Robust forecast methods and monitoring during structural change

Jana Eklund
Barclays, London

George Kapetanios
Bank of England and Queen Mary College, London

Simon Price*
Bank of England, City University, London and CAMA, ANU.

February 11, 2013

Abstract

We examine how to forecast after a recent break. We consider a new approach, monitoring for change and then combining forecasts from two models, one using the full sample and the other solely data from after the identified break point. We compare this to some robust techniques: rolling regressions, forecast averaging over all possible windows and exponentially weighted forecasts. We examine the efficacy of these methods with Monte Carlo experiments where there are single deterministic or multiple stochastic location shifts, and for a large number of UK and US macroeconomic series. No single method is uniformly superior. Monitoring brings only small improvements in forecast performance, so that robust methods are preferred. In some cases, forecast averaging is the best option, with only a small loss of forecast performance in the absence of breaks.

Key words: monitoring, recent structural change, forecast combination, robust forecasts.

JEL Classifications: C100, C590.

1 Introduction

It is widely accepted that structural change is a crucial issue in econometrics and forecasting. By 'structural change', we mean an irregular, discreet and permanent change in a parameter of interest. Clements and Hendry argue forcefully (in e.g. 1998a,b) that it is the main source of forecast error; Hendry (2000) argues that the dominant cause of these failures is the presence of deterministic shifts; Stock and Watson (1996) looked at many forecasting models of a large number of US time series, and found evidence for parameter instability in a substantial proportion. Consequently there are many papers on the identification of breaks, and methods that are robust to them. But the fact that forecasters have to forecast after recent or during changes has received very little attention. Yet this is a pervasive and profound problem facing forecasters who need to generate projections in real time.

Dealing with breaks in a forecast context has two aspects. First, break detection; and subsequently the right forecasting strategy.

*Corresponding author: Bank of England, Threadneedle Street, London EC2R 8AH, UK: simon.price@bankofengland.co.uk. The views expressed are those of the authors, and not necessarily those of the Bank of England or Monetary Policy Committee.

The former activity has a long history. But the question of how to modify the forecasting strategy then arises. One influential contribution is by Pesaran and Timmermann (2007), who consider a number of alternative forecasting strategies in the presence of breaks. They conclude that pooling forecasts generated over a variety of estimation windows provides a reasonably good and robust forecasting performance.

Standard tests, by their nature, require some end-of-sample observations that are assumed to be free of breaks. Typically, between 5% and 15% of the sample size is held back in this way, so real-time detection of very recent breaks is simply impossible. A more fundamental problem, however, is that such tests are not designed for repeated applications. This acute real-time problem of break detection (where the hypothesis of interest is that there has been a recent break) has been tackled in the literature on structural change 'monitoring' (continuing repeated tests for breaks), but has not been integrated with the forecast problem. In this paper, we rectify that omission.

We address two important issues. First, we ask whether the forecaster should attempt to detect and then react to breaks, or instead adopt forecasting strategies that do not rely on break detection but are instead robust to them. Second, we examine two statistical environments. In one case, breaks are unique events (or are rare enough to be treated as such), and in the other they recur. These require different frameworks for analysis.

A new strategy that we explore involves monitoring (doing repeated tests for breaks) and then when a break is detected, combining two models, one using the full sample and the other only post-break data. Clark and McCracken (2009a) write that 'it is possible that using a sample window based on break test estimates could yield better model estimates and forecasts. In practice, however, difficulties in identifying breaks and their timing may rule out such improvements (see, for example, the results in [Clark and McCracken (2009b)]'. We evaluate this in a systematic way.

The alternative to the monitoring method described in the previous paragraph is to use robust models. We examine a set of widely advocated methods for forecasting in the presence of past breaks: model averaging over different estimation windows, rolling windows and exponentially weighted moving average (EWMA) models. Modifying Pesaran and Timmermann (2007), who are motivated by the desire to avoid the need to detect breaks, we consider the forecasting strategies they analyse in the context of recent breaks. Of all the strategies they consider, only forecast averaging translates easily to the current framework. Clark and McCracken (2009a), in their discussion of some related empirical results, write that in a forecast evaluation analysis, after 'aggregating across all models, horizons and variables being forecasted, it is clear that model averaging and Bayesian shrinkage methods consistently perform among the best methods. At the other extreme, the approaches of using a fixed rolling window of observations to estimate model parameters and discounted least squares estimation consistently rank among the worst.' By contrast, rolling regressions are advocated by Giacomini and White (2006). In another related paper, Pesaran and Pick (2011), building on Pesaran and Timmermann (2007), find that forecast averaging over windows is superior to a single estimation

window in almost all cases. They also consider EWMA estimators, and find the results are sensitive to the EWMA tuning parameter.[1]

In Section 2 we spell out our new monitoring approach for forecasting in the presence of recent breaks and describe some robust forecasting strategies. We then present Monte Carlo results in Section 3 in which these alternative methods are evaluated. We apply the methods to a large number of US and UK macroeconomic time series in Section 4, where we find results broadly consistent with the Monte Carlo study. Section 5 concludes.

2 Forecasting strategies

Our modelling framework can be summarised by the general model

$$y_t = \beta_t' x_t + u_t, \qquad t = 1, \ldots, T, \ldots \tag{1}$$

where x_t is a $k \times 1$ vector of predetermined stochastic variables, β_t is a $k \times 1$ vector of parameters and ϵ_t is a martingale difference sequence that is independent of x_t and has finite variance that may depend on t.

We specialise (1) by assuming that our entertained model is characterised by multiple structural breaks of the form

$$y_t = \sum_{i=1}^{b} \mathcal{I}\left(\{T_{i-1} < t \leq T_i\}\right) \beta_i' x_t + \epsilon_t, \qquad t = 1, \ldots, T_1, \ldots, T_b, \ldots, T, \ldots \tag{2}$$

where $\mathcal{I}(\mathcal{A})$ is an indicator variable taking the value one if the event $\mathcal{A}$ occurs and zero otherwise. T denotes the end of the observed sample. Since our main focus is real time forecasting we implicitly assume the existence of data after T. This straightforward model has been analysed extensively in the literature, e.g. by Bai and Perron (1998). The main point of departure from a standard analysis is to assume that some break dates are very close to the end of the sample at time T. The forecaster is aware of the possibility of a break in real time and either actively looks for such a break or adopts a forecasting strategy that is robust to the occurrence of such a break. The former is radically different to standard break detection as such methods cannot detect breaks if $T_b/T \to 1$ as $T \to \infty$.[2] We consider each in turn.

2.1 Forecasting strategies in the presence of a detected recent break

The seminal paper testing for a break at a known point was Chow (1960). Andrews (1993) introduced a methodology that allowed for unknown break-points: one influential paper is Bai and Perron

[1] Unlike us, they do not consider monitoring. They mainly examine forecasts of random walks subject to one-off breaks in the drift and volatility. This set up is effectively a location model, whereas in our applications we also examine parameter shifts in AR models. In our Monte Carlo study, we also explore multiple stochastic breaks as the forecasting period becomes large. This allows an analysis of the realistic scenario of repeated breaks, and provides clear results on the relative performance of competing forecasting methods.

[2] Most tests for breaks assume that $T_b/T \to C$ $T \to \infty$, where $C \in (0, 1)$.

(1998). All these methods test for breaks against specific alternatives. While effective in that case, they are ineffective when the break is not covered by the particular alternative. In addition, by their nature they require some trimming of observations towards the end of the sample. An alternative methodology was the CUSUM approach of Brown, Durbin, and Evans (1975).[3] Its advantage flows from the fact that there are many ways to reject the hypothesis of no structural change. Wald, LM and LR tests are efficient only against specific alternatives. The CUSUM test's usefulness lies partly in the fact that it offers a graphical view of deviations from constancy, but beyond that formal significance tests based on boundary conditions can be constructed for hypotheses likely to be observed in practice. Thus the method is more likely to be robust under different break scenarios. Moreover, there is no sample-trimming problem. However, after detection it is hard to detect the cause of the break.

Whether conventional or CUSUM, the tests are not designed for repeated applications. Consequently, this acute practical problem of break detection (where the hypothesis of interest is that there has been a recent break) has been tackled in the small literature on monitoring for structural change, pioneered by Chu, Stinchcombe, and White (1996). As the forecaster monitors in real time for breaks, she carries out repeated tests. This implies the need for an appropriate asymptotic framework, with critical values that ensure rejection probabilities remain bounded by the significance level when breaks do not occur. The original work has been refined by others, including Kuan and Hornik (1995), Leisch, Hornik, and Kuan (2000) and Zeileis, Leisch, Kleiber, and Hornik (2005). Groen, Kapetanios, and Price (2013) extend the analysis to panel data sets.

The CUSUM detector is specified as follows. In the regression $y_t = \beta' x_t + u_t$ let

$$\hat{\beta}_t = \left(\sum_{i=i}^{t} x_i x_i' \right)^{-1} \left(\sum_{i=1}^{t} x_i y_i \right) \tag{3}$$

be the OLS estimator at time t. Define recursive residuals as

$$\omega_t = \begin{cases} 0 & \text{for } \; t = k \\ \hat{\epsilon}_t / \nu_t^{1/2} & \text{for } \; t = k+1, ..., m, m+1, ... \end{cases} \tag{4}$$

with

$$\hat{\epsilon}_t = y_t - \hat{\beta}_{t-1}' x_t,$$
$$\nu_t = 1 + x_t' \left(\sum_{i=1}^{t-1} x_i x_i' \right)^{-1} x_t.$$

The t-th cumulated sum of recursive residuals is

$$Q_t^m = \hat{\sigma}^{-1} \sum_{i=k}^{n} \omega_i = \hat{\sigma}^{-1} \sum_{i=k}^{k+[(m-k)t]} \omega_i, \quad n = k+1, ..., m, m+1, ... \tag{5}$$

for $(n-k)/(m-k) \leq t \leq (n-k+1)/(m-k)$, where [.] denotes integer part and $\hat{\sigma}^2$ is some consistent estimate of σ^2. An obvious choice for this is the estimate of σ^2 based on the OLS estimate of β

[3] Extended to dynamic models by Krämer, Ploberger, and Alt (1988).

obtained in the initial (assumed break free) period $t = 1, ..., m$. Essentially, we construct our test statistic Q_t^m and if it exceeds some critical value we conclude there has been a structural break.

Regarding the critical value, it is well known (see, e.g., Krämer, Ploberger, and Alt (1988)) that under the null hypothesis, $H_0 : \beta_t = \beta$, for $t = m + 1, ...,$

$$\left\{ t \to m^{-1/2} Q_t^m, \quad t \in [0, \infty) \right\} \Rightarrow \left\{ t \to W(t), \quad t \in (0, \infty) \right\}, \tag{6}$$

where Q_t^m is defined in (5), $\Rightarrow$ denotes the weak convergence of the associated probability measures and $W(t)$ is a standard Brownian motion. This result can be used to motivate the following monitoring scheme

$$\lim_{m \to \infty} \Pr \left\{ |Q_t^m| \geq \sqrt{m} g(n/m), \quad \text{for some } n \geq m \right\} = \Pr \left(|W_j(t)| \geq g(t), \quad \text{for some } t \geq 1 \right) \tag{7}$$

where $W(t)$ is again a standard Brownian motion. In general, the probability on the right hand side (7) does not have a closed form solution for any arbitrary $g(t)$, but there are some specific instances where such a closed form solution is viable. We establish critical values by simulation.[4]

This allows us to monitor for breaks in real time. That then requires a strategy for forecasting when a recent break has been detected. Our new approach, which may be described as monitoring with the option of post-break detection of forecast averaging over different estimation periods (more succintly referred to as 'monitoring' below) is inspired by Pesaran and Timmermann (2007), who provide a detailed analysis of forecasting strategies when breaks occur in the more distant past. But the problem with recent breaks differs from that in their setup, as post-break data are by definition in short supply. As a result the first four of the following strategies suggested by Pesaran and Timmermann are either not straightforwardly applicable or infeasible. For reference, these are listed here: using model (2), estimated over post-window data; trading off the variance against the bias of the forecast by estimating the optimal size of the estimation window; estimating the optimal size of the estimation window using cross-validation;[5] combining forecasts from different estimation windows by using weights obtained through cross-validation as in the previous case; and simple average forecast combination over windows, using equal weights. Our proposal builds on this last suggestion but is tailored to the specific problem.

The forecaster monitors for breaks. As long as none are detected, the forecasts are produced using the model estimated over the whole sample.[6] Once the forecaster detects a break, it is assumed that the break has occurred at that point in time. Thus if $\hat{T}_1$ is the date the break is detected, it is also assumed to be the estimated date at which the break occurred.[7] The forecaster then makes

[4] Specifically, we generate realisations and count the number of times the absolute value of the Brownian motion exceeds g for at least some periods in the simulated sample.

[5] Cross-validation holds back observations at the end of the sample for a post-sample exercise, in this case to establish a minimum MSFE estimation window.

[6] Thus we assume that at the start of the monitoring period the forecaster has considered the possibility of past breaks which have been accommodated by some unspecified method, if found present. We accommodate this in the Monte Carlo design by assuming there is at most one break, and that the forecaster knows this.

[7] The delay in break detection is ignored as it is hard to estimate this bias. See Groen, Kapetanios, and Price (2013) for evidence on its extent.

two judgements, operationalised by the choice of two tuning parameters. The first defines the time elapsed before the model can be reliably estimated post-break. This parameter is referred to as $\underline{\omega}$ in Pesaran and Timmermann (2007) and we retain this notation. The second parameter is a window size $\overline{f}$ that the forecaster deems acceptable for the post-break model to be the sole model used for future forecasting. $\overline{f}$ is then chosen to be the period over which the forecasts of the post-break and the no-break models will be combined. In other words, forecasts will be combined for the period $\hat{T}_1 + \underline{\omega}$ to $\hat{T}_1 + \underline{\omega} + \overline{f}$. The forecasts after $\hat{T}_1 + \underline{\omega} + \overline{f}$ will therefore arise only from the post-break model.

There is a question of how the forecasts from the no-break (i.e., forecasts using all currently available data and ignoring the break) and post-break (using only post-break data) model are to be combined. It is natural that the post-break model should receive increasing weight as new data arrives. We therefore specify that the no-break model will be the sole model used prior to $\hat{T}_1 + \underline{\omega}$ and the post-break model will be the sole model used after $\hat{T}_1 + \underline{\omega} + \overline{f}$. A simple weighting scheme consistent with this choice is one where the weight for the post-break model increases linearly from zero prior to $\hat{T}_1 + \underline{\omega}$ to unity at $\hat{T}_1 + \underline{\omega} + \overline{f}$. That is, the weight for the post-break model at time $\hat{T}_1 + \underline{\omega} + j$ is $j/\left(\overline{f} + 1\right)$, whereas the weight for the no-break model is $1 - j/\left(\overline{f} + 1\right)$, where $j = 0, \ldots, \overline{f}$.

We assume that the forecaster knows there is only a single break. It is reasonable to argue that if breaks occur more frequently than assumed here, the model itself must come under scrutiny. A clear path for addressing this is to endogenise the break process into the model following, e.g., work by either Kapetanios and Tzavalis (2010) or Pesaran, Pettenuzzo, and Timmermann (2007). But an analysis of either course of action is beyond the scope of this paper.

In summary, our new proposal of forecasting based on monitoring is carried out by first constructing a CUSUM-based test, used at each point in time to determine whether a break has occured in real time, taking into account the fact that repeated tests need a specific set of critical values. Once a break is detected, the forecaster waits for a short period till enough data have accumulated after the presumed break date to estimate the model using only post-break data. Then the forecast is produced by combining the post-break model and the model estimated using the whole available sample. After a further sufficiently long period has elapsed after the break, only post-break data are used to estimate the model and produce the forecast. There is no theoretical guide to what the delay periods should be, so we adopt arbitrary values.

2.2 Forecasting strategies that are robust to the presence of a recent break

We recognise monitoring may in practice be problematic. Small breaks are hard to detect; it is not suitable where we expect frequent breaks; breaks are detected with a delay; and estimates of the timing are imprecise.[8] We therefore also consider common existing strategies robust to the presence of recent breaks, all of which involve downweighting past data when estimating the forecasting model.

[8] See Groen, Kapetanios, and Price (2013).

An alternative way to proceed is to disregard the structure in (2) and focus on a robust model such as a random walk or double-differenced model that may be biased but will be less affected by breaks, as Hendry (e.g., 2000) has often suggested. We ignore this approach, preferring to assume that the researcher has a specific view about the break process (2) and the structure of the model conditioning on x_t, (1). Given structural change. older data conveys less information about the current value of model coefficients than more recent data. The downweighting process then makes both intuitive and, given the theoretical analysis of Pesaran and Timmermann (2007), formal sense. The motivation is the familiar one of trading off bias and variance. A long sample reduces the variance of the forecast, but if there are breaks may increase the bias. For this reason, Pesaran and Timmermann (2007) demonstrate that out of sample MSFE may be minimised by choosing a window that opens after the break date. If the break date is known the benefits may be large. The problem is, of course, that the break date is normally unknown, especially for recent breaks, for reasons spelt out above.

We consider three straightforward and easily implementable methods. First, rolling windows offer a simple way to discount past data, where the weights for a window of length m are 0 for $t = 1, \ldots, T - m - 1$ and $1/m$ for $t = T - m, \ldots, T$. The second is based on estimating coefficients using exponentially weighted moving averages, in a regression context also known as discounted least-squares. A detailed description may be found in Harvey (1989). While less common in economics than in other forecasting domains, Clark and McCracken (2009a) observe there is evidence that it may be effective for macro data. The idea is that, unlike rolling windows where only a subset of available observations receive a non-zero weight in estimation, all available observations receive some weight; but older observations receive less. A parameter controls the rate of decline of weighting older observations, which plays a similar role to the rolling window size. The third, advocated by Pesaran and Timmermann (2007), is to combine forecasts using different estimates of the coefficients where these estimates are obtained using all possible contiguous subsets of observations that include the latest available observation.[9]

To be precise, we examine the mean square forecast error MSFE of a one-step-ahead[10] forecast based on a model estimated over the whole period. For the location model we consider the forecast based purely on lagged values of the variable of interest.

$$\hat{y}_{T+1|T} = \hat{\beta}_T, \qquad \text{where} \quad \hat{\beta}_T = \frac{\sum_{t=1}^{T} y_t}{T} \qquad \text{(Full-sample forecast)}, \tag{8}$$

[9] We do not consider time varying coefficient models as an approximation to the type of repeated discreet change we consider. Here the model (1) may be viewed as a measurement equation, augmented by a transition equation in terms of a vector of time varying parameters, β_t. Thus model (1) constitutes a state space model that can be analysed with widely available methods. In practice this can be a computationally intensive and time consuming process. In a multivariate setting it may be infeasible. For example, 10 explanatory variables would require 10 distinct unobserved processes for the time varying coefficients. Specifying and estimating such a model is demanding by most standards and more so if an empirical practitioner is considering several specifications. From a theoretical perspective, that state space model is bilinear, which may represent a stationary process, rather than one of structural change. Thus the time varying approach goes against the nature of the problem we try to address.

[10] We restrict attention on one-step forecasts partly for parsimony and partly because in macroeconomic series the one-step forecast is usually the most important: see evidence for this in Faust and Wright (2007), who show that getting the one-step ahead forecast (the nowcast) right is practically the overwhelmingly dominant issue in US Greenbook forecasts from FRB staff. But our approach can obviously be extended to iterated or direct multi-step forecasts.

versus one that is estimated from a method that discounts early data as discussed above.

$$\tilde{y}_{T+1|T} = \tilde{\beta}_T, \qquad \text{where} \quad \tilde{\beta}_T = \frac{\sum_{t=T-m+1}^{T} y_t}{m},\ m < T, \quad \text{(Rolling forecast)}, \tag{9}$$

$$\bar{y}_{T+1|T} = \frac{1}{T}\sum_{i=1}^{T} \tilde{y}^{(i)}_{T+1|T}, \qquad \text{(Forecast averaging over estimation periods)}, \tag{10}$$

where we denote $\tilde{y}_{T+1|T}$ for a rolling window of size m by $\tilde{y}^{(m)}_{T+1|T}$, and finally

$$\breve{y}_{T+1|T} = \sum_{t=1}^{T} \lambda\,(1-\lambda)^{T-t}\, y_t \qquad \text{(EWMA forecast)} \tag{11}$$

for some $0 < \lambda < 1$.

For the regression based methods, we construct the MSFE of a one-step-ahead forecast based on the OLS regression of the AR process $y_t = \beta' x_t + u_t$ where $x_t = (1, y_{t-1})'$, so that in each case

$$\hat{y}_{T+1|T} = \hat{\beta}' x_T \tag{12}$$

where in the benchmark the equation is estimated over the entire sample, and the estimates for the rolling forecasts and averages over estimation periods are constructed as for the location model.

For the the EWMA based least squares estimator of the regression $y_t = \beta' x_t + u_t$, $t = 1, \ldots, T$, is $\hat{\beta}_{EWMA} = \left(\lambda \sum_{t=1}^{T} (1-\lambda)^{T-t} x_t x_t'\right)^{-1} \lambda \sum_{t=1}^{T} (1-\lambda)^{T-t} x_t y_t$, where λ is a decay parameter. The choice of $0 < \lambda < 1$ is usually arbitrary. Harvey (1989) suggests that λ should lie between 0.05 and 0.3. This matters in practice: see for example Pesaran and Pick (2011). We examine two cases. The first sidesteps the choice by averaging forecasts using $\lambda = 0.1, 0.2, 0.3$, and in the second we use a value at the low end of the range, 0.05.[11] We refer to these as EWMAA (where the final 'A' indicates average) and EWMAL ('L' indicates low decay) respectively.[12]

2.3 Forecast evaluation

For a particular forecast $\hat{y}_t$ the one-step-ahead forecast error is defined as

$$\hat{\varepsilon}_{t+1} = \hat{y}_{t+1|t} - y_{t+1} \tag{13}$$

and associated root mean square forecast error over sample $t = 1, T$ as

$$RMSFE_T = \sum_{t=1}^{T} (\hat{\varepsilon}_t^2/T)^{0.5} \tag{14}$$

The relative RMSFE for model i relative to the benchmark 1 is defined as

$$RRMSFE_T = RMSFE_{i,T}/RMSFE_{1,T} \tag{15}$$

[11] Implying that the weight falls below 5% for lags greater than 60, one of the window lengths reported in the rolling and monitoring approaches.

[12] A more satisfactory way to choose the downweighting parameter would be to make it data dependent. One possibility is explored in Giraitis, Kapetanios, and Price (2013).

We report Diebold-Mariano (DM) tests, which test whether two models have equal forecasting ability. Unadjusted DM tests are inappropriate if the models are nested, but this is not the case here. They are carried out by constructing $\hat{\varepsilon}_{i,t}^2$ and $\hat{\varepsilon}_{j,t}^2$, the forecast errors of models i and j respectively. Then $d_{ij,t} = \hat{\varepsilon}_{i,t}^2 - \hat{\varepsilon}_{j,t}^2$, and the procedure amounts to testing that $E(d_{ij,t}) = 0$ while allowing for the possibility that $d_{ij,t}$ is serially correlated. A standard t-test statistic using a Newey-West correction is used. Further details are given in Diebold and Mariano (1995).

3 Monte Carlo analysis

In this section we present a Monte Carlo study of the forecasting performance of the strategies discussed in Section 2, using three designs. For the first we consider an AR(1) model subject to a single structural change, and in the second and third multiple stochastic breaks in a location and AR(1) model respectively.

The first case is designed for cases where the forecaster believes that breaks are rare, and in practice can be considered as unique events. As we argued in Section 2.1, it may then be reasonable to monitor for a break and react after detection by using a forecast combination strategy. Robust forecasting strategies are also applicable. The second and third designs allows frequent breaks to occur. Consequently, monitoring will not be a good strategy and is not considered.

For each experiment there are 500 Monte Carlo replications. All forecasts are one-step ahead. The benchmark forecast disregards the possibility of a break and uses an AR(1) model estimated over the whole available sample. We compare the forecasts in RRMSFE terms.

3.1 Design of experiments

For the autoregressive experiments, we use an AR(1) model:

$$y_t = \alpha_t + \rho_t y_{t-1} + u_t, \qquad t = 1, \ldots, T_0, \ldots, T_1, \ldots, T. \tag{16}$$

3.1.1 Deterministic single break

We begin with the specification of the single break case. Forecasting and break monitoring start at T_0, which we set to 100. The break occurs at T_1, which is set to 110, and occurs either in the autoregressive parameter or the intercept. These parameters take the value ρ_1 or α_1 up to T_1 and ρ_2 or α_2 thereafter.

That is, the actual data generation process is

$$y_t = \begin{cases} \alpha_1 + \rho_1 y_{t-1} + u_t, & t = 1, \ldots, T_1 - 1 \\ \alpha_2 + \rho_2 y_{t-1} + u_t, & t = T_1, \ldots, T \end{cases} \tag{17}$$

If the intercept or the autoregressive parameter are assumed constant they take the values $\alpha_1 = \alpha_2 = 0$ and $\rho_1 = \rho_2 = 0$ respectively. ρ_1 and ρ_2 take values from the set $\{-0.6, -0.4, -0.2, 0.2, 0.4, 0.6, 0.8\}$, while α_1 and α_2 take values from the set $\{-1.2, -0.4, 0.4, 0.8, 1.2, 1.6\}$.

For brevity we refer to the monitoring followed by averaging over available post-break windows method as 'monitoring'. Monitoring is assumed to cease when a break is detected.[13] Forecasting and evaluation (between T_0 and T) stops at $T = 150$. Averaging occurs during $\hat{T}_1+5$ to $\hat{T}_1+\overline{f}$, where $\overline{f}$ is arbitrarily set at 20 or 60 and $\hat{T}_1$ is the date at which the break is detected. The delay in $\hat{T}_1+5$ is set arbitrarily, but amounts to the effective minimum number of observations necessary to estimate the parameters of the simple models we examine. In practice, were monitoring employed, practitioners would experiment with varying parameters. The arbitrary choice of parameters is uncomfortable, but reflects the practical reality where choices are made in the light of experience.

The robust strategies we consider are: a rolling window where the size of the window is set to M at 20 and 60 periods; forecast averaging of forecasts obtained using parameters estimated over all possible estimation windows which we refer to as 'averaging'; and exponential weighted moving average estimation of the parameters.

3.1.2 Stochastic multiple breaks in a location model

We also examine multiple stochastic breaks. We begin with the simplest case, a location model. Although this appears restrictive, conceptually it can easily be extended to models with strictly exogenous regressors. The specification is based partly on Koop and Potter (2007) and Kapetanios and Tzavalis (2010).

$$y_t = \beta_t + u_t, \qquad t = 1, \ldots, T, \tag{18}$$

where

$$\beta_t = \sum_{i=1}^{t} \mathcal{I}\left(\nu_i = 1\right) w_i, \tag{19}$$

and ν_i is an i.i.d. sequence of Bernoulli random variables taking the value 1 with probability p and 0 otherwise. u_t and w_i are also i.i.d. series independent of each other and ν_i with finite variance denoted by σ_u^2 and σ_w^2 respectively. We assume

$$p = 0.5, 0.33, 0.2, 0.1, 0.05, 0.01$$

implying that breaks occur on average between every 2 and 100 periods. We set $u_t \sim N(0,1)$, and $w_t \sim iidU\left(w_l, w_u\right)$, where

$$\{w_l, w_u\} = \{-1, 1\}, \{-0.9, 0.9\}, \{-0.8, 0.8\}, \{-0.7, 0.7\}, \{-0.6, 0.6\}.$$

Other characteristics of the experiment are the same as for the more general case but the estimated model contains only a constant.

[13] Effectively, we are assuming the forecaster knows the structure of the model (in this as in other respects).

3.1.3 Stochastic multiple breaks in an AR(1) model

In an AR(1) model, for the multiple stochastic case either the autoregressive parameter or the autoregressive model's intercept change as follows:

$$\rho_t = \begin{cases} \rho_{t-1}, & \text{with probability} \quad 1-p \\ \eta_{\rho,t}, & \text{with probability} \quad p \end{cases}$$
$$\alpha_t = \begin{cases} \alpha_{t-1}, & \text{with probability} \quad 1-p \\ \eta_{\alpha,t}, & \text{with probability} \quad p \end{cases}$$

$p = 0.1, 0.05, 0.02, 0.01$ implying that the average duration between breaks varies between 10 and 100 periods. $\eta_{i,t} \sim iid\ U(\eta_{il}, \eta_{iu})$, $i = \rho, \alpha$, where

$$\{\eta_{\rho,l}, \eta_{\rho,u}\} = \{-0.8, 0.8\}, \{-0.6, 0.6\}, \{-0.4, 0.4\}, \{-0.2, 0.2\}$$

and

$$\{\eta_{\alpha,l}, \eta_{\alpha,u}\} = \{-2, 2\}, \{-1.6, 1.6\}, \{-1.2, 1.2\}, \{-0.8, 0.8\}, \{-0.4, 0.4\}.$$

When there are breaks in ρ, $\alpha = 0$, whereas for breaks in α, $\rho = 0$ (leaving the unconditional mean unchanged). The sample size is set to $T = 300$ and forecast evaluation starts at $t = 100$. Other aspects of the specification such as rolling window length are as in the single break case. As there are multiple breaks, only robust forecasting strategies are considered.

3.2 Results for single breaks

In the single break experiments where we are able to evaluate our monitoring approach, we consider breaks in either persistence or the mean. Table 1 reports the former for $\alpha = 0$. For monitoring, in some cases there are gains in forecast performance. However, in most cases the gains are very modest. There are no cases where monitoring leads to worse performance than the benchmark, so it is a conservative forecasting strategy, in the sense that it would tend to do somewhat better than the benchmark in some cases but will not lead to large forecast errors. But on this basis it is hard to recommend it over the full-sample benchmark.

The rolling window methods perform better than monitoring for large breaks. Where they do well, a short window improves the performance. But where they do worst, the opposite is the case. In general, longer windows offer a more conservative strategy. The forecast averaging method outperforms the longer period rolling window in most cases and where it does worse than the benchmark, does not do so by a large margin. In several cases it is best.

By contrast, although the averaged EWMA (EWMAA) does extremely well for some large changes, it does very badly for small changes or no structural change (along the diagonals). It is a risky strategy. The low-discount EWMA (EWMAL) is not so sensitive to small or large breaks. It lies somewhere between the short and long rolling window.

In Table 2 we consider a break in α. The results are qualitatively similar to those in Table 1.

3.3 Results for recurring breaks

We now examine recurring breaks. We exclude the monitoring method as it is inappropriate in this environment.

3.3.1 Location model

Results are reported for the location model in Table 3. The majority of best-performing cases are for the EWMAA. For low probability breaks the EWMAL performs best. The short rolling window is often better than the EWMAL in this parametrisation. Interestingly, the only case where the models fail to beat the full sample benchmark is with the EWMAA for the most infrequent breaks (average duration between breaks 100 periods) and smallest change.

Table 1: RRMSFE for alternative forecasting strategies; Single break in ρ; $\alpha = 0$

$\rho_1 \backslash \rho_2$	-0.6	-0.4	-0.2	0	0.2	0.4	0.6	0.8	-0.6	-0.4	-0.2	0	0.2	0.4	0.6	0.8
	Monitoring ($\overline{f} = 20$)								Monitoring ($\overline{f} = 60$)							
-0.6	1.00	1.00	1.00	1.00	1.00	0.99	0.97	0.92	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.95
-0.4	1.00	1.00	1.00	1.00	1.00	0.99	0.98	0.94	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.97
-0.2	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.95	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.97
0	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.97	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99
0.2	0.99	1.00	1.00	1.00	1.00	1.00	1.00	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99
0.4	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.6	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Rolling Window ($M = 20$)								Rolling Window ($M = 60$)							
-0.6	1.09	1.06	1.00	0.94	0.84	0.74	0.61	0.48	1.01	1.01	0.99	0.96	0.93	0.88	0.82	0.74
-0.4	1.06	1.09	1.06	1.01	0.94	0.82	0.70	0.54	1.01	1.02	1.01	0.98	0.96	0.91	0.84	0.76
-0.2	0.99	1.07	1.09	1.06	1.01	0.93	0.81	0.64	0.98	1.01	1.02	1.01	0.99	0.95	0.89	0.79
0	0.90	1.01	1.07	1.09	1.08	1.02	0.90	0.74	0.93	0.98	1.01	1.02	1.01	0.98	0.94	0.84
0.2	0.80	0.91	1.01	1.07	1.09	1.08	1.00	0.87	0.88	0.94	0.99	1.01	1.02	1.01	0.98	0.90
0.4	0.70	0.84	0.94	1.02	1.08	1.09	1.08	0.97	0.85	0.91	0.95	0.99	1.01	1.02	1.01	0.95
0.6	0.61	0.74	0.84	0.94	1.02	1.08	1.11	1.07	0.81	0.88	0.92	0.96	0.99	1.01	1.02	1.00
0.8	0.53	0.66	0.76	0.86	0.95	1.01	1.08	1.12	0.81	0.86	0.91	0.94	0.96	0.99	1.01	1.02
	Forecast Averaging								EWMAA							
-0.6	1.01	1.00	0.97	0.94	0.89	0.83	0.75	0.67	1.26	1.23	1.14	1.06	0.90	0.75	0.59	0.41
-0.4	1.00	1.01	1.00	0.97	0.94	0.87	0.80	0.70	1.22	1.26	1.23	1.14	1.05	0.87	0.71	0.51
-0.2	0.96	1.00	1.01	1.00	0.97	0.93	0.85	0.75	1.13	1.24	1.27	1.22	1.15	1.03	0.84	0.62
0	0.91	0.97	1.01	1.01	1.00	0.97	0.91	0.81	1.03	1.16	1.24	1.26	1.22	1.14	0.96	0.74
0.2	0.86	0.92	0.97	1.00	1.01	1.00	0.96	0.88	0.89	1.04	1.16	1.23	1.25	1.22	1.08	0.90
0.4	0.80	0.88	0.93	0.98	1.01	1.01	1.00	0.94	0.75	0.92	1.06	1.16	1.23	1.23	1.18	1.02
0.6	0.75	0.83	0.88	0.93	0.97	1.00	1.02	0.99	0.63	0.79	0.92	1.05	1.15	1.22	1.22	1.14
0.8	0.72	0.79	0.85	0.89	0.93	0.97	1.00	1.01	0.52	0.68	0.81	0.93	1.04	1.12	1.19	1.19
	EWMAL															
-0.6	1.04	1.02	0.98	0.91	0.84	0.76	0.64	0.49								
-0.4	1.02	1.04	1.02	0.97	0.91	0.82	0.71	0.55								
-0.2	0.97	1.02	1.04	1.02	0.98	0.89	0.79	0.63								
0	0.88	0.98	1.02	1.04	1.02	0.97	0.88	0.72								
0.2	0.80	0.91	0.98	1.03	1.04	1.01	0.95	0.82								
0.4	0.72	0.84	0.93	0.99	1.02	1.04	1.01	0.92								
0.8	0.66	0.76	0.86	0.93	0.99	1.02	1.03	0.99								
0.8	0.64	0.75	0.82	0.88	0.94	0.99	1.02	1.03								

Notes: EWMAA: Averaging EWMA forecasts with decay parameters of 0.1, 0.2 and 0.3; EWMAL: EWMA with decay parameter 0.05.

Table 2: RRMSFE for alternative forecasting strategies; Single break in α; $\rho = 0$

$\alpha_1 \backslash \alpha_2$	-1.2	-0.8	-0.4	0	0.4	0.8	1.2	1.6	-1.2	-0.8	-0.4	0	0.4	0.8	1.2	1.6
	Monitoring ($\overline{f} = 20$)								Monitoring ($\overline{f} = 60$)							
-1.2	1.00	1.00	0.99	0.95	0.90	0.86	0.83	0.82	1.00	1.00	0.99	0.97	0.95	0.93	0.92	0.91
-0.8	1.00	1.00	1.00	0.98	0.95	0.90	0.85	0.84	1.00	1.00	1.00	0.99	0.97	0.96	0.93	0.92
-0.4	0.98	1.00	1.00	1.00	0.99	0.95	0.90	0.86	0.99	1.00	1.00	1.00	0.99	0.97	0.95	0.93
0	0.95	0.99	1.00	1.00	1.00	0.99	0.95	0.91	0.98	0.99	1.00	1.00	1.00	0.99	0.98	0.95
0.4	0.90	0.95	0.99	1.00	1.00	1.00	0.98	0.95	0.95	0.98	0.99	1.00	1.00	1.00	0.99	0.98
0.8	0.86	0.90	0.95	0.98	1.00	1.00	1.00	0.99	0.93	0.95	0.98	0.99	1.00	1.00	1.00	0.99
1.2	0.83	0.87	0.90	0.95	0.98	1.00	1.00	1.00	0.92	0.93	0.95	0.97	0.99	1.00	1.00	1.00
1.6	0.81	0.83	0.86	0.89	0.94	0.99	1.00	1.00	0.91	0.92	0.93	0.95	0.97	0.99	1.00	1.00
	Rolling Window ($M = 20$)								Rolling Window ($M = 60$)							
-1.2	1.09	1.02	0.88	0.75	0.68	0.62	0.59	0.58	1.02	0.99	0.93	0.87	0.84	0.81	0.79	0.79
-0.8	1.02	1.09	1.03	0.88	0.75	0.67	0.61	0.59	0.99	1.01	0.99	0.93	0.88	0.83	0.81	0.79
-0.4	0.88	1.02	1.09	1.02	0.88	0.76	0.67	0.63	0.93	0.99	1.02	0.99	0.93	0.87	0.84	0.81
0	0.76	0.88	1.02	1.09	1.02	0.88	0.75	0.68	0.88	0.93	0.99	1.02	0.99	0.93	0.88	0.84
0.4	0.67	0.76	0.89	1.03	1.09	1.02	0.88	0.76	0.84	0.88	0.93	0.99	1.02	0.99	0.93	0.88
0.8	0.62	0.67	0.76	0.88	1.03	1.09	1.02	0.88	0.81	0.84	0.88	0.93	0.99	1.02	0.99	0.93
1.2	0.59	0.63	0.67	0.76	0.88	1.02	1.09	1.02	0.80	0.81	0.83	0.87	0.93	0.99	1.02	0.99
1.6	0.57	0.59	0.62	0.67	0.76	0.88	1.02	1.08	0.78	0.79	0.81	0.83	0.87	0.93	0.99	1.02
	Forecast Averaging								EWMAA							
-1.2	1.01	0.98	0.90	0.83	0.79	0.76	0.73	0.73	1.25	1.16	0.97	0.81	0.71	0.63	0.59	0.58
-0.8	0.98	1.01	0.98	0.90	0.84	0.79	0.75	0.74	1.17	1.26	1.17	0.97	0.81	0.70	0.62	0.59
-0.4	0.90	0.98	1.01	0.98	0.90	0.84	0.79	0.76	0.97	1.16	1.26	1.16	0.97	0.81	0.69	0.64
0	0.84	0.91	0.98	1.01	0.98	0.90	0.84	0.79	0.80	0.98	1.17	1.25	1.16	0.97	0.80	0.71
0.4	0.79	0.84	0.91	0.98	1.01	0.98	0.90	0.84	0.70	0.81	0.98	1.17	1.26	1.17	0.97	0.82
0.8	0.76	0.79	0.84	0.90	0.98	1.01	0.97	0.90	0.64	0.70	0.82	0.97	1.16	1.26	1.16	0.98
1.2	0.74	0.76	0.79	0.84	0.90	0.98	1.01	0.98	0.60	0.64	0.69	0.82	0.98	1.16	1.25	1.16
1.6	0.73	0.73	0.75	0.79	0.83	0.91	0.98	1.01	0.57	0.59	0.62	0.70	0.81	0.98	1.17	1.26
	EWMAL															
-1.2	1.04	0.98	0.87	0.77	0.70	0.66	0.64	0.63								
-0.8	0.98	1.04	0.98	0.86	0.77	0.70	0.67	0.65								
-0.4	0.87	0.98	1.04	0.98	0.87	0.77	0.71	0.67								
0	0.77	0.87	0.97	1.04	0.98	0.87	0.77	0.71								
0.4	0.71	0.77	0.87	0.98	1.04	0.98	0.87	0.77								
0.8	0.67	0.70	0.78	0.86	0.98	1.04	0.98	0.87								
1.2	0.64	0.66	0.70	0.77	0.87	0.98	1.04	0.98								
1.6	0.63	0.64	0.67	0.70	0.77	0.86	0.98	1.04								

Notes: EWMAA: Averaging EWMA forecasts with decay parameters of 0.1, 0.2 and 0.3; EWMAL: EWMA with decay parameter 0.05.

3.3.2 AR(1) model

Turning to a more realistic structure, Table 4 reports the results for recurring breaks in persistence ρ in an autoregressive model, for constant α. All the best cases are for forecast averaging and the EWMAL. In marked contrast to the location model results, the EWMAA never performs well. In all cases it performs worse than the alternative methods, and in many cases much worse. Small windows work best for large shifts, but as the size of the shift declines, the small rolling window performance deteriorates so that in most cases it cannot outperform the full sample estimates. The penalty from a short estimation period outweighs the gain from discounting the pre-break period. The higher window rolling case is more robust, in the sense that it both outperforms the full sample benchmark at low break probabilities for larger changes and is close to the benchmark for small changes and lower probabilities. However, the low discount variant EWMAL again performs well for the largest breaks, with a small-change penalty intermediate between the short and long rolling windows. But arguably, forecast averaging is dominant. It is no worse than the benchmark in the worst cases. Consequently the worst-case cost is small, and this method could therefore be described as conservative. In many cases it is the best performer. Forecast averaging never does worse than the benchmark, often does best among the models for smaller breaks, and while sometimes inferior to EWMAL for large breaks is not dramatically so. In that sense it emerges as a successful strategy.

The results in Table 5, where the intercept shifts, reveal less overall diversity, but in general the results are similar to those in Table 4. All the best cases are with forecast averaging or EWMAL. In this case, for smaller breaks the best performer is arguably the forecast average. Where it does worse than the full sample, it is by a small margin. As in Table 4, in no case is the EWMAA best, and tends to be worst, often by wide margins, increasing as the magnitude of changes declines. EWMAL is again best for the larger breaks, but worse than forecast averaging for smaller breaks. We conclude that although no method is unambiguously superior, forecast averaging has the edge over for smaller breaks and is robust in the sense that it is rarely much worse than the benchmark, that EWMAL is good for larger breaks, and that in most circumstances the EWMAA is a poor forecast model.

3.4 Summary

Thus we can draw some tentative conclusions. Results are sensitive to parameter choices, except for the average (where we simply use uniform weights over all possibilities). A monitoring and combination strategy will improve forecast performance and is unlikely to lead to major forecast errors relative to the full sample benchmark; in that sense it is a conservative strategy. But forecast improvements are small. Where we are confident moderately large breaks are likely to occur or are occurring infrequently, rolling windows can be useful. But they may be susceptible to poor forecast performance, the more so the shorter the window. Longer windows make for poorer performance for large breaks but better for small. The averaged EWMAA can provide very large improvements for large breaks but in general is a risky strategy to adopt as it can lead to large errors. The low

Table 3: RRMSFE for forecasting strategies (Location Model); Recurring breaks

$p\backslash$ u_l / u_u	−1 / 1	−0.9 / 0.9	−0.8 / 0.8	−0.7 / 0.7	−0.6 / 0.6	−1 / 1	−0.9 / 0.9	−0.8 / 0.8	−0.7 / 0.7	−0.6 / 0.6
	Rolling Window ($M = 20$)					Rolling Window ($M = 60$)				
0.5	0.18	0.22	0.21	0.25	0.30	0.38	0.40	0.40	0.42	0.43
0.33	0.22	0.22	0.27	0.30	0.37	0.39	0.39	0.42	0.47	0.48
0.2	0.27	0.33	0.33	0.39	0.47	0.45	0.46	0.47	0.51	0.57
0.1	0.41	0.45	0.48	0.55	0.61	0.52	0.57	0.59	0.65	0.70
0.05	0.57	0.61	0.68	0.71	0.76	0.64	0.67	0.71	0.77	0.81
0.01	0.88	0.90	0.93	0.95	0.97	0.89	0.89	0.93	0.94	0.96
	Forecast Averaging					EWMAA				
0.5	0.46	0.49	0.48	0.51	0.54	0.13	0.16	0.17	0.21	0.26
0.33	0.48	0.49	0.52	0.53	0.58	0.17	0.18	0.23	0.26	0.34
0.2	0.52	0.56	0.55	0.59	0.65	0.23	0.29	0.30	0.37	0.45
0.1	0.60	0.63	0.65	0.69	0.73	0.38	0.43	0.47	0.54	0.61
0.05	0.71	0.73	0.77	0.79	0.82	0.56	0.61	0.68	0.73	0.78
0.01	0.90	0.91	0.93	0.94	0.96	0.91	0.94	0.97	0.99	1.01
	EWMAL									
0.5	0.23	0.23	0.27	0.29	0.32					
0.33	0.25	0.27	0.29	0.34	0.39					
0.2	0.31	0.35	0.39	0.44	0.50					
0.1	0.42	0.46	0.53	0.56	0.64					
0.05	0.56	0.61	0.66	0.72	0.77					
0.01	0.85	0.87	0.92	0.92	0.96					

Notes: EWMAA: Averaging EWMA forecasts with decay parameters of 0.1, 0.2 and 0.3; EWMAL: EWMA with decay parameter 0.05.

discount EWMAL is comparable to an intermediate rolling window length or averaging in some respects. Overall, the forecast averaging method emerges as a good compromise between improved forecast performance in the face of large breaks and modest costs in other cases. It also has the advantage of being free of the necessity to make a parameter choice.

4 Empirical application

In this section we examine how our methods would have fared when applied to a large range of UK and US quarterly data series.[14] We are not trying to develop the best methods for particular data sets, but instead trying to get an impression of whether the issues identified above are important in practice. In all cases we transform series to stationarity and employ AR(1) forecasting models. For the UK, we use data on 94 series spanning 1977Q1 to 2008Q2, and examine two forecast evaluation sub-periods within this (1992Q1 to 1999Q4 and 2000Q1 to 2008Q2). For the US, we have data on 97 series from 1960Q1 to 2008Q3, and examine three forecast evaluation sub-periods (1975Q1 to 1986Q2, 1986Q3 to 1997Q4, and 1998Q1 to 2008Q3). For each series, we compare RMSFEs to that from an AR(1) benchmark. The methods we report relate to those in the Monte Carlo study, and are monitoring

[14] We take no account of real-time data revisions.

Table 4: RRMSFE for forecasting strategies (AR Model); Recurring breaks in ρ; $\alpha = 0$

$p\backslash$ $\eta_{\rho,l}$ / $\eta_{\rho,u}$	−0.8 / 0.8	−0.6 / 0.6	−0.4 / 0.4	−0.2 / 0.2	−0.8 / 0.8	−0.6 / 0.6	−0.4 / 0.4	−0.2 / 0.2
	Rolling Window ($M = 20$)				Rolling Window ($M = 60$)			
0.1	0.97	1.04	1.07	1.09	1.00	1.01	1.02	1.02
0.05	0.93	1.01	1.06	1.09	0.96	1.00	1.01	1.02
0.02	0.90	1.00	1.05	1.09	0.93	0.97	1.00	1.02
0.01	0.91	1.02	1.06	1.09	0.91	0.97	1.00	1.02
	Forecast Averaging				EWMAA			
0.1	0.95	0.98	1.00	1.01	1.02	1.14	1.21	1.25
0.05	0.93	0.97	0.99	1.01	1.00	1.12	1.20	1.25
0.02	0.91	0.96	0.99	1.00	0.99	1.12	1.20	1.25
0.01	0.91	0.97	0.99	1.00	1.02	1.16	1.22	1.25
	EWMAL							
0.1	0.92	0.99	1.02	1.04				
0.05	0.89	0.97	1.01	1.04				
0.02	0.87	0.96	1.01	1.03				
0.01	0.90	0.96	1.01	1.04				

Notes: EWMAA: Averaging EWMA forecasts with decay parameters of 0.1, 0.2 and 0.3; EWMAL: EWMA with decay parameter 0.05.

using 40 and 60-period windows (M40 and M60),[15] rolling-window forecasts using 40 and 60-period windows (R40 and R60), averaging across estimation periods (AV) and the exponentially weighted moving average (EWMA). Forecast combination is well known to often improve performance, so we also report the simple average of the robust methods (ROBAV). Detailed results and definitions for each series are given in Eklund, Kapetanios, and Price (2010).

4.1 UK results

An obvious prior question to ask is whether there is evidence of structural breaks in the series we examine. So we begin by performing Bai and Perron (1998) tests for structural breaks (shifts in either constant or autoregressive parameter for an AR(1) process), reported in Table 6.[16] We identify 33 series containing breaks out of the total, so this suggests that structural change was indeed an important issue in the UK over this period. It should be clear that this test uses the full sample and this information would not be available in real time. We also note that conventional significance tests are conservative, designed as they are for inference and low Type 1 errors. For forecasters, the question is not so much are we sure that there has been a break, as is there evidence of a break that might lead to poor forecast performance.

The full set of results is provided in the working paper version of this paper, Eklund, Kapetanios, and Price (2010) for the two periods we examine. They are summarised in Table 7. For the entire

[15] We use a window of 40 observations rather than 20 for the empirical application as it corresponds to a 10-year estimation period.

[16] To be precise, we use the sequential estimation method described in Section 5.2.2 and Proposition 8 of Bai and Perron (1998). When that method finds one or more breaks we say that the relevant series has a break.

Table 5: RRMSFE for forecasting strategies (AR Model); Recurring breaks in α; $\rho = 0$

$p\backslash$ $\eta_{\alpha,l}$ / $\eta_{\alpha,u}$	−2 / 2	−1.6 / 1.6	−1.2 / 1.2	−0.8 / 0.8	−0.4 / 0.4	−2 / 2	−1.6 / 1.6	−1.2 / 1.2	−0.8 / 0.8	−0.4 / 0.4
	Rolling Window ($M = 20$)					Rolling Window ($M = 60$)				
0.1	1.04	1.04	1.04	1.05	1.08	1.02	1.01	1.02	1.01	1.02
0.05	0.94	0.95	0.98	1.02	1.07	0.99	0.99	0.99	1.00	1.02
0.02	0.84	0.88	0.92	0.99	1.06	0.91	0.93	0.94	0.97	1.01
0.01	0.84	0.87	0.93	0.99	1.07	0.88	0.89	0.93	0.97	1.01
	Forecast Averaging					EWMAA				
0.1	0.97	0.97	0.98	0.99	1.00	1.06	1.06	1.10	1.16	1.23
0.05	0.93	0.94	0.95	0.97	1.00	0.97	0.99	1.05	1.13	1.22
0.02	0.88	0.90	0.92	0.96	0.99	0.91	0.96	1.02	1.12	1.22
0.01	0.87	0.89	0.92	0.96	1.00	0.93	0.97	1.05	1.13	1.23
	EWMAL									
0.1	0.96	0.97	0.98	1.00	1.03					
0.05	0.91	0.91	0.93	0.98	1.02					
0.02	0.84	0.85	0.90	0.95	1.01					
0.01	0.82	0.86	0.89	0.96	1.01					

Notes: EWMAA: Averaging EWMA forecasts with decay parameters of 0.1, 0.2 and 0.3; EWMAL: EWMA with decay parameter 0.05.

set of series, we report the mean, the median (giving some indication of skewness), the minimum and maximum, the standard deviation, and skewness of the relative RMSFE. We also report the number of cases in which Diebold-Mariano tests reject equality of performance between a robust method and the full-sample (FS) null at 5% in favour of the robust method (DM(R)), while DM(FS) rejects against the robust method, again at 5% significance level.

In all cases except the EWMAA the mean and median relative RMSFEs are no greater than one. The mean lies uniformly below the medians, consistent with the typically negative skewness, meaning that the gains where there is a break exceed the losses when the series are stable. In the Monte Carlos, we found the results are sensitive to parameter choice and callibrations, and this is the case in the empirical results. On the mean and median criteria in both periods the minima are delivered by forecast averaging, followed by the EWMAL. The EWMAA is not only the worst performer, but on average fails to beat the full sample AR, although in some cases it does extremely well (indicated by the very low values in the 'Minimum' rows). The average of the robust methods ROBAV turns out to have roughly the average RRMSFE performance, a result repeated in the other cases, so in this case averaging does not improve over the best methods. It may be relevant that averaging is itself a robust method, so may be adding little value by robustifying already robust methods. The monitoring method on average beats the benchmark, with a 40 period window outperforming 60 periods. The rolling window does better, especially for the 40 period window. The rolling regressions also deliver low minima, especially for the shorter window. However, if the forecaster gives a high weight to avoiding extreme forecast errors, then using the monitoring method may be the best strategy. The maximum RRMSFE are close to unity in that case, and the variation

Table 6: Series with Identified Breaks: UK data

Private sector output growth	1
GFK index score	1
Stock of net corporate debt	1
Nominal wages per worker	1
Sectoral M4	1
M4 liabilities to private non-financial corporations	1
Net lending to household sector	2
GDP	1
Gross National Income	1
Manufacturing	1
Manufacturing of textile & textile products	1
Manufacturing of leather & leather products	1
Manufacturing of wood & wood products	1
Manufacturing of non-metallic mineral products	1
Manufacturing of basic metals & fabricated prod	2
Manufacturing of electrical & optical equipment	1
Distribution, hotels & catering; repairs	1
Output Index: Total	1
Total adjustment to basic prices	1
GDP at market prices	1
Gross Value Added at factor cost	1
Money stock M4 (end period)	3
Notes & coins in circulation outside Bank of England	1
Total Government benefits paid to household sector	3
General Government: Final consumption expenditure	2
Household final consumption expenditure	2
Durable goods	1
Claimant count rate	1
Whole economy, inc bonus: % change 3 month average	1
Unemployed	3
Economically active	1
Total actual weekly hours worked	2
Imports: Total trade in goods and services excl MTIC fraud	1

in the RRMSFE also smallest. The EWMAA, by contrast, is worst on this criterion. Of the other methods, in the first period the smallest maximum error is for the 60 period rolling window, but the average is only slightly higher. On the formal tests, in the first period the ROBAV is best, followed by the rolling 40 period, and then the forecast average. With the exception of the EWMAA which is selected only slightly more often than would be expected by chance, the other methods are closely comparable. The EWMAA is significantly outperformed by the full-sample forecasts in more cases than it outperforms: by contrast, except for the rolling cases there are only one or two rejections for the other methods. Similar results hold for the second period.

Table 7: Summary for Empirical Results for UK

	M40	M60	R40	R60	AV	EWMAA	EWMAL	ROBAV
First Period (1992Q1 - 1999Q4)								
Mean	0.978	0.984	0.957	0.975	0.918	1.054	0.925	0.925
Median	1.000	1.000	0.974	0.984	0.951	1.056	0.958	0.950
Minimum	0.607	0.692	0.118	0.792	0.155	0.010	0.096	0.093
Maximum	1.050	1.031	1.525	1.235	1.265	2.228	1.229	1.305
Std. Dev.	0.058	0.043	0.170	0.085	0.157	0.300	0.159	0.158
Skewness	-3.783	-4.239	-0.725	0.383	-1.429	0.155	-1.866	-1.605
DM(R)	14	14	18	16	17	6	16	20
DM(FS)	2	2	4	4	1	8	2	2
Second Period (2000Q1 - 2008Q2)								
Mean	0.972	0.980	0.925	0.959	0.903	1.029	0.914	0.905
Median	1 1.000	1.000	0.959	0.987	0.949	1.096	0.963	0.954
Minimum	0.619	0.737	0.006	0.005	0.047	0.005	0.007	0.009
Maximum	1.040	1.025	1.511	1.514	1.301	1.622	1.350	1.354
Std. Dev.	0.065	0.044	0.238	0.218	0.189	0.317	0.203	0.207
Skewness	-2.806	-2.819	-0.676	-0.636	-1.182	-0.525	-1.101	-1.033
DM(R)	12	12	16	16	22	8	19	19
DM(FS)	1	2	2	8	1	9	1	3

Notes: The table reports summary statistics on the set of Relative RMSFEs for alternative forecasting methods for all series. M60: Monitoring using a 60-period window; M40: Monitoring using a 40-period window; R40: Rolling Forecast using a 40-period window; R60: Rolling Forecast using a 60-period window; AV: Averaging across estimation periods; EWMAA: Exponentially Weighted Moving Average (Averaging 3 EWMA forecasts with decay parameters of 0.1, 0.2 and 0.3); EWMAL: EWMA with decay parameter given by 0.05; ROBAV: simple average of the robust methods. DM(R) is the number of series for which the Diebold-Mariano test rejects in favour of the given robust method at the 5% significance level, while DM(FS) is the number of series for which the Diebold-Mariano test rejects against the given robust method at the 5% significance level.

We conclude that over these periods forecast averaging would have been a good strategy, although EWMAL, rolling regressions and monitoring would also have improved forecast performance. However, monitoring would have been a relatively conservative strategy, again in the sense that it would on average offer a small advantage over using the full sample and avoids making large forecast errors, while not offering large improvements in performance. This reflects the difficulty of detecting structural breaks.

Table 8: Series with Identified Breaks: US data

Industrial Production: Consumer Goods	1
Unemployment Rate: All Workers	1
Civilians Unemployed - 15 Weeks & Over	1
1-Year Treasury Constant Maturity Rate	1
Total Reserves of Depository Institutions	1
S&P 500 Finance Total return Index	1

4.2 US results

For the US, far fewer breaks are identified (Table 8), although we note our remarks made above about confidence levels and forecasting. Notwhistanding this, given the weaker evidence for breaks it is not surprising that there are fewer gains to using the methods (Table 9), although there are gains, and relatively more in the third period. Forecast averaging no longer unambiguously emerges as the best average performer, but is nevertheless most often best and the best performer on the formal tests. EWMAA remains both the worst on average and the most variable performer, with the best and worse individual forecasts in each period. The monitoring methods remain conservative in the sense we identified in the UK (small average gains and avoiding very poor performance). However, there was little formal statistical evidence that any model would have helped forecast these series, with the exception of forecast averaging in the third period, where there is some weak evidence in favour.

Table 9: Summary for Empirical Results for US

	M40	M60	R40	R60	AV	EWMAA	EWMAL	ROBAV
First Period (1975Q1 - 1986Q2)								
Mean	1.011	1.005	1.033	1.012	1.032	1.221	1.043	1.033
Median	1.000	1.000	1.033	1.007	1.034	1.212	1.046	1.031
Minimum	0.872	0.905	0.906	0.937	0.889	0.792	0.842	0.899
Maximum	1.171	1.106	1.135	1.355	1.291	2.594	1.440	1.271
Std. Dev.	0.032	0.020	0.042	0.042	0.057	0.212	0.069	0.050
Skewness	0.689	0.077	-0.266	5.515	0.455	2.695	1.594	0.547
DM(R)	2	2	0	0	0	0	0	0
DM(FS)	3	1	7	3	12	10	9	10
Second Period (1986Q3 - 1997Q4)								
Mean	0.990	0.991	0.999	1.040	0.987	1.145	0.987	0.994
Median	1.000	1.000	0.999	1.029	1.008	1.161	1.004	1.010
Minimum	0.815	0.870	0.641	0.798	0.711	0.583	0.686	0.731
Maximum	1.092	1.054	1.284	1.414	1.113	1.732	1.150	1.136
Std. Dev.	0.032	0.024	0.100	0.101	0.070	0.227	0.081	0.070
Skewness	-2.027	-2.210	-0.400	0.978	-1.528	0.014	1.019	-1.347
DM(R)	3	4	2	2	7	3	6	4
DM(FS)	4	2	1	14	2	3	1	1
Third Period (1998Q1 - 2008Q3)								
Mean	0.998	0.991	1.002	0.977	0.952	1.307	0.984	0.960
Median	1.000	1.000	1.025	0.997	0.969	1.104	0.982	0.984
Minimum	0.842	0.877	0.311	0.324	0.513	0.333	0.356	0.376
Maximum	1.623	1.052	2.557	1.626	1.113	15.818	2.384	1.423
Std. Dev.	0.073	0.028	0.212	0.139	0.093	1.540	0.189	0.120
Skewness	6.153	-2.109	3.664	-0.259	-1.595	8.675	3.666	-0.985
DM(R)	1	3	5	3	13	1	4	6
DM(FS)	0	0	6	5	0	6	2	1

Notes: The table reports summary statistics on the set of Relative RMSFEs for alternative forecasting methods for all series. M60: Monitoring using a 60-period window; M40: Monitoring using a 40-period window; R40: Rolling Forecast using a 40-period window; R60: Rolling Forecast using a 60-period window; AV: Averaging across estimation periods; EWMAA: Exponentially Weighted Moving Average (Averaging 3 EWMA forecasts with decay parameters of 0.1, 0.2 and 0.3); EWMAL: EWMA with decay parameter given by 0.05; ROBAV: simple average of the robust methods. DM(R) is the number of series for which the Diebold-Mariano test rejects in favour of the given robust method at the 5% significance level, while DM(FS) is the number of series for which the Diebold-Mariano test rejects against the given robust method at the 5% significance level.

5 Conclusions

A common source of forecast failure is the existence of structural breaks in the data generating process. One characterisation of a break is an abrupt parameter shift. In that context, a natural strategy for a forecaster operating in real time might be to monitor for a break, and then to adopt a robust forecasting strategy until enough data exist to allow the break to be modelled or only post-break data be used. However, the intrinsic difficulty is that by the nature of the exercise there are few observations available either to estimate parameters or to evaluate forecasts. For distant breaks, combinations of differently specified models are known to have good forecast properties, and we use a tailored version of this for the post-break period.

But an alternative is to ignore the discrete nature of the hypothesised structural change and pursue some robust forecasting strategy that effectively allows for time variation in a simple but flexible manner. In general, methods that give more weight to recent observations more than to distant data are likely to be robust. In line with this, we examine a rolling-window estimator, forecast averaging (combination) methods and an exponentially weighted moving average estimator, all avoiding the need to monitor for breaks. There is a cost - discarding data when there has not been a break - but also advantages: there is no delay in recognising a break has occurred, and they may be robust to varying forms of structural change.

In our Monte Carlo exercises which examine single and multiple breaks in an AR process and multiple breaks in a location model, the best methods can vary according to the particular break and parametrisation. Where we explore the monitoring method (only in the single break case) we find the gains are small, although equally the costs in terms of poor forecast performance where there are small breaks are also small. Other methods can do much better where there are large breaks. The results make it hard to recommend a single method and are sensitive to choice of parameters (e.g., the window length or the EWMA decay parameter). However, a version of the EWMA which we examine averaging over several decay values is only a good choice in the location model. The forecast averaging approach, while not always the best, often improves on the full sample benchmark and rarely comes with a large penalty where there are frequent or small breaks.

When we examine AR(1) models using about 200 US and UK time series, we find that for both countries while the averaged EWMA can occasionally do very well, in general it performs poorly and can perform very badly, consistent with the Monte Carlo results. For the UK, where there are relatively many breaks identified in the full sample, the best performing method is forecast averaging, again consistent with the Monte Carlo results, although rolling regressions and a low decay EWMA also beat the benchmark. For both countries monitoring brings only a small improvement in mean forecast performance, which makes it hard to recommend its use. It appears that the speculation of Clark and McCracken (2009a) quoted in the introduction was correct. However, there is a sense that it is a conservative strategy, as it can deliver improved forecast performance but is unlikely to lead to serious forecast failure relative to the benchmark.

References

Andrews, D. W. K. (1993): "Tests for Parameter Instability and Structural Change With Unknown Change Point," *Econometrica*, 61, 821–56.

Bai, J., and P. Perron (1998): "Estimating and Testing Linear Models with Multiple Structural Changes," *Econometrica*, 66, 47–78.

Brown, R. L., J. Durbin, and J. M. Evans (1975): "Techniques for Testing the Constancy of Regression Relationships over Time," *Journal of the Royal Statistical Society Series B*, 37, 149–63.

Chow, G. (1960): "Tests of equality between sets of coefficients in two linear regressions," *Econometrica*, 28, 591–603.

Chu, C.-S. J., M. Stinchcombe, and H. White (1996): "Monitoring Structural Change," *Econometrica*, 64, 1,045–65.

Clark, T. E., and M. W. McCracken (2009a): "Forecasting with Small Macroeconomic VARs in the Presence of Instabilities," in *Forecasting in the Presence of Structural Breaks and Model Uncertainty*, ed. by M. E. Wohar, and D. E. Rapach, pp. 93–147. Amsterdam: Elsevier.

——— (2009b): "Improving forecast accuracy by combining recursive and rolling forecasts," *International Economic Review*, 50(2), 363–95.

Clements, M. P., and D. F. Hendry (1998a): *Forecasting economic time series*. CUP, Cambridge.

Clements, M. P., and D. F. Hendry (1998b): "Intercept corrections and structural change," *Journal of Applied Econometrics*, 11, 475–94.

Diebold, F. X., and R. S. Mariano (1995): "Comparing Predictive Accuracy," *Journal of Business and Economic Statistics*, 13, 253–263.

Eklund, J., G. Kapetanios, and S. Price (2010): "Forecasting in the presence of recent structural change," *Bank of England Working Paper No. 406*.

Faust, J., and J. H. Wright (2007): "Comparing Greenbook and Reduced Form Forecasts using a Large Realtime Dataset," NBER Working Papers 13397, National Bureau of Economic Research, Inc.

Giacomini, R., and H. White (2006): "Tests of Conditional Predictive Ability," 74, 1,545–78.

Giraitis, L., G. Kapetanios, and S. Price (2013): "Adaptive forecasting in the presence of recent and ongoing structural change," *Journal of Econometrics (forthcoming)*.

Groen, J. J. J., G. Kapetanios, and S. Price (2013): "Multivariate Methods for Monitoring Structural Change," *Journal of Applied Econometrics (forthcoming)*.

Harvey, A. C. (1989): *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press.

Hendry, D. F. (2000): "On detectable and non-detectable structural change," *Structural Change and Economic Dynamics*, 11, 45–65.

Kapetanios, G., and E. Tzavalis (2010): "Modeling structural breaks in economic relationships using large shocks," *Journal of Economic Dynamics and Control*, 34(3), 417–36.

Koop, G., and S. Potter (2007): "Estimation and Forecasting in Models with Multiple Breaks," *Review of Economic Studies*, 74, 763–89.

Krämer, W., W. Ploberger, and R. Alt (1988): "Testing for Structural Change in Dynamic Models," *Econometrica*, 56, 1,355–69.

Kuan, C.-M., and K. Hornik (1995): "The Generalized Fluctuation Test: a Unifying View," *Econometric Reviews*, 14, 135–61.

Leisch, F., K. Hornik, and C.-M. Kuan (2000): "Monitoring Structural Changes With the Generalized Fluctuation Test," *Econometric Theory*, 16, 835–54.

Pesaran, M. H., D. Pettenuzzo, and A. Timmermann (2007): "Forecasting Time Series Subject To Multiple Structural Breaks," *Review of Economic Studies*, 73, 1,057–84.

Pesaran, M. H., and A. Pick (2011): "Forecast combination across estimation windows," *Journal of Business and Economic Statistics*, 29, 307–318.

Pesaran, M. H., and A. Timmermann (2007): "Selection of estimation window in the presence of breaks," *Journal of Econometrics*, 137, 134–61.

Stock, J. H., and M. Watson (1996): "Evidence on Structural Instability in Macroeconomic Time Series Relations," *Journal of Business and Economic Statistics*, 14, 11–30.

Zeileis, A., F. Leisch, C. Kleiber, and K. Hornik (2005): "Monitoring structural change in dynamic econometric models," *Journal of Applied Econometrics*, 20, 99–121.