to the forecast period: we focus on the last three years of data, starting in 2007, quarter three, which corresponds more or less to the beginning of the great recession, until the end of the sample. The second check

[7] A similar argument applies if we compare the APL of PPT10 and PPT11 (and also PPT40 with PPT41). In PPT11, we allow zero or one break in the forecast period, whereas in PPT10 case, we allow no break, see Appendix A for details. Therefore the predictive densities of PPT11 are more dispersed by construction than the densities of PPT10. Hence if the observations of the forecast sample are not outliers in the predictive, and the predictive of both models have approximately the same mean at every date, the APL of PPT10 should be larger than the APL of PPT11. We find that this is the case on average for the AR(1) model at horizon one by 0.68 percent (t-stat 5.29) and at horizon four by 0.93 percent (t-stat 6.05); also for the AR(4) model: at horizon one, by 5.42 percent (t-stat 6.65), and at horizon four by 8.40 percent (t-stat. 14.4).

concerns the influence of the prior used in the break models.

6.1 Forecast performance since the middle of 2007

These results were obtained with the same prior as in the previous section. We focus on question 1 since for the other questions the previous answers are unchanged, with the exception that for question 2, using the RMSE criterion, PPT performs significantly better on average than KP at both horizons.

For the RMSE criterion, break models perform better than no-break models in about eighty percent of series at both horizons and on average (see the negative means in Table 11). These differences are significant on average at the five percent level, as the t-statistics in the table reveal. This is stronger than in the results for the last thirty percent of the sample (see subsection 5.1).

For the APL criterion, we find that break models perform better than no-break models in about fifty percent of series at horizon 1, and the (slightly negative) mean difference is not significant. At horizon four, break models dominate in about eighty percent of series and the mean difference (of almost +12 percent) is significant at the one percent level. These conclusions are different from what we found for the last thirty percent of the sample, where the no-break models, especially ROW, were clearly the winners (see subsection 5.2).

We can explain the improved performance of the break models with respect to ROW for the last twelve observations by the same argument as in subsection 5.3, but reversed. Estimated error variances (by ROW) increase at the end of the sample[8] due to the impact of the financial crisis, while break models do not capture this as much (few series have a break around mid-2007).

6.2 Impact of the prior for break models

In Bayesian inference, it is good practice to assess the sensitivity of the results with respect to the informative content of the prior. Thus we have computed again all the results with different sets of prior hyperparameters, one implying a more informative prior (PRIOR M), and the other a less informative prior (PRIOR L) than our intermediate prior (PRIOR I) used for getting all the results reported in the previous (sub)sections. The parameter values

[8] To get an idea of this, compare the two estimates of σ^2 for each series in the last column of Table 3.

Table 11: Performance comparison
on last twelve observations

	RMSE		APL	
% diff.	$h = 1$	$h = 4$	$h = 1$	$h = 4$
Mean	-15.8	-8.51	-0.12	11.96
t-stat	-2.40	-2.19	-0.07	3.58

Source: results available on request. Mean is the mean of percentage differences of the series.

of PRIOR I are given in Appendix A for the PPT model and in Appendix B for the KP model.

All our priors (M, I, L) imply that the unconditional prior expectations are equal to zero for the regression coefficients of the AR(1) or AR(4) equations in each regime since $E(\beta_j) = E[E(\beta_j|\beta_0)] = E(\beta_0)$ and the latter is set to zero. They imply non-existing second moments for the regression coefficients because $Var(\beta_j) = Var[E(\beta_j|\beta_0)] + E[Var(\beta_j|\beta_0)] = Var(\beta_0) + E(B_0)$ and $E(B_0)$ is not finite due to setting the degrees of freedom of the Wishart prior to $m + 1$, with $m = 2$ for AR(1) and 5 for AR(4). However $Var(\beta_0)$ is set to cI_m with $c = 1$ in PRIOR I and by changing the value of c, we can change the tightness of the prior on the regression coefficients.

In PRIOR L, we set $c = 100$, implying standard deviations equal to 10 for β_0, that is ten times larger than the corresponding value in PRIOR I (which has $c = 1$). We are also less informative on error variances of AR equations by setting $\underline{\rho} = 0.01$ and $\underline{d} = 0.01$ (instead of 0.1 for both in PRIOR I) in the PPT model. In the KP model, we set $\underline{V}_\omega = 100$ (instead of 1) and $\underline{\kappa_1} = \underline{\kappa_2} = 0.01$ (instead of 0.5).

In PRIOR M, we set $c = 0.01I_m$, implying a more precise prior (with standard deviations of 0.1) than in PRIOR I. For the other parameters of the prior, the values are the same as in PRIOR I.

Computed by simulation, the highest prior density interval of ninety percent level for each regression coefficient is equal to $(-17, +17)$ for PRIOR L, $(-3.9, +3.9)$ for PRIOR I, and $(-2.6, +2.6)$ for PRIOR M. Notice that if c is set to a smaller value than 0.01, the last interval does not shrink due to the $E(B_0)$ term that is not finite. Compared to the precisions typically implied by the type of data and sample size we use, all these priors are little informative, but PRIOR L is substantially less tight than the other two, while PRIOR M is slightly more concentrated than PRIOR I. In Table 12, we summarize the difference between the results with the three priors, for both criteria and for AR(1) specifications.

Table 12: Performance comparison of three priors for AR(1) models on last thirty percent of sample

		PRIOR M/PRIOR I			PRIOR L/PRIOR I		
horizon		PPT10	PPT11	KP1	PPT10	PPT11	KP1
		RMSE					
1	Mean	-0.10	-0.27	-0.02	0.79	0.51	-1.84
	t-stat	-0.24	-0.68	-0.03	0.91	0.76	-1.22
4	Mean	-0.48	-0.80	1.34	0.62	0.14	-2.76
	t-stat	-0.49	-0.76	1.03	0.77	0.28	-2.12
		APL					
1	Mean	-1.07	-1.07	1.00	-2.20	-2.42	3.12
	t-stat	-1.67	-1.61	0.82	-1.56	-1.77	3.30
4	Mean	-0.92	-0.67	-0.38	-2.34	-5.78	3.26
	t-stat	-1.16	-0.79	-0.23	-1.27	-4.42	3.26

Source: results available on request. Mean is the mean of percentage differences of all series.

For each series, horizon, and forecasting model (among PPT10, PPT11, and KP1), we compute the percentage difference in each criterion value (RMSE and APL) of PRIOR M and PRIOR L relative to PRIOR I. Then we take the average of these values over all series and we test the significance of the mean. For example, the *positive* mean of 0.51 for the RMSE criterion for PPT11 at horizon one indicates that on average the performance is better with PRIOR I than with PRIOR L, by half of a percent. The corresponding t-statistic (0.76) indicates that this is not significant even at the ten percent level. For the APL criterion, a *negative* mean such as -2.42 for PPT11 at horizon one indicates a better performance with PRIOR I than PRIOR L.

For the RMSE criterion, the differences of performance are tiny (nine out of twelve are under one percent) and statistically insignificant: the largest difference is at horizon four for KP1 (2.76 percent in favor of PRIOR L relative to I) and it is the single one that is significant (t-stat -2.12). Globally, for PPT models the mean differences suggest that a more informative prior reduces the RMSE, but of course this observation is conditional on the range of priors we have used.

For the APL criterion, the differences are slightly larger in favor of PRIOR I relative to

M (with one exception for KP1 at horizon one): they are close to one percent but none is significant at the ten percent level. For PRIOR L relative to I, they vary between two and six percent in favor of PRIOR I for PPT, and they are slightly over three percent for KP in favor of PRIOR L. The six t-statistics are larger than one and three of them are significant at the one percent level. Contrary to what we find for the RMSE, there is no evidence that a more (or less) informative prior improves the APL values.

7 Conclusion

In this paper, we have compared various forecasting procedures which allow for structural breaks in a wide variety of common US macroeconomic time series. Our set of forecasting procedures is divided into two groups: ones which formally model the break process (KP, PPT and TVP) and those which do not (rolling and recursive OLS forecasts, and UC-SV).

Our empirical results do not tell one single consistent story, but rather a variety of stories. Most importantly, we have added to the literature establishing the widespread existence of structural breaks in major macroeconomic time series. Our results also show the importance of using a forecasting method which allows for parameter change of some sort. However, perhaps unsurprisingly, we have not established that there is one single forecasting method that always is to be preferred. Each of our methods performs well in some cases, but not as well in others.

One of our findings is that, in terms of predictive likelihoods, it is often the case that rolling (fixed window) forecasts are even better than approaches which formally model the break process. In Section 5.3, we have offered an explanation for this. However, it is worthwhile to expand on this finding. In an effort to produce automatic forecasting procedures, suitable for repeated use with many data sets, this paper has used very simple implementations of the models of KP and PPT. In particular, for each series, we have used the same models (i.e. AR models), with the same prior (a relatively noninformative one) and the break process has been modelled in a very simple way. It is possible that the approaches of KP and PPT are not well-designed for use in such a black box fashion in such simple models. For instance, we have imposed that breaks in AR coefficients and error variance occur at the same time. But in some of the series, it looks to be the case that having separate break processes for the error variance and AR coefficients would be useful (i.e. ensuring more parsimony by allowing breaks in the conditional variance but not in the conditional mean). Also, it is likely that calibrating priors on a case-by-case basis (or using more sophisticated hierarchical priors as in KP) could improve forecast performance. And, the hierarchical structures of KP and PPT will tend to be of most use in more complicated forecasting models (e.g. involving many predictors or with VARs) where rolling or recursive forecasting methods can perform poorly (see, e.g., Korobilis and Koop (2010)) rather than simple univariate AR setups.

In sum, in this paper we have established the importance of structural breaks for forecasting in many macroeconomic time series. However, we also recommend the careful development of appropriate structural break models on a case-by-case basis as opposed to use of an automatic procedure.

Appendix A: Technical Details for PPT Approach

In this appendix, we describe posterior and predictive simulation as well as prior elicitation for our implementation of the PPT approach. Bauwens and Rombouts (2010) provide more details for posterior simulation and computing the marginal likelihood, which is used for choosing the number of breaks.

The model is defined by $y_t = Z_t\beta_{s_t} + \sigma_{s_t}\varepsilon_t$ as in (1) and by the break process which involves $S^T = (s_1, .., s_T)'$ where $s_t \in \{1, 2, .., K\}$ is a state variable and K is the number of in sample regimes. Notice that the last regime is an absorbing state over the sample period, but PPT relax this in the forecast period.

Priors

We use priors of the form:

$$\begin{aligned}
\beta_j|\beta_0, B_0 &\sim N_m(\beta_0, B_0), \\
\beta_0 &\sim N_m(\underline{\mu}_\beta, \underline{V}_\beta), \\
B_0^{-1} &\sim Wishart\left(\underline{\xi}, \underline{B}\right), \\
\sigma_j^{-2}|\upsilon_0, d_0 &\sim Gamma(\upsilon_0, d_0), \\
\upsilon_0 &\sim Gamma(\underline{\lambda}, \underline{\rho}), \\
d_0 &\sim Gamma(\underline{c}, \underline{d}) \\
p_i &\sim Beta\left(\underline{a}, \underline{b}\right).
\end{aligned}$$

In particular, in the forecasting exercise we set $\underline{\mu}_\beta = 0$, $\underline{V}_\beta = I_m$, $\underline{B} = 10I_m$, $\underline{\xi} = m + 1$ (where m is the dimension of Z_t), $\underline{\lambda} = 1$, $\underline{\rho} = 0.1$, $\underline{c} = 1$, $\underline{d} = 0.1$, and $\underline{a} = \underline{b} = 1$. This implies that all priors are proper but little informative.

Posterior simulator

The posterior simulation algorithm is a Gibbs sampler. Given initial conditions, the data, and in each block the other parameters, the sampling is done as follows:

1. Draw S^T using Chib's (1998) algorithm.

2. Draw p_i from $Beta\left(\underline{a}+T_i, \underline{b}+1\right)$ for $i=1, ..., K$, where T_i is the number of observations in regime i.

3. Draw $\beta_i|\sigma_i^2$ from Normal and $\sigma_i^2\ |\beta_i$ from Gamma, for $i=1,2,\ldots,K$.

4. Draw $\beta_0|B_0$ from Normal and $B_0^{-1}|\beta_0$ from Wishart.

5. Draw $d_0|v_0$ from Gamma and $v_0|d_0$ by numerical evaluation and inversion of its cdf.

Appendix B: Technical Details for KP Approach

In this appendix, we describe posterior and predictive simulation as well as prior elicitation for our implementation of the KP approach.

It is convenient to write equation (1) as $y_t = Z_t\beta_{s_t} + \exp\left(\omega_{s_t}/2\right)\varepsilon_t$. The transition probabilities between the states are defined in (5) so that the last diagonal element of the transition matrix is equal to p_K rather than one as in the PPT approach.

Priors

We use priors of the form:

$$\begin{aligned}
\beta_j &\sim N_m\left(\beta_{j-1}, B_0\right)\\
\omega_j &\sim N\left(\omega_{j-1}, \delta\right)\\
\beta_0 &\sim N_m\left(0, \underline{V}_\beta\right)\\
\omega_0 &\sim N\left(0, \underline{V}_\omega\right)\\
B_0^{-1} &\sim Wishart\left(\underline{\xi}, \underline{B}\right)\\
\delta^{-1} &\sim Gamma\left(\underline{\kappa_1}, \underline{\kappa_2}\right)\\
p_{i,i} &\sim Beta\left(\underline{a}, \underline{b}\right).
\end{aligned}$$

In particular, in the forecasting exercise we set $\underline{V}_\beta = I_m$, $\underline{V}_\omega = 1$, $\underline{B} = 10I_m$, $\underline{\xi} = m+1$, $\underline{\kappa_1} = \underline{\kappa_2} = 0.5$, and $\underline{a} = \underline{b} = 1$. This implies that all priors are proper but very little informative.

Posterior simulator

The posterior simulation algorithm is a Gibbs sampler. Given initial conditions, the data, and in each block the other parameters, the sampling is done as follows:

1. Draw S^T using Chib's (1998) algorithm.
2. Draw p_i from $Beta\left(\underline{a} + T_i, \underline{b} + 1\right)$ for $i = 1, ..., K$, where T_i is the number of observations in regime i.
3. Draw $[\beta_{s_t}]_{t=1}^{T}$ using the modified Kalman filter algorithm (see below).
4. Draw $[\omega_{s_t}]_{t=1}^{T}$ using the modified Kalman filter algorithm, after writing the model in appropriate linear state space form using the Kim, Shephard and Chib (1998) algorithm.
5. Draw B_0^{-1} and δ^{-1}, conditional on the draws of β_t and ω_t, using standard expressions.

Modified Kalman filter algorithm

Consider a state-space model of the following form:

$$y_t = z_t a_{s_t} + \varepsilon_t \tag{10a}$$

$$a_j = a_{j-1} + \eta_{s_t} \tag{10b}$$

$$\varepsilon_t \sim N\left(0, \gamma_1^2\right), \eta_j \sim N\left(0, \gamma_2^2\right)$$

conditional on knowing s_t, where (10a) is the measurement equation and (10b) is the state equation, with observed data y_t and unobserved state a_{s_t}. If the errors ϵ_t, η_t are *iid* and uncorrelated with each other, we can use the Kalman filter to estimate the state a.

Let $a_{t|s}$ denote the expected value of a_t and $P_{t|s}$ its corresponding variance, using data up to time s. Given starting values $a_{0|0}$ and $P_{0|0}$, the Kalman filter recursions provide us with initial filtered estimates:

$$\begin{aligned}
a_{t|t-1} &= a_{t-1|t-1} \\
P_{t|t-1} &= \begin{cases} P_{t-1|t-1} + \gamma_2^2 & \text{, if } s_{t-1} \neq s_t \\ P_{t-1|t-1} & \text{, otherwise} \end{cases} && (11) \\
K_t &= P_{t|t-1} z_t' \left(z_t P_{t|t-1} z_t + \gamma_1^2\right)^{-1} && (12) \\
a_{t|t} &= a_{t|t-1} + K_t \left(y_t - z_t a_{t|t-1}\right) \\
P_{t|t} &= P_{t|t-1} - K_t z_t P_{t|t-1}.
\end{aligned}$$

Once we reach the last period ($t = T$) we take the standard draw $a_{s_T} \sim N\left(a_{T|T}, P_{T|T}\right)$. If $s_T = T$ then a break occurs in each observation and we have a full tvp model, so that the Carter and Kohn smoother applies to all observations t. However with structural breaks models it will be the case that $s_T << T$ (i.e. the number of breaks is smaller than the number of observations, i.e. we do not have a full tvp model), we can only simulate a_j for $j = s_T + 1, ..., T$ (i.e. the "out-of-sample breaks" in a) using equation (10b). For $j = 1, ..., s_T$ we can use a standard smoother to get smoothed estimates. To do that, we run the backward recursions for $t = T - 1, ..., 1$:

$$\begin{aligned} a_{t|t+1} &= a_{t|t} + P_{t|t}P'_{t+1|t}\left(a_{t+1} - a_{t|t}\right), \text{ iff } s_{t+1} \neq s_t \\ P_{t|t+1} &= P_{t|t} - P_{t|t}P'_{t+1|t}P_{t|t}, \text{ iff } s_{t+1} \neq s_t \end{aligned}$$

and draw $a_{s_t} \sim N\left(a_{t|t+1}, P_{t|t+1}\right)$ iff $s_{t+1} \neq s_t$.

Appendix C: Predictive Simulator for PPT and KP models

Forecasting with no breaks out-of-sample (PPT model)

Since the PPT model implies that observations following T (the last sample date) are generated from $y_{T+h}|Y_{T+h-1}, \theta_K$ where $\theta_K = (\beta_K, \sigma^2_K)$, i.e. under the last operating regime, we can compute predictive densities as follows:

$$\begin{array}{l} p(y_{T+h}|s_{T+h} = K, s_T = K, Y_T) = \int \dots \int \prod_{j=0}^{h-1} p(y_{T+h-j}|Y_{T+h-1-j}, \theta_K) \\ p(\theta_K|\theta_0, S_T, Y_T)\, p(\theta_0|S_T, Y_T, \underline{A})\, p(S_T|Y_T)\, dy_{T+h-1} \dots dy_{T+1} d\theta_K d\theta_0 dS_{T-1}, \end{array} \tag{13}$$

where the integration is done with respect to S_{T-1} rather than S_T since $s_T = K$. This is implemented by simulation within the Gibbs sampler for the posterior density: for each Gibbs draw of θ_K, θ_0 and S_{T-1}, we generate sequentially future values $y_{T+1}, \dots, y_{T+h}$, each from $y_{T+h-j} \sim p(y_{T+h-j}|Y_{T+h-1-j}, \theta_K)$, and we keep y_{T+h} as a draw of the corresponding predictive density $p(y_{T+h}|s_{T+h} = K, s_T = K, Y_T)$. Doing this for e.g. $h = 4$ provides also the draws of the predictive densities for $h \leq 4$.

Forecasting with breaks out-of-sample (PPT & KP models)

The previous discussion does not allow for a break to occur in the forecast period. In order to allow in the PPT for the possibility of occurrence of one new regime after T, we lift the restriction $p_K = 1$ (something already done in the KP model) and extend the transition matrix to

$$\begin{pmatrix} p_1 & 1-p_1 & 0 & \dots & 0 & 0 & 0 \\ 0 & p_2 & 1-p_2 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & p_{K-1} & 1-p_{K-1} & 0 \\ 0 & 0 & 0 & \dots & 0 & p_K & 1-p_K \\ 0 & 0 & 0 & \dots & 0 & 0 & 1 \end{pmatrix}.$$

Additional regimes can be added by extending further the transition matrix, but here we consider the predictive density subject to one break occurring after date T. Assume that the break occurs at date $T+d$ where d can take any value in the set $\{1, 2, \dots, h\}$. For the predictive simulation of y_{T+h} with $h < d$ (the no-post-sample break case), we proceed as above. For $h \geq d$, the break occurrence implies that $y_{T+h} \sim p(y_{T+h}|Y_{T+h-1}, \theta_{K+1})$ where θ_{K+1} is the parameter characterizing the new regime, and is drawn from its hierarchical prior density $p(\theta_{K+1}|\theta_0)$. The observed sample does not provide information about θ_{K+1} and thus does not directly update this prior, but it does so indirectly by updating the prior information about θ_0 since this is drawn from its posterior distribution in the Gibbs sampler.

Assume first that $h = d = 1$. Then, given θ_0 (drawn in the Gibbs sampler), θ_{K+1} is drawn from $p(\theta_{K+1}|\theta_0)$ and given this draw, y_{T+1} is drawn from $p(y_{T+1}|Y_T, \theta_{K+1})$. This procedure is repeated at each iteration of the Gibbs sampler and delivers a sample of draws from the predictive density $p(y_{T+1}|s_{T+1} = K+1, s_T = K, Y_T)$.

Next assume that $h = 2$ and $d = 1$: y_{T+1} is simulated as explained just above, and y_{T+2} is drawn from $p(y_{T+2}|y_{T+1}, Y_T, \theta_{K+1})$ where y_{T+1} is set at its simulated value and θ_{K+1} is maintained to be the value used for drawing this y_{T+1}. For h larger than 2, one proceeds sequentially in the same way, i.e. freezing θ_{K+1} and using the simulated lagged values y_{T+h-j} ($j = 1, 2, \dots, h-1$) in the conditioning of $p(y_{T+h}|Y_{T+h-1}, \theta_{K+1})$.

Finally if $h \geq d \geq 1$, the values y_{T+j} for $j = 1, 2, \dots, d-1$ are sequentially simulated as in the no-post-sample break case. Then for $j = d$, θ_{K+1} is drawn from $p(\theta_{K+1}|\theta_0)$ and given this draw, y_{T+j} for $j = d, d+1, \dots, h$ are drawn sequentially. The next formula validates this simulation procedure for known break date $T+d$:

$$\begin{array}{l} p(y_{T+h}|\tau_K = T+d, s_{T+h} = K+1, s_T = K, Y_T) = \\ \int \ldots \int \prod_{j=0}^{h-1} p(y_{T+h-j}|Y_{T+h-1-j}, \theta_{K+1} 1_{\{h \geq d\}} + \theta_K 1_{\{h<d\}}) \\ p(\theta_{K+1}|\theta_0, S_T, Y_T)\, p(\theta_K|\theta_0, S_T, Y_T)\, p(\theta_0|S_T, Y_T, \underline{A})\, p(S_T|Y_T) \\ dy_{T+h-1} \ldots dy_{T+1} d\theta_{K+1} d\theta_K d\theta_0 dS_{T-1} \end{array} \tag{14}$$

where $1_{\{h \geq d\}}$ is equal to 1 if $h \geq d$ and 0 otherwise, and $1_{\{h<d\}} = 1 - 1_{\{h \geq d\}}$. To marginalize this density with respect to the break date d, we sum over all values of d as follows: $p(y_{T+h}|s_{T+h} = K+1, s_T = K, Y_T) =$

$$\begin{array}{l} \sum_{d=1}^{h} p(y_{T+h}|\tau_K = T+d, s_{T+h} = K+1, s_T = K, Y_T) \\ \qquad \times \Pr[\tau_K = T+d|s_{T+h} = K+1, s_T = K, Y_T] \end{array} \tag{15}$$

with $\Pr[\tau_K = T+d|s_{T+h} = K+1, s_T = K, Y_T] = p_K^{d-1}(1-p_K)/(1-p_K^h)$. Finally, we can integrate $p(y_{T+h}|s_{T+h}, s_T = K, Y_T)$ with respect to the number of post-sample breaks (0 or 1): $p(y_{T+h}|s_T = K, Y_T) =$

$$\begin{array}{l} p(y_{T+h}|s_{T+h} = K, s_T = K, Y_T) p(s_{T+h} = K|s_T = K, Y_T) \\ p(y_{T+h}|s_{T+h} = K+1, s_T = K, Y_T)[1 - p(s_{T+h} = K|s_T = K, Y_T)] \end{array} \tag{16}$$

where $p(s_{T+h} = K|s_T = K, Y_T) = p_{KK}^h$. This is simulated by drawing s_{T+h} from its discrete distribution, and then y_{T+h} from (13) if $s_{T+h} = K$ and from (15) if $s_{T+h} = K+1$. To sample the discrete distribution, we need a value of p_K. This is simulated in the Gibbs sampler from its full conditional posterior density, which is Beta($\underline{a} + T_K, \underline{b} + 1$), where T_K is the number of observations in regime K according to the sampled S_T vector.

As an example, to implement the simulation of y_{T+1}, we substitute (13), (15) and (14) in (16) and obtain $p(y_{T+1}|s_T = K, Y_T) =$

$$\begin{array}{l} p_K \int \ldots \int p(y_{T+1}|Y_T, \theta_K)\, p(\theta_K|\theta_0, S_T, Y_T)\, p(\theta_0|S_T, Y_T, \underline{A})\, p(S_T|Y_T) \\ \hfill d\theta_K d\theta_0 dS_{T-1} \\ +(1-p_K) \int \ldots \int p(y_{T+1}|Y_T, \theta_{K+1})\, p(\theta_{K+1}|\theta_0, S_T, Y_T)\, p(\theta_K|\theta_0, S_T, Y_T) \\ \hfill p(\theta_0|S_T, Y_T, \underline{A}) p(S_T|Y_T)\, d\theta_{K+1} d\theta_K d\theta_0 dS_{T-1}. \end{array}$$

This formula shows that the simulation for one predictive draw in the KP model, and the PPT model with the possibility of breaks occurring out-of-sample, is performed as follows:

1. Draw S_T, θ_0 and θ_K from the posterior (i.e. use a draw of the Gibbs sampler once it has converged).

2. Draw $p_K \sim Beta(\underline{a} + T_K, \underline{b} + 1)$.

3. Draw $s_{T+1} = K$ or $K + 1$ with respective probabilities $(p_K, 1 - p_K)$.

4. If $s_{T+1} = K$, draw $y_{T+1} \sim p(y_{T+1}|Y_T, \theta_K)$. If $s_{T+1} = K+1$, draw $\theta_{K+1} \sim p(\theta_{K+1}|\theta_0, S_T, Y_T)$ and $y_{T+1} \sim p(y_{T+1}|Y_T, \theta_{K+1})$.

If this is repeated as many times as one iterates in the Gibbs sampler, one obtains as many draws of the predictive of y_{T+1}. Generalizing this algorithm to $h \geq 2$ is not difficult but requires lengthy formulas.

Appendix D: Additional Tables

These tables may not be included in the paper. They are providing detailed results on which some tables in the paper are based. They are referenced in the paper, but these links can be removed without being harmful to the understanding of the paper.

Table 13: Root mean squared errors at horizon 1 on last thirty percent of sample

	series	PPT10	PPT11	PPT40	PPT41	KP1	KP4	TVP1	TVP4	ROW1	ROW4	REC1	REC4	UC-SV
1	GDPC96	0.989	0.991	0.947	**0.945**	0.994	0.967	0.999	0.965	1.048	1.019	1.007	*0.964*	0.608
2	GDPDEF	1.015	1.011	0.969	0.964	1.075	0.983	1.000	**0.944**	1.021	1.058	1.082	*0.969*	0.261
3	PCECC96	1.139	1.135	**0.962**	0.965	1.126	0.983	1.157	1.024	1.070	0.997	1.137	*0.989*	0.419
4	PCECTPI	1.027	1.030	1.007	1.003	1.029	1.029	**0.942**	0.963	*0.953*	0.969	1.037	1.008	0.439
5	GPDIC96	1.011	1.017	**0.996**	0.998	1.006	0.999	1.024	1.009	1.011	1.043	1.018	*0.999*	0.354
6*	OPHPBS	0.831	0.836	0.804	0.802	0.844	0.803	0.858	0.837	0.900	0.911	0.831	**0.800**	0.844
7	ULCNFB	0.896	0.898	0.824	0.823	0.941	0.829	**0.800**	0.804	*0.813*	0.826	0.930	0.827	1.275
8	CPIAUCSL	0.702	0.700	0.725	0.719	0.689	1.253	0.699	**0.675**	0.714	0.707	0.698	*0.676*	0.986
9	PPIFCG	0.730	0.730	0.705	0.698	0.721	1.301	0.722	**0.671**	0.754	0.728	0.729	*0.677*	2.499
10	TB3MS	0.937	0.939	0.936	0.936	**0.919**	1.004	1.033	1.019	*0.976*	1.057	1.051	1.068	0.381
11	GS10	0.844	0.848	**0.804**	0.808	0.849	0.808	0.849	0.830	0.847	0.810	0.847	*0.807*	0.433
12	M1SL	0.712	0.714	0.867	0.823	0.755	0.722	**0.711**	0.713	0.751	0.827	*0.713*	0.719	1.709
13	M2SL	0.759	0.757	0.672	*0.671*	0.760	1.290	0.762	0.690	0.781	0.704	0.759	**0.668**	0.957
14*	UTL11	0.817	0.812	**0.528**	0.529	0.768	0.549	0.631	0.687	0.793	0.552	0.771	*0.547*	0.151
15*	SP500	0.926	**0.922**	0.935	0.930	0.924	1.235	0.928	0.945	0.944	1.045	*0.923*	0.949	0.798
16	INDPRO	0.925	0.923	1.062	1.034	0.935	**0.911**	0.992	0.963	0.970	1.011	0.980	*0.925*	1.035
17*	HOUST	0.844	0.849	0.794	0.791	0.851	0.805	**0.663**	1.003	0.863	0.851	0.851	*0.806*	0.090
18	AHEMAN	0.884	0.885	0.925	0.923	0.972	0.911	**0.881**	0.879	*0.886*	0.911	1.123	0.958	0.319
19	UNRATE	0.966	0.966	0.967	0.969	0.976	**0.942**	1.117	1.135	1.001	1.063	*0.959*	0.974	0.235
20	PAYEMS	0.859	0.857	0.789	0.788	0.880	**0.769**	0.956	0.927	0.897	0.860	0.889	*0.866*	0.273
21	EXUSUK	0.898	0.898	*0.876*	0.879	0.900	0.880	0.906	0.903	0.921	0.902	0.894	**0.874**	0.406
22*	PMI	0.821	0.818	0.731	0.732	0.829	**0.721**	0.796	0.745	0.864	0.784	0.822	*0.722*	0.379
23*	NAPMNOI	0.866	0.863	0.823	0.821	0.864	*0.801*	0.879	0.820	0.909	0.892	0.863	**0.799**	0.568

See Table 4 for model definitions. Values in the last column are the RMSE for the UC-SV model. Values in other columns are the RMSE values for each model in the column header, divided by the value for the UC-SV model. For each series, the smallest value across all models is in bold. If this global minimum is in the category PPT+KP+TVP, the value in italics is the minimum across all other models. If the global minimum is in these other models, the value in italics is the minimizer across the PPT+KP+TVP models.

Table 14: Root mean squared errors at horizon 4 on last thirty percent of sample

	series	PPT10	PPT11	PPT40	PPT41	KP1	KP4	TVP1	TVP4	ROW1	ROW4	REC1	REC4	UC-SV
1	GDPC96	0.849	0.852	0.848	0.847	0.857	0.872	0.832	**0.824**	*0.863*	0.900	0.864	0.856	0.823
2	GDPDEF	1.116	1.105	1.032	1.028	1.172	1.020	1.046	1.012	1.148	1.011	1.208	1.035	**0.268**
3	PCECC96	1.031	1.041	0.948	0.944	1.035	**0.972**	1.000	0.984	1.066	1.057	1.040	*0.989*	0.550
4	PCECTPI	1.039	1.023	0.991	0.995	1.019	0.999	**0.855**	0.871	*0.857*	0.865	1.037	1.000	0.511
5	GPDIC96	**0.779**	0.779	0.789	0.789	0.782	0.784	0.781	0.781	0.823	1.010	*0.785*	0.795	0.497
6*	OPHPBS	0.824	0.822	0.820	0.822	*0.817*	0.817	0.854	0.860	0.850	0.857	0.819	**0.816**	0.869
7	ULCNFB	0.947	0.951	0.957	0.959	0.983	0.961	**0.921**	0.939	*0.925*	1.008	1.033	0.967	1.105
8	CPIAUCSL	**0.836**	0.840	0.920	0.884	0.847	0.842	0.837	0.837	*0.843*	0.871	0.838	0.843	0.886
9	PPIFCG	0.795	0.795	0.780	0.799	0.788	0.771	0.791	*0.777*	0.793	0.810	0.792	**0.775**	2.419
10	TB3MS	0.776	0.776	0.776	*0.772*	0.781	0.781	0.795	0.793	0.807	0.885	0.782	**0.771**	0.592
11	GS10	0.753	0.754	**0.750**	0.750	0.759	0.761	0.758	0.759	0.770	0.773	0.754	*0.752*	0.474
12	M1SL	0.855	**0.852**	1.003	0.917	0.876	0.893	0.854	0.875	0.860	0.889	*0.854*	0.888	1.508
13	M2SL	0.835	0.834	0.835	0.833	**0.832**	1.538	0.834	0.834	0.835	0.846	*0.833*	0.838	0.875
14*	UTL11	1.133	1.093	**0.937**	0.939	0.952	0.943	1.116	1.213	1.105	1.015	0.961	*0.952*	0.377
15*	SP500	0.824	0.828	1.085	0.953	0.829	0.832	**0.824**	0.829	0.843	0.885	*0.828*	0.830	0.949
16	INDPRO	0.813	0.810	1.028	0.848	0.834	0.806	0.773	**0.770**	0.964	0.893	*0.784*	0.785	1.868
17*	HOUST	0.981	**0.980**	1.009	1.006	0.983	1.040	1.226	1.395	1.405	1.374	*0.989*	1.051	0.230
18	AHEMAN	*0.816*	0.835	0.890	0.892	0.961	0.894	0.836	0.867	0.814	**0.813**	1.426	0.972	0.350
19	UNRATE	0.865	0.863	0.862	**0.860**	0.887	0.908	0.865	0.881	0.995	1.066	*0.845*	0.870	0.386
20	PAYEMS	0.994	0.994	0.969	0.962	1.009	0.959	0.932	**0.926**	1.151	1.100	*0.962*	1.020	0.559
21	EXUSUK	0.870	0.866	0.850	0.855	0.852	0.857	**0.845**	0.846	0.865	0.926	*0.850*	0.850	0.469
22*	PMI	0.825	0.817	0.754	0.761	0.815	*0.745*	1.026	0.989	0.893	0.905	0.805	**0.743**	0.743
23*	NAPMNOI	0.800	0.800	0.757	0.787	0.787	*0.723*	0.992	0.952	0.854	0.724	0.777	**0.711**	1.022

See Table 4 for model definitions. Values in the last column are the RMSE for the UC-SV model. Values in other columns are the RMSE values for each model in the column header, divided by the value for the UC-SV model. For each series, the smallest value across all models is in bold. If this global minimum is in the category PPT+KP+TVP, the value in italics is the minimum across all other models. If the global minimum is in these other models, the value in italics is the minimizer across the PPT+KP+TVP models.

Table 15: Average predictive likelihoods at horizon 1 on last thirty percent of sample

	series	PPT10	PPT11	PPT40	PPT41	KP1	KP4	TVP1	TVP4	ROW1	ROW4	REC1	REC4	UC-SV
1	GDPC96	**1.753**	1.734	1.680	1.663	1.008	1.678	1.724	1.661	1.747	1.657	1.327	1.348	0.289
2	GDPDEF	*3.068*	3.019	2.858	2.821	2.970	2.957	2.060	2.085	**3.306**	3.280	2.505	2.630	0.423
3	PCECC96	1.415	1.403	1.488	1.474	1.441	1.607	1.528	1.692	1.783	*1.932*	**1.425**	1.487	0.333
4	PCECTPI	2.264	2.247	2.408	2.391	2.370	*2.476*	1.893	1.916	**2.704**	2.586	2.328	2.490	0.388
5	GPDIC96	1.925	1.923	1.864	1.857	*1.978*	1.941	1.804	1.754	**2.354**	2.251	1.859	1.893	0.371
6*	OPHPBS	1.509	1.502	1.551	1.552	1.521	*1.576*	1.575	1.533	**1.626**	1.583	1.525	1.588	0.230
7	ULCNFB	1.403	1.408	1.370	1.355	1.337	1.393	1.454	*1.479*	**1.563**	1.492	1.275	1.375	0.218
8	CPIAUCSL	2.200	2.201	2.333	2.303	2.287	2.422	1.927	2.068	2.519	**2.608**	2.282	2.499	0.288
9	PPIFCG	1.558	1.561	1.597	1.599	1.582	1.666	1.738	*1.675*	1.508	1.545	1.615	**1.690**	0.157
10	TB3MS	2.142	2.117	2.106	2.083	*2.193*	2.022	2.131	2.016	**2.211**	2.097	1.178	1.218	0.378
11	GS10	*1.952*	1.927	1.932	1.900	1.783	1.871	1.619	1.708	2.110	**2.167**	1.882	1.922	0.335
12	M1SL	1.672	1.672	1.571	1.561	1.546	1.614	*1.712*	1.608	1.561	1.467	1.712	**1.742**	0.177
13	M2SL	1.796	1.796	1.942	*1.926*	1.812	1.146	1.876	**2.109**	1.934	2.054	1.808	1.983	0.249
14*	UTL11	6.347	6.268	*6.856*	6.687	4.916	6.191	1.562	1.568	6.744	**8.336**	4.846	6.199	0.466
15*	SP500	1.626	1.616	1.564	1.567	**1.645**	1.571	1.528	1.520	1.608	1.556	*1.634*	1.588	0.268
16	INDPRO	1.357	1.346	1.217	1.196	1.311	1.130	**1.439**	*1.437*	1.335	1.245	0.895	0.972	0.257
17*	HOUST	7.475	7.386	*7.583*	7.548	7.558	7.811	1.587	1.523	**9.639**	9.497	7.603	7.818	0.485
18	AHEMAN	**2.552**	2.519	2.266	2.240	2.361	2.255	1.772	1.856	*2.550*	2.513	1.616	1.941	0.374
19	UNRATE	2.785	2.749	2.622	2.603	2.845	*2.856*	2.166	2.172	**3.328**	3.135	2.674	2.637	0.433
20	PAYEMS	3.423	*3.368*	3.136	3.064	3.161	2.860	2.172	2.063	**3.370**	3.244	2.375	2.464	0.419
21	EXUSUK	2.008	1.998	1.864	1.866	*2.127*	1.928	1.900	1.902	**2.181**	2.042	2.031	2.019	0.388
22*	PMI	2.131	2.122	2.307	2.275	2.334	*2.403*	1.533	1.597	2.596	**2.775**	2.147	2.389	0.371
23*	NAPMNOI	1.873	1.875	1.894	1.887	1.891	*1.993*	1.430	1.418	1.989	1.915	1.895	**2.004**	0.310

See Table 4 for model definitions. Values in the last column are the average predictive likelihoods (APL) for the UC-SV model. Values in other columns are the APL values for each model in the column header, divided by the value for the UC-SV model. For each series, the largest value across all models is in bold. If this global maximum is in the category PPT+KP+TVP, the value in italics is the maximum across all other models. If the global maximum is in these other models, the value in italics is the maximizer across the PPT+KP+TVP models.

Table 16: Average predictive likelihoods at horizon 4 on last thirty percent of sample

	series	PPT10	PPT11	PPT40	PPT41	KP1	KP4	TVP1	TVP4	ROW1	ROW4	REC1	REC4	UC-SV
1	GDPC96	*2.590*	2.437	2.405	2.134	2.366	2.194	2.581	2.424	**2.601**	2.399	1.971	1.958	0.182
2	GDPDEF	3.584	3.229	3.712	3.340	3.207	*3.771*	2.486	2.663	3.995	**4.399**	2.580	3.301	0.256
3	PCECC96	2.185	2.106	2.127	2.020	2.183	2.184	*2.368*	2.246	**2.628**	2.448	2.210	2.223	0.198
4	PCECTPI	2.540	2.449	2.873	2.675	2.720	*2.885*	2.382	2.371	3.670	**3.734**	2.575	2.919	0.228
5	GPDIC96	3.231	3.127	3.156	2.945	*3.324*	3.191	3.119	3.070	**3.992**	3.762	3.187	3.189	0.211
6*	OPHPBS	2.070	2.023	2.093	1.956	*2.101*	2.087	2.055	1.936	**2.286**	2.161	2.092	2.113	0.165
7	ULCNFB	1.889	1.860	1.752	1.670	1.667	1.803	*1.973*	1.971	**2.066**	1.918	1.661	1.812	0.151
8	CPIAUCSL	*3.242*	3.167	3.142	2.960	3.269	3.216	2.733	2.793	**3.425**	3.217	3.345	3.275	0.184
9	PPIFCG	2.009	1.974	1.961	1.874	2.033	2.026	**2.186**	2.035	1.970	1.912	*2.064*	2.060	0.113
10	TB3MS	**3.099**	2.864	2.941	2.647	2.911	2.682	3.051	2.915	2.900	2.535	1.996	1.999	0.211
11	GS10	*3.047*	2.936	2.913	2.632	2.761	2.762	2.672	2.651	**3.278**	3.278	2.977	2.933	0.209
12	M1SL	*2.051*	2.001	1.857	1.749	1.889	1.904	2.030	1.973	1.932	1.781	2.101	**2.106**	0.137
13	M2SL	2.651	2.564	2.644	2.507	2.632	1.556	2.768	*2.845*	**2.878**	2.752	2.668	2.703	0.168
14*	UTL11	*4.959*	4.252	3.866	3.425	4.201	3.560	0.852	0.739	**5.042**	4.621	4.235	3.564	0.247
15*	SP500	**2.567**	2.485	2.450	2.312	2.550	2.505	2.315	2.275	2.494	2.222	*2.565*	2.530	0.164
16	INDPRO	1.830	1.729	1.667	1.506	1.720	1.566	**1.969**	1.883	*1.795*	1.641	1.361	1.382	0.142
17*	HOUST	6.753	6.181	5.865	5.341	*6.777*	5.882	0.911	0.865	7.298	**7.843**	6.859	5.959	0.254
18	AHEMAN	*4.375*	3.985	3.510	3.156	3.889	3.460	2.756	2.720	**4.405**	4.317	2.100	2.857	0.215
19	UNRATE	3.970	3.792	3.783	3.559	*4.300*	3.883	3.625	3.592	**5.090**	4.492	3.919	3.838	0.241
20	PAYEMS	**3.506**	3.214	2.877	2.542	3.106	2.657	2.106	1.885	*3.284*	2.788	2.474	2.292	0.246
21	EXUSUK	3.090	3.124	2.986	2.822	3.248	2.966	*3.381*	3.291	**3.365**	3.221	3.308	3.242	0.225
22*	PMI	2.289	2.152	2.216	2.034	*2.509*	2.334	1.065	1.054	**2.760**	2.277	2.381	2.392	0.196
23*	NAPMNOI	2.270	2.122	2.198	2.332	2.332	*2.376*	1.115	1.108	**2.474**	2.213	2.393	2.472	0.158

See Table 4 for model definitions. Values in the last column are the average predictive likelihoods (APL) for the UC-SV model. Values in other columns are the APL values for each model in the column header, divided by the value for the UC-SV model. For each series, the largest value across all models is in bold. If this global maximum is in the category PPT+KP+TVP, the value in italics is the maximum across all other models. If the global maximum is in these other models, the value in italics is the maximizer across the PPT+KP+TVP models.

References

Ang, A., and G. Bekaert (2002): "Regime switches in interest rates," *Journal of Business and Economic Statistics*, 20, 163–182.

Bauwens, L., and J. Rombouts (2010): "On marginal likelihood computation in change-point models," *Computational Statistics & Data Analysis*, in press.

Chib, S. (1998): "Estimation and comparison of multiple change-point models," *Journal of Econometrics*, 86, 221–241.

Clark, T., and M. McCracken (2009): "Improving forecast accuracy by combining recursive and rolling forecasts," *International Economic Review*, 50, 363–395.

Clements, M., and D. Hendry (1998): *Forecasting economic time series*. Cambridge University Press, Cambridge.

D'Agostino, A., L. Gambetti, and D. Giannone (2009): "Macroeconomic forecasting and structural change," ECARES working paper 2009-20.

Eklund, J., G. Kapetanios, and S. Price (2009): "Forecasting in the presence of recent and recurring structural change," Mimeo, Queen Mary University London, Department of Economics.

Giordani, P., and R. Kohn (2008): "Effcient Bayesian inference for multiple change-point and mixture innovation models," *Journal of Business and Economic Statistics*, 26, 66–77.

Koop, G., and S. Potter (2007): "Estimation and forecasting with multiple breaks," *Review of Economic Studies*, 74, 763–789.

——— (2009): "Prior elicitation in multiple change-point models," *International Economic Review*, 50, 751–772.

Korobilis, D., and G. Koop (2010): "Forecasting inflation using dynamic model averaging," Working paper available at http://personal.strath.ac.uk/gary.koop/.

Maheu, J., and S. Gordon (2008): "Learning, forecasting and structural breaks," *Journal of Applied Econometrics*, 23, 553–583.

Marcellino, M., J. H. Stock, and M. W. Watson (2006): “A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series,” *Journal of Econometrics*, 135, 499–526.

Meese, R., and J. Geweke (1984): “A comparison of autoregressive univariate forecasting procedures for macroeconomic time series,” *Journal of Business and Economic Statistics*, 2, 191–200.

Pesaran, M. H., D. Pettenuzzo, and A. Timmermann (2006): “Forecasting time series subject to multiple structural breaks,” *Review of Economic Studies*, 73, 1057–1084.

Pesaran, M. H., and A. Timmermann (2007): “Selection of estimation window in the presence of breaks,” *Journal of Econometrics*, 137, 134–161.

Stock, J. H., and M. W. Watson (1996): “Evidence on structural instability in macroeconomic time series relations,” *Journal of Business & Economic Statistics*, 14, 11–30.

——— (2007): “Why has U.S. inflation become harder to forecast?,” *Journal of Money, Credit & Banking*, 39, 3–33.