MSc Econometrics and Management Science:
Quantitative Finance

Dynamic Term Structure Modelling:
A Forecasting Perspective

A Thesis Submitted in Partial Fulfilment of the Requirements
for the Degree of Master of Science

Author:
J.L. Yeh, 430672

Supervisor ESE:
Prof.dr. M. van der Wel

Supervisor KPMG:
P. Bosschaart

Second Assessor ESE:
Dr. R. Lange

November 21, 2019

Abstract

This paper examines the predictability of the term structure of US treasuries. We consider a range of term structure models in which we include richer dynamics. The starting point of our term structure models is the Nelson-Siegel framework. We extend this framework in three ways: (i) allow interactions with the macro-economy, (ii) include a time-varying unconditional mean, and (iii) incorporate Markov-switching dynamics. Moreover, we explore the merits of combining forecasts of individual term structure models by considering two types of combination schemes: (a) simple weighting schemes, and (b) time-varying weighting schemes. We find that no individual model consistently outperforms a no-change forecast. However, we do obtain more accurate and more robust forecasts by combining individual forecasts. Furthermore, we find the performance of both individual models and forecast combinations to be highly dependent on the forecasting period. Next, we barely find predictability over short horizons whereas longer horizons show promising results. All in all, we demonstrate how combining forecasts can improve the overall performance.

Keywords— Term structure; Nelson-Siegel; Macroeconomic factors; Markov-switching; Forecast combinations; State space models

Contents

1 Introduction

In the field of finance, modelling interest rate dynamics has attracted a good amount of attention from academics as well as practitioners. The task at hand has emerged as notorious for the challenges it poses in terms of fitting and predicting. The term structure of interest rates, that is, the relation between yields and time to maturity, is closely related both in the cross-section and over time. This follows from the fact that long yields are essentially risk-adjusted averages of expected future short rates. Despite the close relationship, short maturity yields are inclined to react differently from long maturity yields with respect to news shocks, which makes accurately describing the term structure rather difficult. The ever-growing term structure literature is one of the most important fields of research considering the size of the fixed income market and its many applications. Accurately describing the term structure and its dynamics is therefore of utmost importance given the applications of many fields in finance. These fields include, but are not limited to, derivative pricing, financial risk management, asset allocation, and conducting monetary policy.

The main objective of this study is to examine the out-of-sample forecasting performance of several term structure models. With this objective in mind, we particularly focus on the importance of model uncertainty in forecasting interest rates. We do so by expanding on research done by De Pooter, Ravazzolo, and Van Dijk (2010) and deviate in two distinct ways. First, we consider several extension of the Nelson-Siegel class. More specific, we consider extensions à la Svensson (1994) allowing for more flexibility. This choice is motivated by the popularity this class has attracted among financial practitioners and central banks (Diebold and Rudebusch, 2013). In addition, we consider extensions relaxing some assumptions in the spirit of Van Dijk, Koopman, Van der Wel, and Wright (2014), as well as Bernadell, Coche, and Nyholm (2005). The former relaxes the assumption of constant mean and the latter proposes a regime-switching extension of the Nelson-Siegel model. Second, we look into whether more sophisticated methods of combining forecasts, such as Billio, Casarin, Ravazzolo, and Van Dijk (2013), who consider time-varying combinations, can further improve on forecasting interest rates.

Over the last decades, academics have produced a rich literature concerning term structure models. This vast literature has seen significant improvements in modelling and understanding the term structure of interest rates. However, there is no clear-cut superior model. Instead, the term structure literature can generally be divided into two different classes.

The first class concerns the theoretical literature. This particular approach can be divided into two different subclasses. The first one relates to the general equilibrium framework which took off by the seminal contributions of Vasicek (1977) and Cox, Ingersoll Jr, and Ross (1985). They propose the affine term structure model which relate bond prices to the instantaneous short rate. These models gained significant interest due to the generalisation of Duffie and Kan (1996). Dai and Singleton (2000) provide a classification of the range of admissible and identifiable models. Despite the theoretical background and popularity, it turns out that it does poorly in forecasting (Duffee, 2002). Besides the disappointing forecasting performance, affine term structure models are notorious for posing difficulties in estimation.

Several researchers have pointed out the existence of multiple likelihood maxima which have near identical fit, but with contrasting implications for economic behaviour (Christensen, Diebold, and Rudebusch, 2011). Within the theoretical framework, the second approach imposes no-arbitrage restrictions. The earliest and most notable contributions are by Hull and White (1990) and Heath, Jarrow, and Morton (1992). These contributions focus on the cross-sectional dimension to ensure no-arbitrage restrictions. Although the imposition of the no-arbitrage condition has strong economic foundations, they render these models useless for forecasting purposes which is due to the strong cross-sectional focus (Duffee, 2011).

The second class is of statistical nature. Based on the premise that financial assets typically contain a factor structure, several statistical methods can and have been exploited. The most popular example is the class of Nelson and Siegel (1987), who imposes a three-factor structure on the yields. They show that their proposed factor structure fit the yield curve remarkably well. Subsequently, several extensions have been proposed to extend the flexibility, see for example Svensson (1994) and Björk and Christensen (1999). De Pooter (2007) provides an extensive review of the Nelson-Siegel model and several extensions. He finds that extensions allowing for more flexibility improve both in-sample and out-of-sample performance. The Nelson-Siegel class owes its popularity to its parsimony, flexibility, and mathematical properties (Diebold and Rudebusch, 2013). The original Nelson-Siegel model is a static approach to term structure modelling. An influential contribution by Diebold and Li (2006) has been a turning point in term structure literature. They reinterpreted the Nelson-Siegel factors as level, slope, and curvature. Subsequently, they proposed to make the factors dynamic through autoregressive processes. They report very promising results for their proposed dynamic Nelson-Siegel model relative to competitive benchmarks. From a theoretical aspect, this class has one important caveat. Namely, it allows for arbitrage. Efforts by Christensen et al. (2011) filled this gap by proposing an arbitrage-free Nelson-Siegel model.

One of the latest developments in term structure literature has been linking the yield curve to the macro-economy. The first study to do so is Ang and Piazzesi (2003). In their study, they augment the affine term structure model with macroeconomic variables. More importantly, they find that the inclusion of macroeconomic information improves forecasts of interest rates. Regarding the Nelson-Siegel class, Diebold, Rudebusch, and Aruoba (2006) have augmented this class with macroeconomic variables and also report promising results in terms of in-sample fit.

In spite of the substantial amount of term structure literature, little is said about model uncertainty. De Pooter et al. (2010) point out that the yield curve exhibits distinct dynamics in subperiods. For this reason, it seems near impossible for a single model to consistently outperform competitive benchmarks. One way to account for distinct dynamics is by taking a Markov-switching approach. Markov-switching models belong to a class of models which allow more complex nonlinear structures. More specifically, it is able to describe distinct dynamics during different regimes. Intuitively, the idea of multiple regimes is very appealing as economic theory suggests that the economy exhibit distinctive states. Indeed, previous literature has documented the existence of multiple states in many financial applications including the term structure of interest rates, see for example Ang and Bekaert (2002), Bansal and Zhou (2002), and

Dai, Singleton, and Yang (2007). Alternatively, model uncertainty can be tackled by combining forecasts of several term structure models. De Pooter et al. (2010) propose to do so and find encouraging results.

Our empirical research is based on monthly data of constant maturity yields of US government zero-coupon bonds for different maturities and spans from October 1993 through April 2019. We extend this data set with a large panel of 128 macroeconomic variables covering the same period. In light of all the aforementioned literature and results therein, we take the statistical route by adopting the Nelson-Siegel framework in state space form as our baseline model. Specifically, we consider the Nelson-Siegel model as described in Diebold and Li (2006) and an adjusted version of the Svensson (1994) model introduced in De Pooter (2007). We extend this framework in several directions where previous literature documented results. First, we include macroeconomic factors in our baseline model. The macroeconomic factors are constructed by means of principal component analysis (PCA). Similar to Exterkate, Dijk, Heij, and Groenen (2013), we expand the state vector with the macroeconomic factors. Hereby, we only allow the macroeconomic factors to interact with the Nelson-Siegel framework factors and not directly with the yields. Next, we incorporate a time-varying mean in the Nelson-Siegel framework. To incorporate a time-varying mean, we consider a similar approach to Van Dijk et al. (2014). That is, we capture slowly moving changes by incorporating an exponential smoothing recursion in the unconditional mean. Again, this is achieved by expanding the state vector to include the exponential smoothing recursion. Lastly, we account for multiple regimes by including Markov-switching dynamics. Related to Bernadell et al. (2005) and motivated by economic foundations, we include Markov-switching dynamics in the mean of the slope.

In terms of in-sample fit, we find that more flexible specifications of the Nelson-Siegel framework fit the cross-sectional yields better on average. Also, we find clear economic interpretations of a low inflation state and high inflation state in our Markov-switching model. As all the extensions impose different structures on the dynamics of the factors rather than the cross-section of yields, we do not find noticeable differences in in-sample fit.

The results of our forecasting study for individual models can be summarised by the following observations. First, we find no consistent performance for individual models relative to the no-change forecasts. Next, we find barely any predictability over short forecasting horizons. On the other hand, longer horizons do show predictable patterns. Third, macroeconomic factors do not seem to have information that could be exploited. Although, we do find it to be helpful during times of turmoil, at least to some extent. Lastly, the short end of the yield curve is more difficult to forecast in comparison to longer maturity yields. As we find no clear-cut winner in our individual models, we continue by considering combinations of forecasts of individual models. Here, we distinguish between two methods of combining forecasts. First, we consider simple averaging over forecasts methods. Second, we consider time-varying weights, where the weights evolve with a random walk over time. This is done by adopting a state space form, where the weights are the unobserved states. The weights in a general state space model are unbounded which, in some cases, results in very extreme weights. Therefore, we also consider restricted time-varying weights by transforming the weights in a nonlinear way. Consequently, the general state space model becomes nonlinear. To deal with the nonlinearity, we extract the weights with particle filtering methods. By combining forecasts, we find substantial improvements in overall performance. Especially, we are able to construct more stable forecasts and more capable to cope with recessions. Short horizons remain difficult to predict relative to a no-change forecast whereas longer horizons show a higher degree of predictability. To put these results in perspective with the individual model performance, we are able to construct more robust forecasts compared to individual models, in the sense that the overall performance is more stable by combining forecasts relative to a no-change benchmark. The performance of individual models is affected quite heavily by the forecasting period. We can mitigate this, to a certain extent, by combining forecasts of individual models which, in turn, results in a more consistent performance over the complete forecasting sample. Moreover, we obtain more accurate forecasts over longer forecast horizons by combining forecasts of individual models where the gains are particularly strong at the short end of the yield curve. Altogether, our results provide a compelling claim to account for model uncertainty.

The remainder of this paper is structured as follows. In Section 2 we introduce the Nelson-Siegel framework and our extensions. Section 3 describes the estimation procedures for our models. Section 4 elaborates on the data for our empirical analysis. We discuss in-sample results in Section 5. Section 6 presents the forecasting study and methodology for combining forecasts. In Section 7 we finalise and conclude our main findings.

2 Models

This section is subdivided in two different parts. In Subsection 2.1 we introduce the Nelson-Siegel framework of yield curve models and discuss two popular variants, and Subsection 2.2 continues by stating the Nelson-Siegel framework in state space form and presents several extensions to account for richer dynamics.

2.1 Nelson-Siegel Class

In the upcoming sections we elaborate on the functional form of the Nelson-Siegel class and highlight several appealing features of this class. We follow the formulation of Diebold and Li (2006) for the Nelson-Siegel framework.

2.1.1 Nelson-Siegel Model

Let the yield be denoted by $y_t(\tau)$ at time t with maturity τ. Then at any given time t, the yield curve can be represented as a smooth function of τ. The Nelson-Siegel model is then given by

$$y_t(\tau) = \beta_{1,t} + \beta_{2,t}\left(\frac{1 - e^{-\lambda\tau}}{\lambda\tau}\right) + \beta_{3,t}\left(\frac{1 - e^{-\lambda\tau}}{\lambda\tau} - e^{-\lambda\tau}\right), \tag{1}$$

where $\boldsymbol{\beta}_t = (\beta_{1,t},\, \beta_{2,t},\, \beta_{3,t})'$ are the factors level, slope, and curvature, respectively. The fixed parameter,

λ, is the decay parameter and determines the exponential rate of decay of the slope and curvature loading.

At first sight, one can argue that this functional form to approximate the yield curve is somewhat arbitrary. However, upon further inspection, this form contains several appealing features. First, it is a parsimonious approximation of the yield curve requiring only three components to describe the cross-section of yields, for a fixed decay parameter. A parsimonious approximation is key as (i) it supports smoothness, (ii) it prevents in-sample overfitting, and (iii) estimation remains tractable and trustworthy. Smoothness is motivated by empirical evidence that yields are (mostly) smooth functions of maturity. In general, overfitting tends to result in poor forecasts, and tractable and trustworthy estimation is always a welcome addition. Despite the small number of components, it is a very flexible approximation as well. This functional form is able to capture typical empirically observed shapes of the yield curve. Besides the parsimonious and flexible approximation, the Nelson-Siegel factors can also be interpreted from an economic point of view. Consider the limiting behaviour of the standard Nelson-Siegel form:

$$\lim_{\tau \downarrow 0} y_t(\tau) = \beta_{1,t} + \beta_{2,t}; \qquad \lim_{\tau \to \infty} y_t(\tau) = \beta_{1,t}. \tag{2}$$

From the limits one can easily conclude that the short-term interest rate, i.e. when $\tau \downarrow 0$, is affected by the first and second factor, and the long-term interest rate, i.e. when $\tau \to \infty$, is affected only by the first factor. When we define the slope of the yield curve as the difference between the long-term interest rate and short-term interest rate, we have that the slope converges to $-\beta_{2t}$ for a given t. The third factor loading approaches zero in both extreme limiting cases of τ and is positive for intermediate values of τ. Therefore, $\beta_{3,t}$ is commonly interpreted as the curvature factor.

2.1.2 Adjusted Svensson Model

The Svensson (1994) four-factor model is another workhorse in practice and proposes an extra hump-shape factor with an additional decay parameter. The extra hump factor allows for more flexibility, however it has been argued in the case that when both decay parameters are similar, multicolinearity arises (De Pooter, 2007). To remedy this problem, De Pooter (2007) proposes an adjusted Svensson model as follows

$$y_t(\tau) = \beta_{1,t} + \beta_{2,t}\left(\frac{1-e^{-\lambda_1\tau}}{\lambda_1\tau}\right) + \beta_{3,t}\left(\frac{1-e^{-\lambda_1\tau}}{\lambda_1\tau} - e^{-\lambda_1\tau}\right) + \beta_{4,t}\left(\frac{1-e^{-\lambda_2\tau}}{\lambda_2\tau} - e^{-2\lambda_2\tau}\right), \tag{3}$$

where the first three terms remain the same in interpretation as in Equation (1) and the fourth term is the second curvature factor with corresponding rate of exponential decay, λ_2. The second curvature factor allows for term structure shapes with multiple local maxima or minima. Again, the additional curvature factor primarily affects medium-term maturities, but has a faster rate of decay than the first curvature factor. Therefore, the limiting results of (2) also holds for this model.

2.2 General State Space Representation

As the two different Nelson-Siegel specifications are nested, we proceed with the general set-up. Let us observe a set of yields with n different maturities at time t, we can then simply estimate the yield curve via the following regression model

$$y_t(\tau_i) = \Lambda(\tau_i)\boldsymbol{\beta}_t + \varepsilon_{it} \qquad \text{for } i = 1, ..., n, \tag{4}$$

and $t = 1, ..., T$. The factor loadings for maturity τ_i are given by $\Lambda(\tau_i) = \Big[1, \big(\frac{1-e^{-\lambda\tau_i}}{\lambda\tau_i}\big), \big(\frac{1-e^{-\lambda\tau_i}}{\lambda\tau_i} - e^{-\lambda\tau_i}\big)\Big]$ and $\Lambda(\tau_i) = \Big[1, \big(\frac{1-e^{-\lambda_1\tau_i}}{\lambda_1\tau_i}\big), \big(\frac{1-e^{-\lambda_1\tau_i}}{\lambda_1\tau_i} - e^{-\lambda_1\tau_i}\big), \big(\frac{1-e^{-\lambda_2\tau_i}}{\lambda_2\tau_i} - e^{-2\lambda_2\tau_i}\big)\Big]$ for the Nelson-Siegel specification and adjusted Svensson specification, respectively, $\boldsymbol{\beta}_t$ is a $(K \times 1)$ vector with the factors, and ε_{it} is the error term. The error terms are assumed to be independent with mean zero and constant variance. It is therefore possible to apply cross-sectional least squares to obtain estimates for the factors. This general exposition only serves as an introduction to the state space formulation.

Diebold et al. (2006) recognise that $\boldsymbol{\beta}_t$ can be treated as a latent vector and subsequently represent the Nelson-Siegel framework in a state space model. The Nelson-Siegel framework can be captured in the following state space representation

$$\mathbf{y}_t = \boldsymbol{\Lambda}\boldsymbol{\beta}_t + \boldsymbol{\varepsilon}_t, \qquad \boldsymbol{\varepsilon}_t \sim \text{NID}(\mathbf{0}, \boldsymbol{\Sigma}_\varepsilon), \tag{5}$$

$$\boldsymbol{\beta}_t = \boldsymbol{\mu} + \boldsymbol{\Phi}(\boldsymbol{\beta}_{t-1} - \boldsymbol{\mu}) + \boldsymbol{\eta}_t, \qquad \boldsymbol{\eta}_t \sim \text{NID}(\mathbf{0}, \boldsymbol{\Sigma}_\eta). \tag{6}$$

In the observation equation, Equation (5), $\mathbf{y}_t = [y_t(\tau_1), y_t(\tau_2), ..., y_t(\tau_n)]'$ is the observation vector containing n yields, $\boldsymbol{\Lambda}$ is an $(n \times K)$ matrix containing the factor loadings, and $\boldsymbol{\varepsilon}_t$ is an $(n \times 1)$ disturbance vector. The two Nelson-Siegel specifications of the preceding sections are special cases of Equation (5) with the adjusted Svensson model having an additional factor. Note that the loading matrix, $\boldsymbol{\Lambda}$, is fixed when the decay parameter(s) is (are) fixed. Diebold and Li (2006) suggest to fix it at 0.0609 which maximises the loading at a 30-month maturity. We take a slightly different approach. That is, we do consider the decay parameter(s) fixed, but we include it (them) in the parameter vector to be estimated.

The dynamics of the factors are modelled through the transition equation, Equation (6). We follow Diebold et al. (2006) and specify a first-order autoregressive (AR) process for the factors. This can be done either by a univariate AR(1) process or multivariate VAR(1) model. We consider AR(1) processes as a VAR(1) would introduce additional estimation error and this is not necessarily desirable for forecasting purposes. Thus, we have a $(K \times 1)$ mean vector, $\boldsymbol{\mu}$, and a diagonal $(K \times K)$ transition matrix, $\boldsymbol{\Phi}$. The diagonal elements of the transition matrix, $\boldsymbol{\Phi}$, are denoted by ϕ_{jj} for $j = 1, ..., K$. Throughout this thesis, we assume the covariance matrices, $\boldsymbol{\Sigma}_\varepsilon$ and $\boldsymbol{\Sigma}_\eta$, to be diagonal to reduce the number of parameters to be estimated. We label the general dynamic Nelson-Siegel and dynamic adjusted Svensson models by **NS** and **AS**, respectively.

2.2.1 Macroeconomic Information

The Nelson-Siegel framework approaches the yield curve from a statistical perspective. We now allow the yield curve factors to interact with the macro-economy. We do so in a data-rich environment by extracting a small set of macroeconomic factors from a large panel of macroeconomic variables. Incorporating macroeconomic information can be done in a fairly straightforward manner in the state space framework. We follow Exterkate et al. (2013) and extend the state vector, $\boldsymbol{\beta}_t$, to include macroeconomic factors. That is, the state vector is now given by $(\beta_{1,t}, ..., \beta_{K,t}, f_{1,t}, ..., f_{p,t})'$, where K is the number of yield curve factors and p denotes the number of macro factors. We allow for a unidirectional interaction between yields and macro factors. This restriction is imposed to constrain the number of parameters to estimate. Moreover, Diebold et al. (2006) find that the causal relationship from macro factors to yields is stronger than vice versa, and Exterkate et al. (2013) report superior results for the restricted model as opposed to the unrestricted counterpart. Thus, we consider the following structure for the transition matrix

$$\boldsymbol{\Phi}^{\text{FA}} = \left(\begin{array}{c|c} \text{Diagonal} & \text{Unrestricted} \\ \hline \text{Zero} & \text{Diagonal} \end{array}\right),$$

where the blocking represents partitioning of the state vector in $\boldsymbol{\beta}_t$ and $\boldsymbol{f}_t$.

Due to the extended state vector, we also have to modify the structure of the general state space model. As we only allow the macro factors to interact with the yield curve factors, we have to modify the general state space representation. In the general representation, Equation (5) remains the same and we have the following structure for the transition equation

$$\begin{pmatrix} \boldsymbol{\beta}_t \\ \boldsymbol{f}_t \end{pmatrix} = \left(\mathbf{I}_{K+p} - \boldsymbol{\Phi}^{\text{FA}}\right) \begin{pmatrix} \boldsymbol{\mu} \\ \mathbf{0}_p \end{pmatrix} + \boldsymbol{\Phi}^{\text{FA}} \begin{pmatrix} \boldsymbol{\beta}_{t-1} \\ \boldsymbol{f}_{t-1} \end{pmatrix} + \boldsymbol{\eta}_t, \qquad \boldsymbol{\eta}_t \sim \text{NID}(\mathbf{0}, \boldsymbol{\Sigma}_\eta), \tag{7}$$

where $\boldsymbol{\Phi}^{\text{FA}}$ is as defined above and the dimension of $\boldsymbol{\eta}$ and $\boldsymbol{\Sigma}_\eta$ are increased as appropriate. We denote the factor-augmented model by **NS-X** and **AS-X** for the Nelson-Siegel and adjusted Svensson variants, respectively.

2.2.2 Shifting Endpoints

In the Nelson-Siegel framework, under the assumption that $|\phi_{jj}| < 1$, the autoregressive process given in the transition equation implies a stationary and mean-reverting process with constant unconditional mean equal to μ_j.

Van Dijk et al. (2014) show that this assumption is not appropriate and propose to incorporate a time-varying mean. They do so by specifying the mean as a local level model, where the mean follows an exponential smoothing recursion. That is,

$$\boldsymbol{\mu}_{t+1} = \alpha\boldsymbol{\beta}_t + (1-\alpha)\boldsymbol{\mu}_t, \qquad \text{for } t = 1, 2, ..., \tag{8}$$

where $0 < \alpha < 1$ is the smoothing parameter and starting with $\boldsymbol{\mu}_1 = \boldsymbol{\beta}_1$. This recursion can be rewritten in

$$\boldsymbol{\mu}_{t+1} = \alpha \sum_{l=0}^{t-2} (1-\alpha)^l \boldsymbol{\beta}_{t-1} + (1-\alpha)^{t-1}\boldsymbol{\beta}_1. \tag{9}$$

It then follows that the mean at $t+1$ is an exponentially weighted moving average of past factor values. By substituting the mean in the state equation, we obtain

$$\boldsymbol{\beta}_t = \boldsymbol{\omega}\boldsymbol{\beta}_{t-1} + (\mathbf{I}_K - \boldsymbol{\omega})\boldsymbol{\mu}_{t-1} + \boldsymbol{\eta}_t, \tag{10}$$

where $\boldsymbol{\omega} = \boldsymbol{\alpha} + \boldsymbol{\Phi}$ and $\boldsymbol{\alpha} = \alpha\mathbf{I}_K$. The conditional expectation of the factors are now a weighted average of the previous factor realisation and the unconditional expectation.

The smoothness of the shifting endpoints and yield forecasts crucially depend on the smoothing parameter, α. The lower the α is, the smoother the yield forecasts will be. To illustrate this, consider the case when α gets close to zero, as l increases, the exponential decay converges more gradually to zero. This, in turn, will result in more smoothness of the yield forecasts. Van Dijk et al. (2014) report little sensitivity to the precise choice of the smoothing parameter. We have performed similar robustness analysis and find that, in terms of Root Mean Squared Prediction Error (RMSPE), the forecasts are barely different for roughly $\alpha > 0.05$. Lower values of α yielded worse performance. Similarly to Van Dijk et al. (2014), we choose to set $\alpha = 0.1$.

To account for such dynamics in the state space framework, we have to modify the general state space representation. This is done by expanding the transition equation to include the time-varying mean. Again, the measurement equation remains the same and we obtain the following transition equation

$$\begin{pmatrix} \boldsymbol{\beta}_t \\ \boldsymbol{\mu}_t \end{pmatrix} = \begin{pmatrix} \boldsymbol{\omega} & \mathbf{I}_K - \boldsymbol{\omega} \\ \boldsymbol{\alpha} & \mathbf{I}_K - \boldsymbol{\alpha} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_{t-1} \\ \boldsymbol{\mu}_{t-1} \end{pmatrix} + \boldsymbol{\eta}_t^{\text{ES}}, \qquad \boldsymbol{\eta}_t^{\text{ES}} \sim \text{NID}(\mathbf{0}, \boldsymbol{\Sigma}_\eta^{\text{ES}}). \tag{11}$$

The mean does not contain any noise, and therefore $\boldsymbol{\eta}_t^{\text{ES}}$ and $\boldsymbol{\Sigma}_\eta^{\text{ES}}$ are specified as

$$\boldsymbol{\eta}_t^{\text{ES}} = \begin{pmatrix} \boldsymbol{\eta}_t \\ \mathbf{0}_K \end{pmatrix}, \qquad \boldsymbol{\Sigma}_\eta^{\text{ES}} = \begin{pmatrix} \boldsymbol{\Sigma}_\eta & \mathbf{0}_{K\times K} \\ \mathbf{0}_{K\times K} & \mathbf{0}_{K\times K} \end{pmatrix}.$$

This model is denoted by **NS-ES** and **AS-ES** for the Nelson-Siegel and adjusted Svensson variants, respectively.

2.2.3 Markov-Switching Model

We take a similar approach to Bernadell et al. (2005) and include a Markov-switching structure. They connect economic theory with the yield curve. More specifically, their reasoning is based on the Taylor (1993) rule where the slope of different states is linked to the central bank's policy. To illustrate this, consider the case when the economy is in a state of low inflation levels, then the central bank will decrease the short rate to stimulate the economy. This results in the yield curve to be more upward sloping. Conversely, when the economy is in a state of high inflation levels, then the central bank will increase the short rate which causes the yield curve to flatten or even invert. This reasoning is built upon the premise that the long rate is constant relative to the short rate. This premise is supported by the Fisher equation which states that the nominal yields equals the sum of expected real interest rate and the inflation rate. In the long run, the real rate should be more stable as it equals the growth of the economy. In the short run, the central bank's policy influences the real rate causing it to be less stable.

We consider a two-state regime-switching model where we allow the slope factor to switch between the two states. By doing so, the model can account for distinct slopes in different phases of the business cycle. Let us now introduce the switching mechanism we employ in our empirical analysis. Denote state $S_t \in \{1, 2\}$ as the latent variable for the regime at time t. Then the transition probability matrix is given by

$$\mathbf{P} = \begin{pmatrix} p_{11} & p_{21} \\ p_{12} & p_{22} \end{pmatrix} = \begin{pmatrix} p_{11} & 1 - p_{22} \\ 1 - p_{11} & p_{22} \end{pmatrix}, \tag{12}$$

where $p_{ij} = \Pr[S_t = j | S_{t-1} = i]$ is the probability of moving from state i at time $t - 1$ to state j at time t.

The distinct regimes are modelled through the transition equation. More specifically, we only allow different regimes in the slope factor which we do through the mean vector, $\boldsymbol{\mu}$, of the general transition equation. In the state space representation, the transition equation is then modified as

$$\boldsymbol{\beta}_t = \boldsymbol{\mu}_{S_t} + \boldsymbol{\Phi}(\boldsymbol{\beta}_{t-1} - \boldsymbol{\mu}_{S_t}) + \boldsymbol{\eta}_t, \qquad \boldsymbol{\eta}_t \sim \text{NID}(\mathbf{0}, \boldsymbol{\Sigma}_\eta), \tag{13}$$

where $\boldsymbol{\mu}_{S_t} = (\mu_1, \mu_{2,S_t}, \mu_3)'$, or $\boldsymbol{\mu}_{S_t} = (\mu_1, \mu_{2,S_t}, \mu_3, \mu_4)'$ depending on the Nelson-Siegel model, or adjusted Svensson model, respectively. We note that this specification is modified in several ways compared to the formulation of Bernadell et al. (2005). First, we consider a two-state instead of the three-state model proposed by Bernadell et al. (2005). We motivate this choice by noting that a three-state model requires substantially more data to reliably estimate, and to retain tractability. Besides, we let switching occur endogenously instead which is also the second difference between Bernadell et al. (2005). In their approach, the transition matrix is determined via macroeconomic factor realisations with threshold values to alternate between transition probability matrices. Lastly, we do specify an autoregressive process for the slope factor.[1] The Markov-switching model is denoted by **NS-MS** and **AS-MS** for the Nelson-Siegel and adjusted Svensson variants, respectively.

[1] We have tested the specification with autoregressive process against without autoregressive process for the slope factor. The results were convincingly in favour of the specification with autoregressive process. Therefore, we proceed with this specification.

3 Estimation Procedures

In Subsection 2.2 we have described and formulated our models in state space form. We now turn to the estimation of these models. Conventionally, state space models are estimated via numerical procedures utilising the Kalman filter. We consider a hybrid of several estimation procedures. That is, we combine the two-step approach of Diebold and Li (2006) with the Expectation-Maximisation (EM) algorithm for state space models to obtain starting values for a traditional gradient-based method. This hybrid procedure is suggested by Diebold and Rudebusch (2013) and basically provides highly accurate starting values for the gradient-based method which then should be able to quickly move to an optimum.[2]

We outline the general hybrid estimation procedure in the systemic scheme below and elaborate in more detail afterwards:

1. Estimate $\hat{\boldsymbol{\beta}}_t$ from cross-sectional regressions for $t = 1, ..., T$.

2. Fit a VAR(1) model to the estimated factors, $\{\hat{\boldsymbol{\beta}}_t\}_{t=1}^{T}$, of step 1 according to the transition equation specification for each model.

3. Perform 50 iterations of the EM algorithm for state space models using the estimates of step 2 as initialisation values for the transition equation.[3]

4. Use the estimates from step 3 as input for traditional gradient-based optimisation method.

We discuss the first two steps of the estimation procedure in Subsection 3.1. The third and fourth steps are elaborated on in Subsections 3.2 and 3.3, respectively. In essence, we apply the hybrid estimation procedure to all the models except for the shifting endpoints model. For this specific model, we skip the third step which we explain in Subsection 3.2.

3.1 Diebold and Li (2006) Two-Step Approach

The first two steps essentially summarise the two-step procedure of Diebold and Li (2006). These two steps are easy to perform when the decay parameters are predetermined. We choose to initialise the decay parameter in the Nelson-Siegel model with the Diebold and Li (2006) value, 0.0609. For the adjusted Svensson model we initialise the decay parameters as $\lambda_1 = 0.0609$ and $\lambda_2 = 1.5\lambda_1$. The choice for λ_2 is such that we impose the second hump to be after the first hump. Moreover, the estimation procedure is robust against initial choices for the decay parameters as we quickly find sensible estimates for the decay parameters in the EM algorithm.

[2] We note that they merely suggest this procedure and do not apply it. In our empirical application, we experienced a substantial amount of numerical difficulties implementing the models with only traditional gradient-based methods when the starting values were too different from the optimum. Thereby requiring a huge computational effort which rendered traditional gradient-based methods infeasible within a reasonable amount of time for this thesis. Therefore, we resort to this hybrid estimation procedure.

[3] From limited experimentation we found that the relative improvement in likelihood decreases significantly after a couple dozen of iterations. For most models, the most significant improvements are even within the first 10 iterations. We proceed with 50 iterations as this provides a good balance between accurate starting values and computation time.

3.2 Step 3: EM Algorithm

The third step makes use of the EM algorithm, as described in for example Shumway and Stoffer (1982), which is tailored to each model and relies on analytical solutions. We have unobserved factors in the state space model. In essence, the EM algorithm for state space models presumes the state vector containing the latent factors as known, then estimation becomes straightforward. The EM algorithm maximises the following joint likelihood

$$\begin{aligned} L(\mathbf{y}_{1:T}, \boldsymbol{\beta}_{0:T}|\boldsymbol{\theta}) = &\frac{T}{2}\log|\boldsymbol{\Sigma}_{\varepsilon}^{-1}| - \frac{1}{2}\sum_{t=1}^{T}(\mathbf{y}_t - \boldsymbol{\Lambda}(\lambda)\boldsymbol{\beta}_t)'\boldsymbol{\Sigma}_{\varepsilon}^{-1}(\mathbf{y}_t - \boldsymbol{\Lambda}(\lambda)\boldsymbol{\beta}_t) \\ &+ \frac{T}{2}\log|\boldsymbol{\Sigma}_{\eta}^{-1}| - \frac{1}{2}\sum_{t=1}^{T}[(\boldsymbol{\beta}_t - \boldsymbol{\mu}) - \boldsymbol{\Phi}(\boldsymbol{\beta}_{t-1} - \boldsymbol{\mu})]'\boldsymbol{\Sigma}_{\eta}^{-1}[(\boldsymbol{\beta}_t - \boldsymbol{\mu}) - \boldsymbol{\Phi}(\boldsymbol{\beta}_{t-1} - \boldsymbol{\mu})]. \end{aligned} \tag{14}$$

However, the state vector is not known, but, given a fixed parameter set, we can optimally extract it via the Kalman filter and then run the Kalman smoother over the extracted state vector. Conditional on the smoothed state vector, we update the previous parameter set with analytical solutions. Strictly speaking we do not use analytical solutions for all parameters. This is because the decay parameter is highly nonlinear prohibiting us from obtaining analytical solutions. Instead, we numerically maximise the decay parameter conditional on all other parameters being fixed.[4] For brevity, we only provide the updating formulas of the EM algorithm here. The expressions for the Kalman filter and smoother can be found in Appendix B.1, and derivations for the updating formulas can be found in Appendix B.3. The updating formulas are

$$\begin{aligned} \boldsymbol{\mu} &= (\mathbf{I}_K - \boldsymbol{\Phi})^{-1}\frac{1}{T}\sum_{t=1}^{T}[\hat{\boldsymbol{\beta}}_{t|T} - \boldsymbol{\Phi}\hat{\boldsymbol{\beta}}_{t-1|T}], \\ \boldsymbol{\Phi}_i^u &= \left(\sum_{t=1}^{T}(\hat{\beta}_{i,t|T} - \mu_i)(\hat{\boldsymbol{\beta}}_{t-1|T} - \boldsymbol{\mu})^{u\prime} + P_{t,t-1|T}^{(i,u)}\right)\left(\sum_{t=1}^{T}(\hat{\boldsymbol{\beta}}_{t-1|T} - \boldsymbol{\mu})^{u}(\hat{\boldsymbol{\beta}}_{t-1|T} - \boldsymbol{\mu})^{u\prime} + P_{t-1|T}^{(u,u)}\right)^{-1}, \\ \boldsymbol{\Sigma}_{\eta} &= \frac{1}{T}\sum_{t=1}^{T}\Big((\hat{\boldsymbol{\beta}}_{t|T} - \boldsymbol{\mu})(\hat{\boldsymbol{\beta}}_{t|T} - \boldsymbol{\mu})' + P_{t|T} - \boldsymbol{\Phi}[(\hat{\boldsymbol{\beta}}_{t-1|T} - \boldsymbol{\mu})(\hat{\boldsymbol{\beta}}_{t|T} - \boldsymbol{\mu})' + P_{t-1,t|T}] \\ &\quad - [(\hat{\boldsymbol{\beta}}_{t|T} - \boldsymbol{\mu})(\hat{\boldsymbol{\beta}}_{t-1|T} - \boldsymbol{\mu})' + P_{t,t-1|T}]\boldsymbol{\Phi}' + \boldsymbol{\Phi}[(\hat{\boldsymbol{\beta}}_{t-1|T} - \boldsymbol{\mu})(\hat{\boldsymbol{\beta}}_{t-1|T} - \boldsymbol{\mu})' + P_{t-1|T}]\boldsymbol{\Phi}'\Big), \\ \boldsymbol{\Sigma}_{\varepsilon} &= \frac{1}{T}\sum_{t=1}^{T}\Big(\mathbf{y}_t\mathbf{y}_t' - \boldsymbol{\Lambda}(\lambda)\hat{\boldsymbol{\beta}}_{t|T}\mathbf{y}_t' - \mathbf{y}_t\hat{\boldsymbol{\beta}}_{t|T}'\boldsymbol{\Lambda}(\lambda)' + \boldsymbol{\Lambda}(\lambda)[\hat{\boldsymbol{\beta}}_{t|T}\hat{\boldsymbol{\beta}}_{t|T}' + P_{t|T}]\boldsymbol{\Lambda}(\lambda)'\Big), \end{aligned} \tag{15}$$

where $\hat{\boldsymbol{\beta}}_{t|T}$ and $P_{t|T}$ are the smoothed state vector and accompanying smoothed uncertainty, respectively. We denote $P_{t,t-1|T}$ as the smoothed covariance between two consecutive smoothed state vectors. As we impose constraints on the transition matrix, we estimate the unconstrained part of the transition matrix row-by-row denoted by $\boldsymbol{\Phi}_i^u$, where u selects the unconstrained part. Next, the superscript u selects the elements in $(\hat{\boldsymbol{\beta}}_{t-1|T} - \boldsymbol{\mu})$ corresponding to the unconstrained part. Further, the double superscript in $P_{t,t-1|T}^{(i,u)}$ represents the i-th row and the columns of u in $P_{t,t-1|T}$. Similarly, the double superscript

[4] Admittedly, we do increase the computational effort, but we found the increase to be near negligible and got highly accurate starting values for the loading parameter. Besides, we only perform 50 iterations.

in $P^{(u,u)}_{t-1|T}$ represents the rows and columns of u in $P_{t-1|T}$. The minimisation problem for the loading parameter becomes

$$\min \sum_{t=1}^{T} \mathrm{Tr}\big[(\mathbf{y}_t - \boldsymbol{\Lambda}(\lambda)\hat{\boldsymbol{\beta}}_{t|T})(\mathbf{y}_t - \boldsymbol{\Lambda}(\lambda)\hat{\boldsymbol{\beta}}_{t|T})' + \boldsymbol{\Lambda}(\lambda)P_{t|T}\boldsymbol{\Lambda}(\lambda)'\big]. \tag{16}$$

The initial parameter set uses the estimates of step 2 for the transition equation, and the variances of both the observation and transition equations are initialised at 1. After updating the parameters with the previous parameter set, we can run the Kalman filter and smoother again with the updated parameter set. We perform 50 iterations of this procedure.

We have now provided the EM algorithm for the general state space model. As our models have different specifications and dynamics, we now give an outline of how the EM algorithm has to be modified to account for each model's needs. First, we note that for the factor-augmented model we do not need to make any modifications to the EM algorithm as described above. Essentially, the factor-augmented model extends the state vector, thereby changing the structure of the transition matrix. Our EM algorithm can account for differences in the transition matrix structure. Therefore, we can proceed with the general EM algorithm, where we only have to adjust the state vector and transition matrix. Next, it is not possible to derive an EM algorithm for the shifting endpoints model. To understand the underlying cause, recall that the conditional expectation of the mean is an exponentially weighted moving average of past factor values. Consequently, we need terms which are not produced by the smoothing procedure. Therefore, it is not possible to use the EM algorithm for the shifting endpoints model. Thus for this model, we proceed by estimating this model via the two-step approach and traditional gradient-based optimisation method. Finally, for the Markov-switching model, we need to make some adjustments to account for the regimes which we do in the upcoming section.

3.2.1 Markov-Switching EM Algorithm

We have introduced a first-order Markov process in the Markov-switching model. Consequently, the joint likelihood has to account for the switching process. We first define some new variables to facilitate notation. Denote the "Kronecker delta", $\tilde{\delta}_{ij}(t)$, as a random variable which equals 1 if $S_t = i$ and $S_{t-1} = j$, and $\tilde{\delta}_j(t)$ as the random variable which equals 1 if $S_t = j$. We now maximise the following joint likelihood

$$L(\mathbf{y}_{1:T}, \boldsymbol{\beta}_{0:T}, S_{1:T}|\boldsymbol{\theta}, \mathbf{P}, \boldsymbol{\rho}) = \sum_{t=1}^{T}\left[\sum_{i,j=1}^{M} \tilde{\delta}_{ij}(t)\Big(\log[f_i(\mathbf{y}_t, \boldsymbol{\beta}_t|\boldsymbol{\theta})p_{ij}]\Big)\right] + \sum_{j=1}^{M} \tilde{\delta}_j(0)\log\rho_j, \tag{17}$$

where M denotes the number of states, and $f_i(\mathbf{y}_t, \boldsymbol{\beta}_t|\boldsymbol{\theta})$ is the joint density of the observation and transition equation. That is,

$$\begin{aligned} f_i(\mathbf{y}_t, \boldsymbol{\beta}_t|\boldsymbol{\theta}) &= (2\pi)^{-n/2}|\boldsymbol{\Sigma}_\varepsilon|^{-1/2}\exp\left\{-\frac{1}{2}(\mathbf{y}_t - \boldsymbol{\Lambda}\boldsymbol{\beta}_t)'\boldsymbol{\Sigma}_\varepsilon^{-1}(\mathbf{y}_t - \boldsymbol{\Lambda}\boldsymbol{\beta}_t)\right\} \\ &\quad \times |\boldsymbol{\Sigma}_\eta|^{-1/2}\exp\left\{-\frac{1}{2}[(\boldsymbol{\beta}_t - \boldsymbol{\mu}_i) - \boldsymbol{\Phi}(\boldsymbol{\beta}_{t-1} - \boldsymbol{\mu}_i)]'\boldsymbol{\Sigma}_\eta^{-1}[(\boldsymbol{\beta}_t - \boldsymbol{\mu}_i) - \boldsymbol{\Phi}(\boldsymbol{\beta}_{t-1} - \boldsymbol{\mu}_i)]\right\}, \end{aligned} \tag{18}$$

where $\boldsymbol{\mu}_i$ denotes the mean in state i. After taking the expectation, we maximise the following quantity

$$\mathrm{E}[L(\mathbf{y}_{1:T}, \boldsymbol{\beta}_{0:T}, S_{1:T}|\boldsymbol{\theta}, \mathbf{P}, \boldsymbol{\rho})|\mathcal{I}_T] = \sum_{t=1}^{T}\left[\sum_{i,j=1}^{M} p^*_{ij}(t)\left(\mathrm{E}\Big[\log[f_i(\mathbf{y}_t, \boldsymbol{\beta}_t|\boldsymbol{\theta})p_{ij}]\Big|\mathcal{I}_T\Big]\right)\right] + \sum_{j=1}^{M} p^*_j(0)\log\rho_j, \quad (19)$$

where $\mathcal{I}_T$ denotes the information set until time T, $p^*_{ij}(t) = \Pr[S_t = i, S_{t-1} = j|\mathcal{I}_T]$ and $p^*_i(t) = \Pr[S_t = i|\mathcal{I}_T]$.

For the Markov-switching state space model, the standard Kalman procedure will not suffice. Therefore, we make use of the Kim (1994) algorithm for approximate filtering. This algorithm combines the Kalman filter and Hamilton filter. Again, for brevity, we only provide the updating formulas of the EM algorithm. The Kim filter and derivations for the updating formulas can be found in Appendices B.2 and B.3.1, respectively. We do not have to make adjustments to the updating formulas for $\boldsymbol{\Sigma}_\varepsilon$ and the minimisation problem of the decay parameters as those are independent of the regimes. We start with the updating formulas for the initial state distribution and transition probabilities which are given by

$$\hat{\boldsymbol{\rho}} = \hat{\xi}_{0|T} \qquad \text{and} \qquad p_{kl} = \frac{\sum_{t=1}^{T} p^*_{kl}(t)}{\sum_{t=1}^{T} p^*_l(t-1)}, \quad (20)$$

where $\hat{\xi}_{0|T}$ is the smoothed initial state. Next, we have to update the elements of the mean vector individually to account for the regime-dependent mean of the slope factor. We update the mean of the slope factor in state k as

$$\mu_{2,k} = \frac{\sum_{t=1}^{T} p^*_k(t)(\beta_{2,t|T} - \phi_{22}\beta_{2,t-1|T})}{(1-\phi_{22})\sum_{t=1}^{T} p^*_k(t)}, \quad (21)$$

whereas the mean of factor i is updated as

$$\mu_i = (1-\phi_{ii})^{-1}\frac{1}{T}\sum_{t=1}^{T}(\beta_{i,t|T} - \phi_{ii}\beta_{i,t-1|T}), \quad (22)$$

where $i \neq 2$. At last, the remaining updating formulas for $\boldsymbol{\Phi}^u_i$ and $\boldsymbol{\Sigma}_\eta$ are

$$\begin{aligned}
\boldsymbol{\Phi}^u_i =& \left(\sum_{t=1}^{T}\sum_{k=1}^{M} p^*_k(t)(\hat{\beta}_{i,t|T} - \mu_{i,k})(\hat{\boldsymbol{\beta}}_{t-1|T} - \boldsymbol{\mu}_k)^{u\prime} + P^{(i,u)}_{t,t-1|T}\right) \\
&\times \left(\sum_{t=1}^{T}\sum_{k=1}^{M} p^*_k(t)(\hat{\boldsymbol{\beta}}_{t-1|T} - \boldsymbol{\mu}_k)^{u}(\hat{\boldsymbol{\beta}}_{t-1|T} - \boldsymbol{\mu}_k)^{u\prime} + P^{(u,u)}_{t-1|T}\right)^{-1}, \\
\boldsymbol{\Sigma}_\eta =& \frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{M} p^*_k(t)\Big((\hat{\boldsymbol{\beta}}_{t|T} - \boldsymbol{\mu}_k)(\hat{\boldsymbol{\beta}}_{t|T} - \boldsymbol{\mu}_k)' + P_{t|T} - \boldsymbol{\Phi}[(\hat{\boldsymbol{\beta}}_{t-1|T} - \boldsymbol{\mu}_k)(\hat{\boldsymbol{\beta}}_{t|T} - \boldsymbol{\mu}_k)' + P_{t-1,t|T}] \\
&- [(\hat{\boldsymbol{\beta}}_{t|T} - \boldsymbol{\mu}_k)(\hat{\boldsymbol{\beta}}_{t-1|T} - \boldsymbol{\mu}_k)' + P_{t,t-1|T}]\boldsymbol{\Phi}' + \boldsymbol{\Phi}[(\hat{\boldsymbol{\beta}}_{t-1|T} - \boldsymbol{\mu}_k)(\hat{\boldsymbol{\beta}}_{t-1|T} - \boldsymbol{\mu}_k)' + P_{t-1|T}]\boldsymbol{\Phi}'\Big),
\end{aligned} \quad (23)$$

respectively. For the initial parameter set of the Markov-switching model we have to specify regime-specific mean for the slope factor. We distinguish between the two regimes by specifying a different mean as $\mu_{2,2} = 2 + \mu_{2,1}$, where $\mu_{2,1}$ is the estimated mean from the two-step approach. This way we ensure different slopes in each regime, and that regime 1 corresponds to a low inflation state and regime 2 to a high inflation state. We hereby also avoid potential identification problems. We initialise the transition probabilities at 0.95 and the initial state distribution at the long-term average. That is, $p_{11} = p_{22} = 0.95$ and $\boldsymbol{\rho} = (1/2, 1/2)'$. Again, we find that the estimation procedure is robust against precise choice for the initialisation as the EM algorithm quickly moves to reasonable estimates.

3.3 Step 4: Gradient-Based Optimisation

Lastly, step 4 optimises the likelihood function based on the prediction error decomposition. That is,

$$L(\boldsymbol{\theta}) = -\frac{nT}{2}\log(2\pi) - \frac{1}{2}\sum_{t=1}^{T}\log|\mathbf{F}_t| - \frac{1}{2}\sum_{t=1}^{T}\mathbf{v}_t'\mathbf{F}_t^{-1}\mathbf{v}_t, \tag{24}$$

where $\mathbf{v}_t$ is the prediction error at time t and $\mathbf{F}_t$ denotes the corresponding variance. These are both produced by the Kalman filter and further details are provided in Appendix B.1.

The factor-augmented model and shifting endpoints model do not need any adjustments to the likelihood function, Equation (24). For the Markov-switching model, we have to account for the regimes again. The approximate likelihood is as follows

$$L(\boldsymbol{\theta}) = \sum_{t=1}^{T}\log\left[\sum_{j=1}^{M}\sum_{i=1}^{M} f(\mathbf{y}_t|\mathcal{I}_{t-1}, S_t = j, S_{t-1} = i) \times \Pr[S_t = j, S_{t-1} = i|\mathcal{I}_{t-1}]\right], \tag{25}$$

where $f(\cdot)$ denotes a density function and the conditional density of $\mathbf{y}_t$ is a modified version of the prediction error decomposition. That is,

$$f(\mathbf{y}_t|\mathcal{I}_{t-1}, S_t = j, S_{t-1} = i) = (2\pi)^{-n/2}|\mathbf{F}_t^{(i,j)}|^{-1/2}\exp\left\{-\frac{1}{2}\mathbf{v}_t^{(i,j)\prime}\,\mathbf{F}_t^{(i,j)\,-1}\mathbf{v}_t^{(i,j)}\right\}. \tag{26}$$

This quantity is a by-product of the Kalman filter and Kim smoother, and further details are provided in Appendix B.2.

We feed the last updated parameter set from step 3 to the gradient-based optimisation method and subsequently optimise the likelihood given by either (24) or (25).

4 Data

In this section we discuss the data which serve as input for the described models in Subsection 2.2. Specifically, we discuss the yield data in Subsection 4.1 and the macroeconomic dataset in Subsection 4.2. In addition, we discuss the extracted factors from the panel of macroeconomic variables.

4.1 Yield Data

In our empirical analysis we monthly data of constant maturity yields of US government zero-coupon bonds. The data set is comprised of end-of-month yields and the sample period ranges from October 1993 to April 2019. The data is readily available in daily format from the Federal Reserve Economic

Database (FRED) of the Federal Reserve Bank of St. Louis.[5] We cover maturities of 3, 6, 12, 24, 36, 60, 84, 120, 240, and 360 months. We exclude maturities of 1 and 2 months due to their high sensitivity to policy changes.

Our sample period is chosen such that we obtain a balanced data set. This leaves us with $n = 10$ yield series with $T = 307$ time series observations per series. One side note has to be made. The 360-month yield series is actually not available for the whole sample period. This is due to that the Treasury department discontinued this yield series from February 19, 2002 through February 8, 2006. Instead, the Treasury department published extrapolation factors to compute 360-month yields. In this period we make use of their extrapolated yields. After this period of discontinuance they started publishing the regular series again from February 9, 2006 onward.[6]

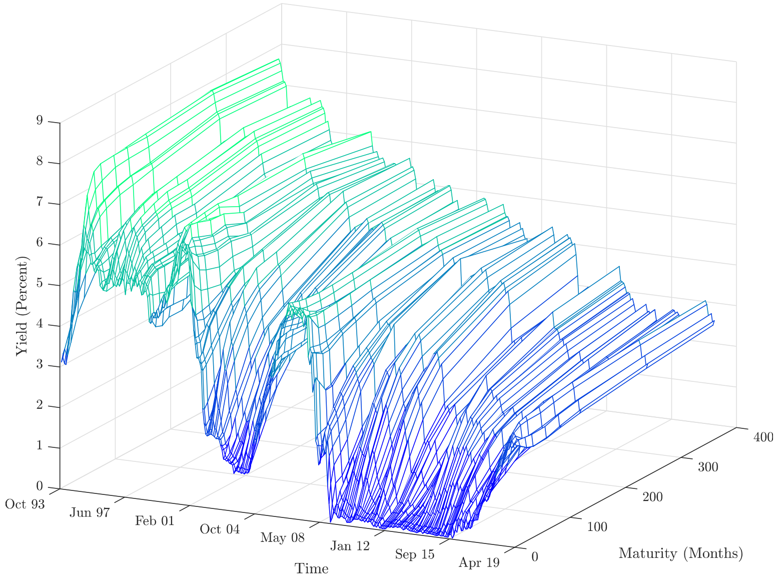


Figure 1: Plot of the three-dimensional surface of constant maturity yields over the period October 1993 - April 2019.

We provide a three-dimensional visualisation of the yield curve data in Figure 1. In addition, we present descriptive statistics for the yield series and empirical based proxies for the yield curve factors in Table 1. Recall that the level factor is the long-term yield, and the slope factor is the difference between the long-term yield and short-term yield. Then we define the level proxy as $y_t(360)$ and the slope proxy as $y_t(360) - y_t(3)$. Lastly, a proxy for the curvature of the yield curve can be defined as $[y_t(\tau^*) - y_t(3)] - [y_t(360) - y_t(\tau^*)]$ for some medium-term maturity, τ^*, and for a given t. Similar to Diebold and Li (2006), we choose the 24-month yield.

Figure 1 highlights how the yield curve takes on different shapes and dynamics over time. It also shows how the level fluctuates the most with a downward trend in the long term, as well as fluctuations in the slope and curvature, albeit to a lesser extent than the level. It can also be seen that the term

[5] https://www.treasury.gov/resource-center/data-chart-center/interest-rates/Pages/default.aspx

[6] For more details, see https://www.treasury.gov/resource-center/data-chart-center/interest-rates/Pages/TextView.aspx?data=longtermrate.

structure exhibit nonlinear dynamics which is especially visible around recession periods. Namely, we observe sharp decreases around the dot-com bubble of 2000 and the financial crisis of 2008. During these periods the term structure becomes less stable as the sudden changes in dynamics affect the volatility of interest rates. Moreover, there will likely be a higher degree of model uncertainty during these periods and therefore motivates the choice to account for model uncertainty.

Table 1: Descriptive statistics for the yield curve

Maturity (Months)	Mean	Std	Skew	Kurt	Min	Max	JB-p	$\hat{\rho}(1)$	$\hat{\rho}(12)$	$\hat{\rho}(30)$
3	2.460	2.167	0.277	1.462	0.000	6.380	0.000	0.995	0.823	0.437
6	2.593	2.204	0.268	1.472	0.030	6.510	0.000	0.996	0.823	0.442
12	2.723	2.206	0.261	1.521	0.090	7.200	0.000	0.995	0.833	0.479
24	2.986	2.185	0.271	1.619	0.200	7.690	0.001	0.993	0.841	0.543
36	3.190	2.101	0.265	1.693	0.300	7.800	0.001	0.992	0.842	0.582
60	3.576	1.915	0.246	1.836	0.590	7.830	0.002	0.989	0.837	0.630
84	3.887	1.782	0.241	1.909	0.980	7.840	0.003	0.987	0.830	0.653
120	4.129	1.635	0.239	2.027	1.460	7.910	0.005	0.985	0.818	0.660
240	4.649	1.577	0.096	2.000	1.780	8.100	0.007	0.985	0.828	0.687
360 (level)	4.686	1.417	0.198	2.138	2.180	7.990	0.010	0.983	0.806	0.665
Slope	2.226	1.375	-0.158	1.902	-0.640	4.570	0.003	0.976	0.552	-0.105
Curvature	-1.175	1.225	-0.107	2.006	-3.570	1.810	0.007	0.966	0.633	0.225

NOTE: This table presents the summary statistics of the monthly yields at different maturities over the complete sample period. We also present statistics for the yield curve's level, slope, and curvature. We define the level as the 360-month yield, the slope as the difference between the 360-month and 3-month yields, and the curvature is constructed by taking twice the 24-month yield, and subtracting the sum of the 3-month and 360-month yields from it. The three right columns display the sample auto-correlations for 1, 12, and 30 lags.

Table 1 confirms several stylised facts of the yield curve. Namely, we have that (i) the average yield curve is upward sloping and concave, (ii) volatility generally decreases with maturity, and (iii) yields of all maturities are persistent with long-term yields being more persistent. The second stylised fact also supports the motivation for the Markov-switching model, in the sense that long-term yields are more stable relative to short-term yields which was the basis for the reasoning behind this model. Further, we reject the Jarque-Bera test for normality for all maturities, as well as the proxies for the yield curve factors. Instead, we find that the yields are platykurtic meaning that the distribution of yields shows a lack of outliers. In addition, we find that the yields are positively skewed. Lastly, we find that the empirical factor proxies are very persistent as well, where the level proxy is more persistent than the slope and curvature proxies.

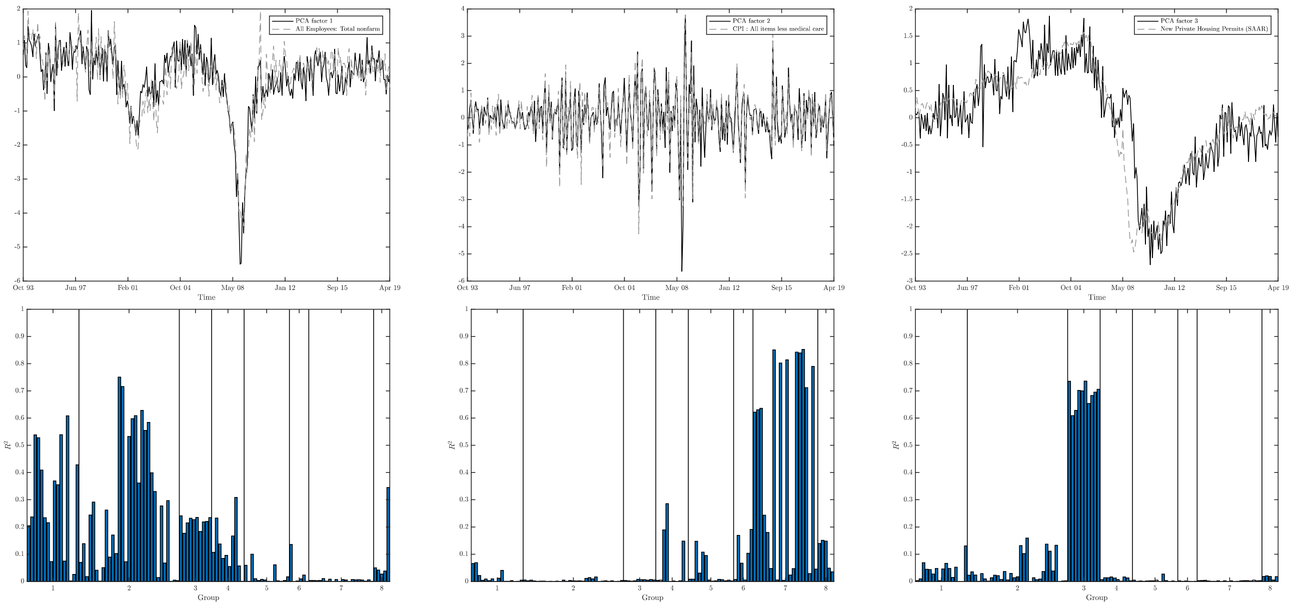
4.2 Macroeconomic Data

We consider a monthly macroeconomic data set covering 128 variables grouped in eight categories. The data set stems from McCracken and Ng (2016) and can be obtained from the FRED.[7] The macroeconomic variables are classified in the following eight groups: (1) output and income, (2) labour market, (3)

[7] https://research.stlouisfed.org/econ/mccracken/fred-databases/

housing, (4) consumption, orders, and inventories, (5) money and credit, (6) interest and exchange rates, (7) prices, and (8) stock market. This data set is frequently maintained and updated monthly by the Federal Reserve Bank of St. Louis. Additionally, McCracken and Ng (2016) show that this data set, in a statistical sense, contains the same predictive content as the vintage Stock and Watson (2005) data set.

Similar to De Pooter et al. (2010), we exclude all series related to interest rates and interest rate spreads except for the federal fund rates since it is closely related to the federal fund target rates. They argue that it could help capture movement in the short end of the yield curve as the federal fund target rates is the key instrument for monetary policy. After removing those series, we are left with 112 series. To obtain stationary series, we follow the transformations of McCracken and Ng (2016). We provide an overview of the macroeconomic series and their corresponding transformations in Table A.1 in Appendix A. Next, following Stock and Watson (2005), we replace outliers by the median of the last five observations, where we define an outlier as observations with absolute median deviations exceeding six times the interquartile range.



(a) PCA factor 1 (b) PCA factor 2 (c) PCA factor 3

Figure 2: The upper three figures are time series plots of the first three standardised principal components extracted from the macroeconomic dataset and the standardised individual macro series closest to the factor over the period October 1993 - April 2019. The lower three figures show the R^2 when the individual macro series are regressed on the factors. We show the results for the first three principal components in panels (a), (b), and (c), respectively. The groups are classified as (1) output and income, (2) labour market, (3) housing, (4) consumption, orders, and inventories, (5) money and credit, (6) interest and exchange rates, (7) prices, and (8) stock market.

Similar to previous literature, we extract a small number of factors of the large panel of macroeconomic data. We do so by applying principal component analysis (PCA). Over the full sample, the first ten extracted macro factors explain only 56% of the total variation in the panel of macroeconomic variables. In turn, the first three factors explain over half of the first ten factors. In particular, the first, second, and third factors explain 16%, 10%, and 8%, respectively. These results suggest that the factors only extract a fraction of the total variation of the macroeconomic dataset. One of the reasons could be that the panel of macroeconomic variables is too diverse. Again, McCracken and Ng (2016) show that the predictive content is statistically indistinguishable from the vintage Stock and Watson (2005) data set, we therefore proceed with the first three principal components. Besides parsimony and tractability, we choose the first three principal components as the remaining individual principal components capture 5% or less of the total variation.

To further analyse the extracted factors, we show time series plots of the first three standardised factors together with the individual standardised macroeconomic series that resembles the factor the most in Figure 2. In addition, we regress the individual macro series on the factors and plot the corresponding R^2 such that we can label the extracted factors from an economic point of view.

The first factor loads heaviest on the first and second categories which are output and income, and labour market, respectively. It can therefore be interpreted as a real activity/employment factor. Next, the explanatory power of the second factor is concentrated in category seven which mostly contains inflation measures. Thus, we can interpret this as the inflation factor. Lastly, the third factor is concentrated around the housing category and is therefore interpreted as such. The time series plots confirm these economic interpretations of the factors.

5 In-Sample Results

Before we analyse the underlying dynamics of the yield curve, we consider the in-sample fit of all our models. To this end, we examine several aspects of our models. In specific, we showcase the versatility of the Nelson-Siegel class and how the Markov-switching model links with economic theory. In addition, we assess in-sample fit by analysing filtered error statistics of all our models.

Figure 3 displays the yield curve of selected dates based on the data and the model-implied yield curves. This figure highlights the variety of yield curve shapes the Nelson-Siegel framework can replicate. On a visual inspection, the adjusted Svensson variants appear to provide the better fit which can be attributed to the additional hump factor allowing for more flexibility. In addition, we barely find any notable differences in terms of average fit for both the Nelson-Siegel and adjusted Svensson variants. This is mainly due to our model specifications. We have assumed that the cross-section of yield can be explained by only a small number of factors. However, all our alternative model specifications impose different structures on the dynamics of the factors rather than the yields. Therefore, one would expect notable differences to be in forecasting. The results of the average yield curve and the model-implied average yield curve also barely show any noticeable differences across model specifications. To conserve space, we show the average fitted yield curve in Appendix C. Further, we notice that all models have trouble fitting the 20-year yield as it is relatively high compared to the 30-year yield resulting in a local maximum.

As the Nelson-Siegel class has been studied thoroughly in the literature, we only briefly discuss the results of the Nelson-Siegel class before discussing the Markov-switching model. In Figure 4 we plot the filtered factor estimates for both the Nelson-Siegel and adjusted Svensson models. Additionally, we plot

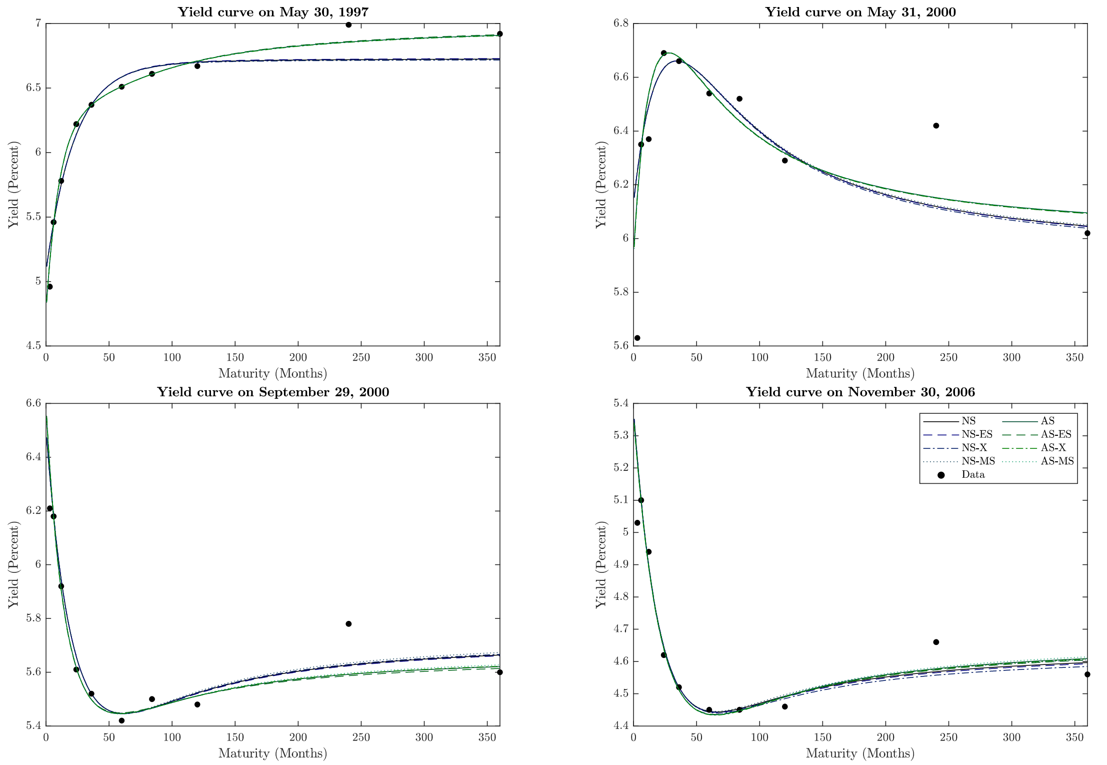


Figure 3: Actual yields and fitted yield curves for selected months. We show model-implied yield curves for all our models.

their empirical counterparts. We only show plots of the plain vanilla variants of the Nelson-Siegel class as there are barely noticeable differences between the different model specifications. Figure 4 shows how well the Nelson-Siegel class of models match their empirical factor counterparts. Besides the close match, we also observe sharp changes in factor estimates in and around recession periods which is especially visible in the slope factor. To put this in perspective with our different model specifications, this especially motivates the Markov-switching model as well as the shifting endpoints model. Those models account for nonlinearities in the factors, at least to some extent, which should aid in capturing the dynamics of the factors better.

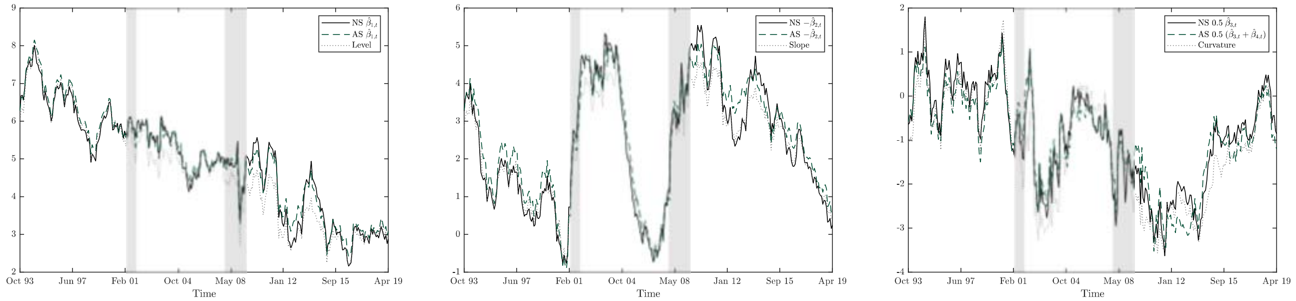


Figure 4: Time series plots of the filtered factor estimates for both the Nelson-Siegel and adjusted Svensson variants. In addition, we plot the empirical counterpart of the factor estimates. We define the level as the 360-month yield, the slope as the difference between the 360-month and 3-month yields, and the curvature is constructed by taking twice the 24-month yield, and subtracting the sum of the 3-month and 360-month yields from it. The shaded areas denote NBER recession periods.

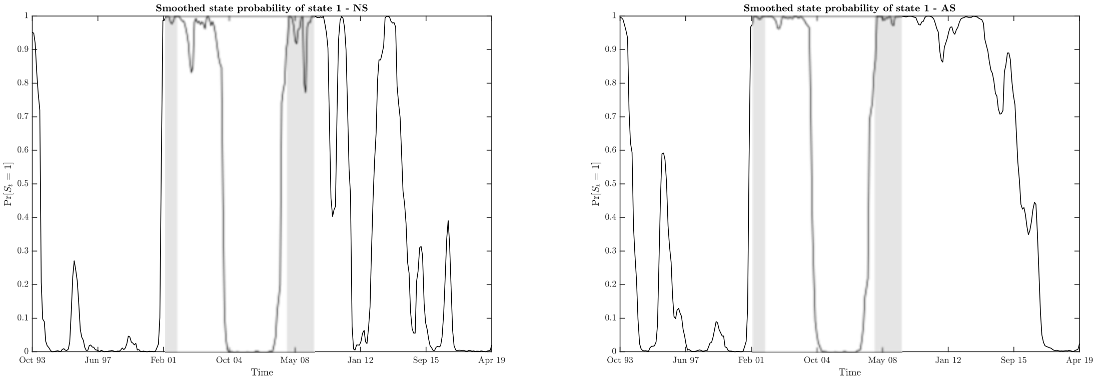


Figure 5: The left panel shows the smoothed state probabilities for the regime-switching Nelson-Siegel model. The right panel shows the smoothed state probabilities for the regime-switching Adjusted Svensson model. The shaded areas denote NBER recession periods.

To gauge to what extent the states of the Markov-switching model hold economic interpretations, we show plots of the smoothed probabilities of being in state 1 in Figure 5, and the average implied yield curve corresponding to a state in Figure 6, for the Nelson-Siegel and adjusted Svensson models. The latter is computed by separating the yield curve based on the smoothed probabilities. Similar to Hevia, Gonzalez-Rozada, Sola, and Spagnolo (2015), we split the sample with a dummy variable which equals 1 if $\Pr[S_t = 1|\mathcal{I}_T] > 0.5$ and 0 otherwise.

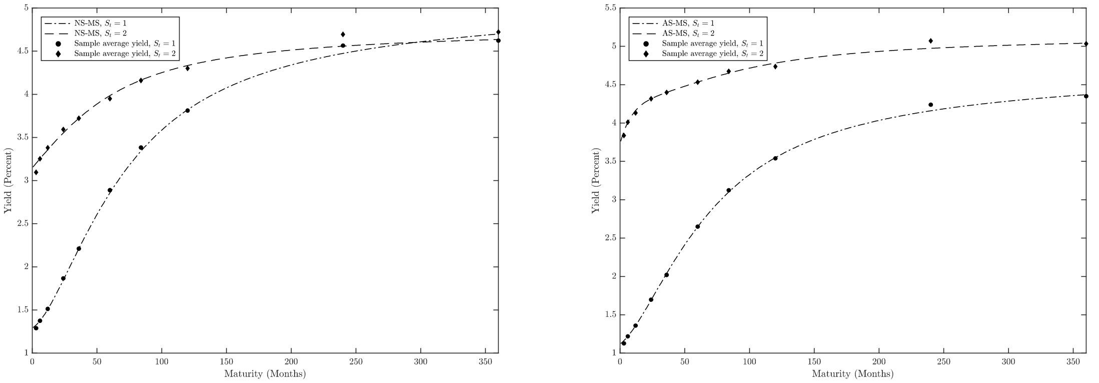


Figure 6: Plots of model-implied average regime specific yield curves and data. The sample is split according to a dummy variable which equals 1 if $\Pr[S_t = 1|\mathcal{I}_T] > 0.5$ and 0 otherwise, where $S_t = 1$ is the low inflation state and $S_t = 2$ is the high inflation state. The left panel shows the average model-implied fit for the Nelson-Siegel model and the right panel for the adjusted Svensson model.

Figure 5 shows that the smoothed probabilities coincide with the beginning of NBER recession periods which are given by the shaded bars. The smoothed probabilities also show persistence which implies that once we enter a state, we remain in that state for a prolonged period, where the adjusted Svensson model shows more persistence than the Nelson-Siegel model. From an economic perspective, persistence is also reassuring as the state of the economy is closely linked to the evolution of the business cycle which also progresses gradually.

Figure 6 confirms how the average shape of the yield curve based on the smoothed probabilities is connected to the state of the economy. On the one hand, it highlights how state 1 matches a low inflation

Table 2: Descriptive statistics in-sample fit

Maturity	NS	NS-ES	NS-X	NS-MS	AS	AS-ES	AS-X	AS-MS
Mean error								
3-month	-8.09	-8.09	-8.08	-8.08	-4.49	-4.37	-4.46	-4.40
6-month	0.00	0.00	0.00	0.00	0.14	0.20	0.17	0.20
1-year	1.68	1.68	1.67	1.67	-1.42	-1.40	-1.40	-1.39
2-year	3.36	3.36	3.34	3.34	0.37	0.38	0.37	0.36
5-year	-1.23	-1.26	-1.21	-1.23	0.03	0.02	0.03	0.03
10-year	-2.38	-2.50	-2.34	-2.45	-2.68	-2.77	-2.66	-2.67
Standard deviation								
3-month	12.08	12.07	12.09	12.07	10.55	10.44	10.45	**10.39**
6-month	0.00	**0.00**	0.00	0.00	0.38	0.54	0.47	0.55
1-year	7.39	7.39	7.40	7.38	5.92	5.91	**5.90**	5.91
2-year	5.26	5.26	5.26	5.26	1.42	1.43	1.42	**1.40**
5-year	6.51	6.51	6.50	6.50	2.11	2.12	**2.08**	2.10
10-year	4.49	4.59	**4.44**	4.66	5.19	5.18	5.17	5.20
Root mean squared error								
3-month	14.54	14.53	14.54	14.53	11.46	11.32	11.36	**11.29**
6-month	0.00	**0.00**	0.00	0.00	0.41	0.57	0.50	0.58
1-year	7.58	7.58	7.58	7.56	6.09	6.07	6.07	**6.07**
2-year	6.24	6.24	6.23	6.23	1.47	1.48	1.46	**1.44**
5-year	6.63	6.63	6.61	6.62	2.11	2.12	**2.08**	2.10
10-year	5.08	5.23	**5.01**	5.26	5.85	5.88	5.82	5.84
Mean absolute error								
3-month	11.07	11.06	11.07	11.06	8.25	8.19	8.15	**8.12**
6-month	0.00	**0.00**	0.00	0.00	0.31	0.43	0.38	0.44
1-year	5.82	5.81	5.81	5.80	4.42	4.43	**4.40**	4.41
2-year	5.06	5.06	5.05	5.05	1.12	1.11	1.11	**1.10**
5-year	5.40	5.41	5.40	5.39	1.59	1.57	**1.57**	1.58
10-year	4.10	4.21	**4.05**	4.26	4.72	4.73	4.71	4.71

NOTE: (continued on next page)

state characterised by a steep yield curve. On the other hand, it shows how state 2 corresponds to a high inflation state with a flat yield curve. The economic interpretation is less strong in the Nelson-Siegel model which is mainly due to the lack of persistence in smoothed probabilities. This suggests that the Nelson-Siegel specification has more difficulties identifying the current state of the economy.

In Table 2 we present detailed in-sample descriptive statistics for all models. In particular, we report the mean error, standard deviation, root mean squared error (RMSE), mean absolute error (MAE), minimum error, maximum error, and autocorrelation statistics for the filtered errors for selected maturities. The bold numbers indicate the best performing model assessed by standard criteria. Table 2 confirms that there are barely any differences across the different model specifications for both the Nelson-Siegel and adjusted Svensson variants. We find the Nelson-Siegel variants to be better in fitting the 3-month and 10-year yields whereas the adjusted Svensson variants fit the 6-month, 1-year, 2-year, and 5-year

Table 2: Descriptive statistics in-sample fit (continued)

Maturity	NS	NS-ES	NS-X	NS-MS	AS	AS-ES	AS-X	AS-MS
				Minimum error				
3-month	-65.47	-65.47	-65.71	-65.43	-61.85	**-60.63**	-61.69	-60.84
6-month	0.00	0.00	**0.00**	0.00	-0.93	-1.37	-1.13	-1.21
1-year	-30.43	-30.45	-30.36	-30.39	-28.90	**-28.37**	-28.70	-28.65
2-year	-9.87	-9.83	-9.88	-9.88	-3.62	-3.69	-3.55	**-3.54**
5-year	-18.89	-18.92	-18.82	-18.72	-5.96	-5.98	**-5.93**	-5.94
10-year	-13.31	-13.46	**-12.98**	-13.66	-15.25	-15.47	-15.14	-15.17
				Maximum error				
3-month	20.49	20.49	20.49	20.43	19.64	19.56	**19.41**	19.55
6-month	0.00	**0.00**	0.00	0.00	2.02	2.83	2.45	2.88
1-year	33.67	33.71	33.77	33.69	27.01	**26.81**	27.12	26.92
2-year	17.28	17.26	17.59	17.11	5.34	5.34	5.26	**5.24**
5-year	22.18	22.20	21.24	22.60	7.13	7.37	7.16	**7.11**
10-year	22.66	22.58	**19.84**	23.65	26.23	25.94	25.45	26.35
				$\hat{\boldsymbol{\rho}}_1$				
3-month	0.714	0.714	0.714	0.714	0.652	0.657	0.651	0.653
6-month	0.418	0.420	0.416	0.425	0.341	0.326	0.342	0.334
1-year	0.728	0.728	0.728	0.727	0.692	0.696	0.691	0.695
2-year	0.749	0.749	0.750	0.749	0.349	0.360	0.345	0.350
5-year	0.795	0.792	0.795	0.791	0.503	0.499	0.493	0.505
10-year	0.595	0.587	0.589	0.581	0.770	0.768	0.770	0.767
				$\hat{\boldsymbol{\rho}}_{12}$				
3-month	0.196	0.195	0.196	0.196	0.260	0.265	0.259	0.261
6-month	0.174	0.191	0.187	0.185	0.158	0.151	0.154	0.171
1-year	0.254	0.252	0.253	0.252	0.325	0.322	0.319	0.327
2-year	0.145	0.142	0.143	0.145	0.166	0.163	0.167	0.165
5-year	0.158	0.154	0.154	0.159	0.151	0.141	0.146	0.153
10-year	0.379	0.369	0.376	0.380	0.487	0.482	0.486	0.488

NOTE: Summary statistics for selected maturities of the in-sample fit based on filtered errors over the full sample from October 1993 - April 2019. The summary statistics are presented in basis points. Bold numbers highlight the best performing model for a maturity. The first and twelfth residual autocorrelation coefficients are denoted by $\hat{\boldsymbol{\rho}}_1$ and $\hat{\boldsymbol{\rho}}_{12}$, respectively.

yields better. The same pattern emerges in the persistence of the errors. Overall, both variants provide an accurate in-sample fit with average errors between -8.1 and 3.4 basis points for the Nelson-Siegel variant, and between -4.5 and 0.4 basis points for the adjusted Svensson variant. In addition, we find quite stable errors in terms of standard deviation ranging anywhere from 0 to 12.1 basis points. Judging from Table 2, the largest errors originate from the short-end of the yield curve with the 3-month yield and with the 1-year yield. Both variants provide the most accurate fit for medium-term maturities with the exception of the 6-month yield.

All in all, we find the Nelson-Siegel class to fit the data very accurately, where the adjusted Svensson model is marginally more accurate. Next, we barely observe any notable differences between the model specifications in terms of in-sample fit. Especially the extensions to the basic Nelson-Siegel and adjusted Svensson models perform slightly better in-sample.

6 Out-of-Sample Forecasting

In the previous section we have discussed the in-sample results of the yield curve models. This section shifts the focus from in-sample to out-of-sample performance. We divide this section in five parts. In Subsection 6.1 we elaborate on our forecasting setup and how forecasts are constructed. Next, in Subsection 6.2 we move on to forecasting combinations as a tool to account for model uncertainty. In Subsections 6.3 and 6.4 we discuss the statistical criteria we assess the forecasting performance with and the results, respectively. Lastly, in Subsection 6.5 we examine the robustness of the forecasting performance.

6.1 Forecasting Procedure

To examine the predictive ability of our models, we consider a recursive out-of-sample forecasting study. To this end, we split the full data sample in two different subsamples. We take the first ten years of data as initial estimation period, i.e. October 1993 - September 2003. Next, we have the forecasting period with the remaining observations, i.e. October 2003 - April 2019. We apply an expanding window in our recursive forecasting study. That is, we start with the initial estimation period to obtain parameters. Next, we use those parameters to make h-step ahead predictions of the h-step ahead yields, where h denotes the forecasting horizon. We then include the next observation to the estimation period and repeat these steps up to and including the last observation. We consider forecasts for 3, 6, 12, 24, 60, 84, and 120-month maturities, at horizons $h = 1, 3, 6, 12$ months ahead. In Figure 7 we show the cross-section of a subset of the yields we forecast, the initial estimation period, and subsequent subsamples we analyse for robustness. In particular, it is interesting to see how the models perform under different economic circumstances. The first subsample period ranges from October 2003 - October 2008, the second subsample period is the low interest rate environment from November 2008 - April 2015, and the last subsample period analyses the period from May 2015 onward. The subsamples can be distinguished with the dashed and dash-dotted line in Figure 7.

We construct forecasts by first iterating the transition equation forward to obtain factor predictions and then substituting the predictions in the measurement equation. Consider the general state space representation, based on (6), iterated factor forecasts for horizon h are constructed as

$$\hat{\boldsymbol{\beta}}_{t+h|t} = \hat{\boldsymbol{\mu}} + \hat{\boldsymbol{\Phi}}^{h}(\hat{\boldsymbol{\beta}}_{t|t} - \hat{\boldsymbol{\mu}}), \tag{27}$$

where $\hat{\boldsymbol{\beta}}_{t|t}$ is the filtered factor estimate at time t. The extensions to construct iterated forecasts for the factor-augmented model and the shifting endpoints model are trivial. The iterated forecasts for the Markov-switching model, on the other hand, are less trivial. To construct optimal h-step ahead forecasts,

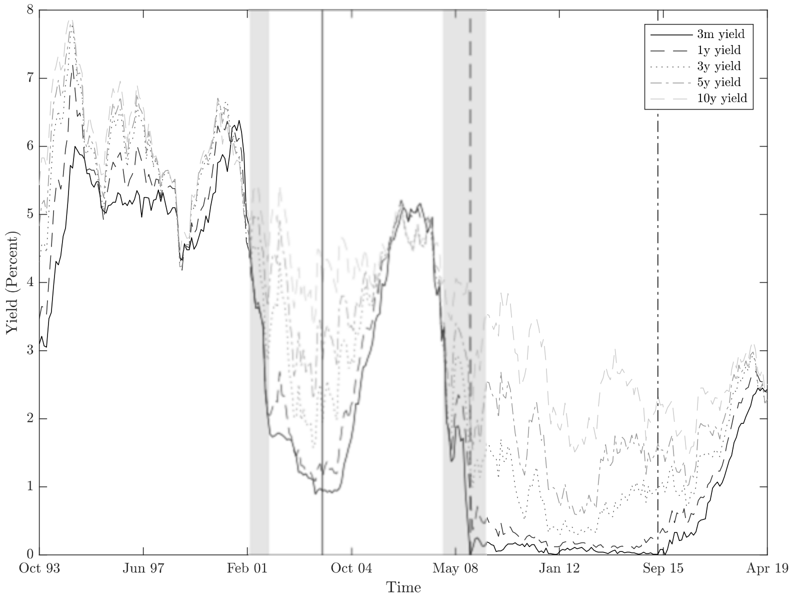


Figure 7: Time series plots of selected maturity yields over the period October 1993 - April 2019. We indicate the start of the forecasting period with the solid line. The dashed line and dash-dotted line splits the forecasting period in three different subsamples. The three subsample periods are from October 2003 - October 2008, November 2008 - April 2015, and May 2015 - April 2019, respectively. The shaded areas denote NBER recession periods.

we consider the expectation of the h-step ahead factor realisation conditional on the information set of time t. That is,

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{t+h|t} = \mathrm{E}[\boldsymbol{\beta}_{t+h}|\mathcal{I}_t] &= \mathrm{E}[\hat{\boldsymbol{\mu}}_{S_{t+h}} + \hat{\boldsymbol{\Phi}}(\boldsymbol{\beta}_{t+h-1} - \hat{\boldsymbol{\mu}}_{S_{t+h}})|\mathcal{I}_t] \\ &= \sum_{i=1}^{2} \Pr[S_{t+h} = i|\mathcal{I}_t]\big(\hat{\boldsymbol{\mu}}_i + \hat{\boldsymbol{\Phi}}(\hat{\boldsymbol{\beta}}_{t+h-1|t} - \hat{\boldsymbol{\mu}}_i)\big).\end{aligned} \tag{28}$$

Besides forecasts up until $\hat{\boldsymbol{\beta}}_{t+h|t}$, we also need to make predictions of the state probability at time $t+h$. This is done in a similar fashion as the prediction step of the Hamilton (1989) filter. That is,

$$\hat{\xi}_{t+h|t} = \hat{\mathbf{P}}^h \hat{\xi}_{t|t}, \tag{29}$$

where $\hat{\xi}_{t|t}$ denotes the filtered state estimate at time t, and $\hat{\mathbf{P}}$ is the estimated transition probability matrix.

6.2 Forecast Combinations

So far, we have only considered individual forecasts of several term structure models. However, previous forecasting literature shows that diversifying forecasts can yield significant gains in out-of-sample performance. Moreover, De Pooter et al. (2010) find promising results in combining individual model forecasts. By combining forecasts, one can account for model uncertainty and thereby make more robust forecasts. We consider two different types of forecast combination schemes: (i) simple combination schemes, and (ii) time-varying combination schemes.

The general combination set-up is based on a linear combination of L models. Denote the combined forecast for a h-month horizon and a yield with maturity τ_i as $\hat{y}_{t+h|t}(\tau_i)$, and define the corresponding forecast of model l as $\hat{y}_{l,t+h|t}(\tau_i)$. Then the combined forecast can be written as

$$\hat{y}_{t+h|t}(\tau_i) = \sum_{l=1}^{L} w_{l,t+h|t}(\tau_i)\hat{y}_{l,t+h|t}(\tau_i), \tag{30}$$

where $w_{l,t+h|t}(\tau_i)$ denotes the weight for the forecast of model l at time t. In the upcoming sections we discuss various methods to assign weights.

6.2.1 Simple Combination Schemes

The first scheme is a simple equally weighted forecast. That is, $w_{l,t+h|t}(\tau_i) = 1/L$ for $l = 1, ..., L$. In forecasting literature, this simple scheme has proven to be a difficult to beat benchmark. In addition, we consider a forecast scheme based on trimmed means which essentially removes "outliers" before computing the mean. This way, we reduce the influence of "outliers". We compute the trimmed mean with 5% symmetric trimming. The trimmed mean forecast scheme is enforced to trim at least one forecast. Thus, we remove the largest and smallest forecast prior to computing the mean.

6.2.2 Time-Varying Combination Schemes

On the one hand, the simple combination schemes provide a good starting point in combining forecasts. On the other hand, it would be more appropriate to assign weights to forecasting models based on their past performance. Intuitively, one would rather capitalise on the forecasts of models that have performed better in the recent past, than on the poor performing forecasts. To take the past performance of a forecasting model into account, one could choose to explicitly model the structure of the combination weights, or not. We consider the former approach.

To explicitly model the time-variation in the combination weights, we allow the weights to evolve smoothly via a random walk. This approach is used by, e.g. Sessions and Chatterjee (1989), LeSage and Magura (1992), and Stock and Watson (2004). Consider the following state space representation

$$y_{t+h}(\tau_i) = \mathbf{w}_t(\tau_i)'\hat{\mathbf{y}}_{t+h|t}(\tau_i) + \varsigma_{t+h}, \tag{31}$$

$$\mathbf{w}_t(\tau_i) = \mathbf{w}_{t-1}(\tau_i) + \boldsymbol{\vartheta}_t, \tag{32}$$

where $\mathbf{w}_t(\tau_i) = [w_{1,t+h|t}(\tau_i), ..., w_{L,t+h|t}(\tau_i)]'$, $\hat{\mathbf{y}}_{t+h|t}(\tau_i) = [\hat{y}_{1,t+h|t}(\tau_i), ..., \hat{y}_{L,t+h|t}(\tau_i)]'$, $\boldsymbol{\vartheta}_t = (\vartheta_{1,t}, ..., \vartheta_{L,t})'$, and the error terms are assumed to be orthogonal and normally distributed. Note that this is a stand-alone state space model from the models for $\hat{\mathbf{y}}_{t+h|t}(\tau_i)$. The time-varying weights can be extracted via the Kalman filter. In this state space model, the only unknown parameters are the variances of the error terms. In essence, these parameters can be estimated, however, we follow the approach of Stock and Watson (2004) and set the relative variance to $\text{var}(\vartheta_{l,t})/\text{var}(\varsigma_{t+h}) = \phi^2/L^2$, where ϕ is predetermined. The Kalman filter is initialised with the equally weighted combination scheme and zero variance. The degree of time-variation in the combination weights depend on the value of ϕ, where a smaller value for ϕ corresponds to less time-variation. In our study we analyse three values for ϕ, i.e. $\phi = 0.01, 0.05, 0.1$.[8] We label this weighting scheme as $\mathbf{TVW}(\phi)$.

When the relative variance is set too large, the preceding scheme allows for too much time-variation. Moreover, the time-varying weights are not bounded which results in extreme weights. To resolve this issue, one can restrict the weights to be on the unit interval. Billio et al. (2013) propose to do this for combining density forecasts. We adapt their methodology to combining point forecasts. First, introduce a new set of real-valued latent variables, $\mathbf{x}_t = (x_{l,t}, ..., x_{L,t})' \in \mathcal{X}$, and consider the following multivariate logistic transformation

$$w_{l,t} = \frac{\exp\{x_{l,t}\}}{\sum_{j=1}^{L} \exp\{x_{j,t}\}}, \qquad l = 1, ..., L. \tag{33}$$

Next, we need to specify dynamics for the new latent processes. We consider two alternative approaches. First, similar to the unrestricted time-varying combination scheme, we assume the weights to follow a random walk process. That is,

$$\mathbf{x}_t = \mathbf{x}_{t-1} + \boldsymbol{\upsilon}_t, \tag{34}$$

where $\boldsymbol{\upsilon}_t \sim \mathrm{N}(0, \boldsymbol{\Sigma}_\upsilon)$ with $\boldsymbol{\Sigma}_\upsilon = \mathrm{diag}(\sigma^2_{1,\upsilon}, ..., \sigma^2_{L,\upsilon})$. The second approach extends the random walk process with a learning mechanism. We essentially let the weights be explicitly driven by past performance. This way, one takes the distribution of the forecast errors into account and can subsequently assign weights more appropriately to recent errors. We do so by including an exponentially weighted moving average of forecast errors to the random walk process. Let H denote the moving window length, then the weighted moving average of forecast errors of model l is given by

$$e_{l,t} = (1 - \tilde{\alpha}) \sum_{i=1}^{H} \tilde{\alpha}^{i-1} V\big(y_{t+h}(\tau_i), \hat{y}_{l,t+h|t}(\tau_i)\big), \tag{35}$$

where $0 < \tilde{\alpha} < 1$ is the smoothing parameter and $V(\cdot)$ is a loss function for the forecast error. Essentially, one is free to choose whatever loss function one deems appropriate. We consider a fairly standard quadratic loss function. Define the distance vector, $\mathbf{e}_t = (e_{1,t}, ..., e_{L,t})'$, then the adaptive weighting scheme is given by

$$\mathbf{x}_t = \mathbf{x}_{t-1} - \Delta\mathbf{e}_t + \boldsymbol{\upsilon}_t, \tag{36}$$

where $\Delta\mathbf{e}_t = \mathbf{e}_t - \mathbf{e}_{t-1}$, and $\boldsymbol{\upsilon}_t \sim \mathrm{N}(0, \boldsymbol{\Sigma}_\upsilon)$ with $\boldsymbol{\Sigma}_\upsilon = \mathrm{diag}(\sigma^2_{1,\upsilon}, ..., \sigma^2_{L,\upsilon})$.

The latent weights, $\mathbf{w}_t$, are now constructed as a nonlinear combination of the latent variables, $\mathbf{x}_t$. The restricted time-varying combination scheme can therefore be summarised in a nonlinear state space model. Define the parameter vector as $\boldsymbol{\theta} = (\log\sigma^2_\varsigma, \log\sigma^2_{1,\upsilon}, ..., \log\sigma^2_{L,\upsilon}) \in \Theta$ and define the augmented

[8] The performance of this time-varying combination scheme strongly declined for higher values of ϕ.

state space vector $\mathbf{z}_t = (\mathbf{x}_t, \boldsymbol{\theta}) \in \mathcal{Z}$, where $\mathcal{Z} = \mathcal{X} \times \Theta$ is the augmented state space. Then we have the following distributional state space form

$$y_t|\mathbf{z}_t, \tilde{\mathbf{y}}_t \sim f(y_t|\mathbf{z}_t, \tilde{\mathbf{y}}_t), \tag{37}$$

$$\mathbf{z}_t|\mathbf{z}_{t-1}, y_{1:t-1}, \tilde{\mathbf{y}}_{1:t-1} \sim f(\mathbf{z}_t|\mathbf{z}_{t-1}, y_{1:t-1}, \tilde{\mathbf{y}}_{1:t-1}), \tag{38}$$

$$\mathbf{z}_0 \sim f(\mathbf{z}_0), \tag{39}$$

where we let $y_t = y_{t+h}(\tau_i)$ and $\tilde{\mathbf{y}}_t = \hat{\mathbf{y}}_{t+h|t}(\tau_i)$, for brevity. These densities can be used to define the predictive and filtering densities as

$$f(\mathbf{z}_{t+1}|y_{1:t}, \tilde{\mathbf{y}}_{1:t}) = \int_{\mathcal{Z}} f(\mathbf{z}_{t+1}|\mathbf{z}_t, y_{1:t}, \tilde{\mathbf{y}}_{1:t}) f(\mathbf{z}_t|y_{1:t}, \tilde{\mathbf{y}}_{1:t}) d\mathbf{z}_t, \tag{40}$$

$$f(\mathbf{z}_{t+1}|y_{1:t+1}, \tilde{\mathbf{y}}_{1:t+1}) = \frac{f(y_{t+1}|\mathbf{z}_{t+1}, \tilde{\mathbf{y}}_{t+1}) f(\mathbf{z}_{t+1}|y_{1:t}, \tilde{\mathbf{y}}_{1:t})}{f(y_{t+1}|y_{1:t}, \tilde{\mathbf{y}}_{1:t})}, \tag{41}$$

respectively. Unfortunately, it is not possible to solve the predictive and filtering densities analytically for the restricted time-varying combination model. To circumvent this problem, we resort to sequential Monte Carlo (SMC) methods (see Doucet, De Freitas, and Gordon (2001) for more details and an introduction to SMC methods) to approximate filtering. In particular, we employ a regularised particle filter. The key idea behind particle filters is to represent the posterior distribution by a set of random samples weighted accordingly from the posterior distribution. One is then able to perform inference based on the set of random samples. A prevalent problem in particle filtering is the degeneracy problem. To reduce this problem, it was suggested to include a resample move in the particle filter algorithm. Nonetheless, resampling came with their own problems. Specifically, the problem of loss of diversity in particles which was due to samples being drawn from discrete distributions. This can result in "particle collapse" where all particles would take the same point in the state space leaving us with a poor representation of the posterior distribution (Arulampalam, Maskell, Gordon, and Clapp, 2002). To avoid this problem, we use a regularised version of the particle filter. Below we give an outline of the SMC algorithm based on Casarin, Grassi, Ravazzolo, and van Dijk (2015).

Define $\Xi_t = \{\mathbf{z}_{t-1}^i, \gamma_{t-1}^i\}_{i=1}^N \in \mathcal{Z}^N$ to be the particle set containing N weighted random variables. At each iteration $t+1$, we assume that the weighted particle set, Ξ_t, approximating the filtering density $f(\mathbf{z}_t|y_{1:t}, \tilde{\mathbf{y}}_{1:t})$ is available. Then the empirical distribution of the filtering density $f(\mathbf{z}_t|y_{1:t}, \tilde{\mathbf{y}}_{1:t})$ is approximated by

$$f_N(\mathbf{z}_t|y_{1:t}, \tilde{\mathbf{y}}_{1:t}) = \sum_{i=1}^{N} \gamma_t^i \delta_{\mathbf{x}_t^i}(\mathbf{x}_t) K_h(\boldsymbol{\theta}_t - \boldsymbol{\theta}_t^i), \tag{42}$$

where $\delta_x(y)$ represents the Dirac's mass centred at x, and $K_h(x) = h^{-n_\theta} K(x/h)$ is the rescaled kernel density, $K(\cdot)$, with $h > 0$ a smoothing factor (kernel bandwidth). We denote the dimension of $\boldsymbol{\theta}$ by n_θ and the kernel density is a symmetric probability density function satisfying

$$\int xK(x)dx = 0 \qquad \text{and} \qquad \int ||x||^2 K(x)dx < \infty. \tag{43}$$

We can now approximate the predictive density of $\mathbf{z}_{t+1}$, conditional on $y_{1:t}$ and $\tilde{\mathbf{y}}_{1:t}$, Equation (40), as

$$f_N(\mathbf{z}_{t+1}|y_{1:t},\tilde{\mathbf{y}}_{1:t}) = \sum_{i=1}^{N} \gamma_t^i f(\mathbf{x}_{t+1}|\mathbf{x}_t,\boldsymbol{\theta}_{t+1},y_{1:t},\tilde{\mathbf{y}}_{1:t})\delta_{\mathbf{x}_t^i}(\mathbf{x}_t)K_h(\boldsymbol{\theta}_{t+1}-\boldsymbol{\theta}_t^i). \tag{44}$$

After observing the information at time $t+1$, we can update the states via an approximation of the filtering density, Equation (41), as follows

$$f_N(\mathbf{z}_{t+1}|y_{1:t+1},\tilde{\mathbf{y}}_{1:t+1}) \propto \sum_{t=1}^{N} \gamma_t^i f(y_{t+1}|\mathbf{z}_{t+1},y_{1:t},\tilde{\mathbf{y}}_{1:t}) f(\mathbf{x}_{t+1}|\mathbf{x}_t^i,\boldsymbol{\theta}_{t+1},y_{1:t},\tilde{\mathbf{y}}_{1:t})K_h(\boldsymbol{\theta}_{t+1}-\boldsymbol{\theta}_t^i). \tag{45}$$

We proceed with the following steps, at $t=0$, we initialise the particle filter with the independent particle set $\Xi_0 = \{\mathbf{z}_0^i, \gamma_0^i\}_{i=1}^N$ comprised of N iid random variables, $\mathbf{z}_0^i$, with corresponding random weights, γ_0^i. For the prior of the log variances we assume a lognormal random walk process. At $t>0$, given Ξ_t, for $i=1,...,N$:

1. Generate $\boldsymbol{\theta}_{t+1}^i$ from $\boldsymbol{\theta}_{t+1}^i \sim \text{LN}(\boldsymbol{\theta}_t^i, b^2)$, where b denotes the smoothing factor of the regularisation step.

2. Generate $\mathbf{x}_t$ from $\mathbf{x}_{t+1}^i \sim \text{N}(\mathbf{x}_t^i, \tilde{\boldsymbol{\Sigma}}_v)$. To include the learning mechanism, we simply generate $\mathbf{x}_t$ from $\mathbf{x}_{t+1}^i \sim \text{N}(\mathbf{x}_t^i - \Delta\mathbf{e}_t, \tilde{\boldsymbol{\Sigma}}_v)$.

3. Update the weights as follows

$$\tilde{\gamma}_{t+1}^i \propto \gamma_t^i \text{exp}\left\{-\frac{(y_{t+1}-\sum_{l=1}^L w_{l,t}^i \tilde{y}_{l,t+1})^2}{2\tilde{\sigma}_\varsigma^2}\right\}, \tag{46}$$

where $w_{l,t}^i = \exp\{x_{l,t}^i\}/\sum_{j=1}^L \exp\{x_{j,t}^i\}$ and $\gamma_{t+1}^i = \tilde{\gamma}_{t+1}^i/\sum_{j=1}^N \tilde{\gamma}_{t+1}^j$, for $i=1,...,N$.

4. Evaluate the Effective Sample Size (ESS) and resample if the ESS is below a certain threshold. The ESS is a measure for the overall efficiency of an importance sampling algorithm and is defined as

$$\text{ESS}_{t+1} = \frac{N}{1+N\sum_{i=1}^N \Big(\gamma_{t+1}^i - N^{-1}\sum_{i=1}^N \gamma_{t+1}^i\Big)^2 / \Big(\sum_{i=1}^N \gamma_{t+1}^i\Big)^2}. \tag{47}$$

If $\text{ESS}_{t+1} < c$, simulate $\{\tilde{\mathbf{z}}_{t+1}^i\}_{i=1}^N$ from $\{\mathbf{z}_{t+1}^i, \gamma_{t+1}^i\}_{i=1}^N$ (multinomial resampling) and set $\Xi_{t+1} = \{\tilde{\mathbf{z}}_{t+1}^i, \frac{1}{N}\}_{i=1}^N$. If $\text{ESS}_{t+1} \geq c$, set $\Xi_{t+1} = \{\mathbf{z}_{t+1}^i, \gamma_{t+1}^i\}_{i=1}^N$.

We employ the SMC algorithm with a slightly adjusted version of the suggested settings by Casarin et al. (2015). In specific, we use 1000 particles, $b=0.01$, $c=0.7$, $H=12$, $\tilde{\alpha}=0.95$, and the variances are initialised at 0.3. Contrary to the previous time-varying weighting scheme, we estimate the variances in this scheme. The reason for this is that we have a nonlinear state space model which makes it unclear how to set the relative variances.[9] Therefore, we proceed by estimating the variances. The exact initialisation of the variances does not considerably affect the results, provided that the variances are not initialised at too high values. Next, the combination of 1000 particles and an ESS threshold of 0.7 provides a good balance between precision and computation time. In terms of the settings for the learning mechanism, we find that the results are not noticeably affected by the moving window length. Therefore, we choose to include the last 12 observations in the moving window which is effectively the past year. For the smoothing parameter, $\tilde{\alpha}$, we find the results to be robust to exact initialisation for larger values of $\tilde{\alpha}$, whereas smaller values result in worse performance. On a final note, we have established these settings with limited experimentation as this scheme is computationally very expensive. Thus, we do not rule out additional improvements. We label these weighting schemes by **Con. TVW** and **Con. TVW**$(\tilde{\alpha})$, where the former is without a learning mechanism and the latter includes the learning mechanism.

6.3 Forecasting Evaluation

To assess the predictive ability of candidate forecasts, we consider several performance metrics relative to a benchmark. The random walk forecast is notorious for being hard-to-beat. Especially for term structure models as yield are close to stationary. Therefore, it serves as a useful benchmark. The random walk model can be formulated as

$$y_t(\tau_i) = y_{t-1}(\tau_i) + \nu_t(\tau_i), \qquad \nu_t(\tau_i) \sim \mathrm{N}\big(0, \sigma_\nu(\tau_i)\big), \tag{48}$$

where we denote $y_t(\tau_i)$ as the yield for maturity τ_i at time t. The h-step ahead forecast for the random walk model is simply the last observed yield, i.e. $\hat{y}_{t+h|t}(\tau_i) = y_t(\tau_i)$. We label these forecasts as **RW**.

Similar to De Pooter et al. (2010), we also report the RMSPE and the Trace Root Mean Squared Prediction Error (TRMSPE). The latter is a multivariate extension of the RMSPE, in the sense that it summarises the performance of all yields in one single statistic. It is an adaptation of the trace Mean Squared Error (MSE) as defined in Christoffersen and Diebold (1998). More specifically, for each horizon, we compute the TRMSPE by taking the square root of the sum of the MSPE of all yields considered. Although this performance metric is useful, it is still limited, in that we summarise the performance in one statistic over the complete forecasting period. To compensate for this limitation, we consider the Trace Cumulative Squared Prediction Error (TCSPE). This metric allows us to graphically analyse the model's performance over time relative to a benchmark for which we take the random walk forecasts. For model l, we calculate at time T the TCSPE as

$$\mathrm{TCSPE}_{l,T} = \sum_{t=1}^{T} \left[\sum_{i=1}^{n} \big(y_{t+h}(\tau_i) - \hat{y}_{\mathrm{RW},t+h|t}(\tau_i)\big)^2 - \big(y_{t+h}(\tau_i) - \hat{y}_{l,t+h|t}(\tau_i)\big)^2 \right], \tag{49}$$

where $y_{t+h}(\tau_i)$ is the observed maturity τ_i yield at time $t+h$ and $\hat{y}_{l,t+h|t}(\tau_i)$ denotes the corresponding

[9] We have tried several combinations similar to Stock and Watson (2005), however, the results were worse than the current settings.

forecast for model l. An increasing series indicates outperformance of model l relative to the benchmark, and vice versa.

To evaluate the statistical significance of out-of-sample forecasts, we perform the Diebold and Mariano (1995) test for equal predictive accuracy. We test all our candidate forecasts against the random walk forecast. A small discussion for the applicability of the Diebold-Mariano (DM) test is in order. It has been pointed out that the standard critical values for the DM-test are not valid anymore for nested models under an expanding window (West, 2006). This does not pose a problem for the individual model forecasts as we test against the random walk forecasts. However, we do include the random walk forecasts in the model set for the forecast combinations. Therefore, we do not necessarily have non-nested forecasts. Although this essentially renders the critical values of the standard DM-test invalid, it has been noted by Diebold (2015) that the standard DM-test critical values are an asymptotically trustworthy approximation. We therefore proceed with the standard DM-test. Though, results should be interpreted with caution.

6.4 Forecasting Results

We present the forecasting results of the individual models and forecast combinations for the complete forecasting period October 2003 - April 2019 in Tables 3 and 4, respectively. In particular, the first row of each panel presents the (T)RMSPE for the random walk. The subsequent rows in a panel present the relative (T)RMSPE of a model with respect to the random walk. Each panel presents the results for one of the forecasting horizons $h = 1, 3, 6, 12$. Outperformance of a model relative to the random walk forecasts are highlighted with bold numbers, and significant outperformance is denoted by $(\cdot)^5$ and $(\cdot)^1$ at significance levels of 95% and 99%, respectively. In addition, we present plots of the TCSPE for each model and each horizon in Figures 8 and 9 for the individual models and forecast combinations, respectively. Note that in Figure 9d we have excluded the results of TVW(0.10) as the results were very extreme and distorted the remaining results. Instead, we provide the TCSPE plot including the results of TVW(0.10) in Figure C.2 of Appendix C.

6.4.1 Individual Models Results

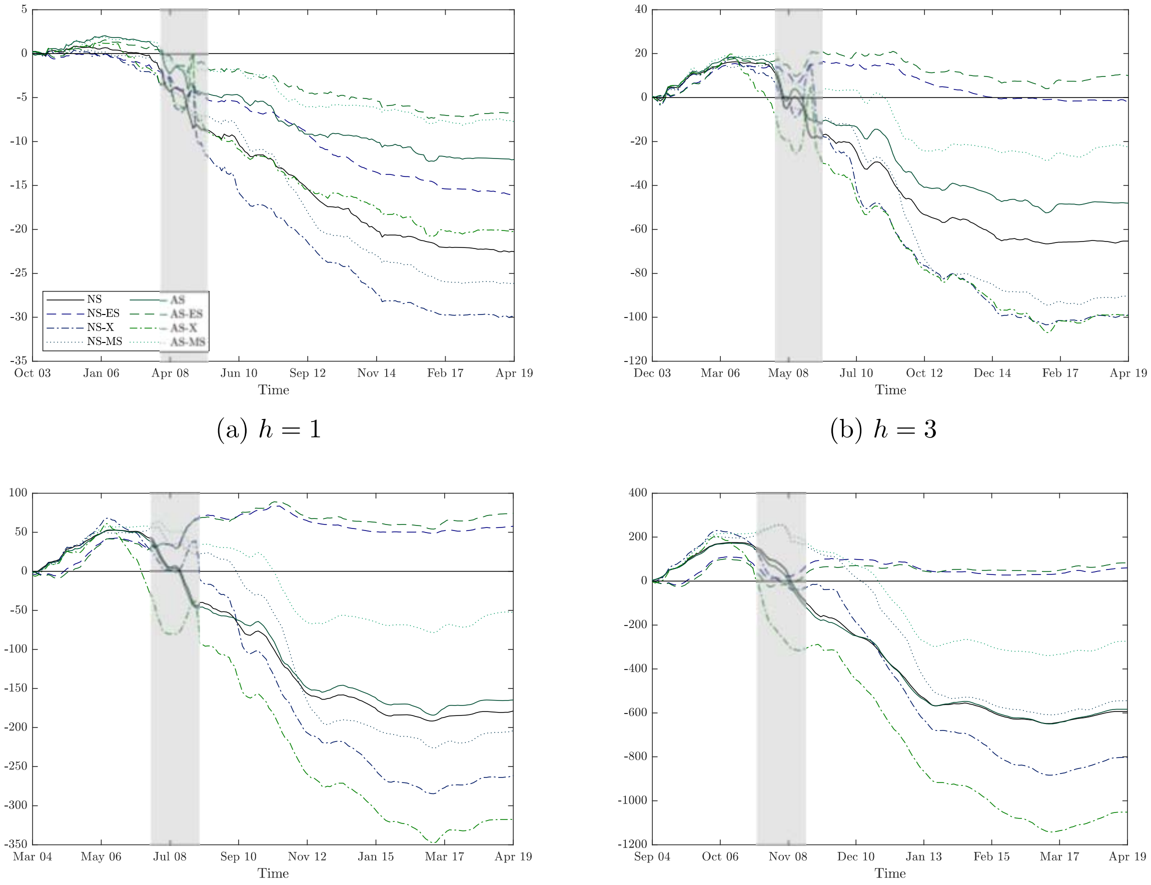
In examining the TRMSPE of the different forecasting horizons of Table 3, we do not find a consistently outperforming model. Especially at a 1-month horizon we find no model outperforming the no-change forecasts as all TRMSPEs are larger than 1. As noted by Diebold and Rudebusch (2013), over such a short horizon, yields barely have sufficient time for mean reversion and thus no-change forecasts are unlikely to be beaten. On the other hand, at higher horizons, we do find the shifting endpoints models to be performing well, where the adjusted Svensson version does marginally better. Again, using the same intuition, over longer horizons yields exhibit mean reversion which can be captured by models and no-change forecasts are beaten on a more consistent basis.

Table 3: Out-of-sample forecasting results, October 2003 - April 2019

	TRMSPE	**RMSPE**						
Maturity	All	3m	6m	1y	2y	5y	7y	10y
Panel A: 1 month								
RW	67.91	19.04	17.85	17.55	20.05	23.75	24.26	23.74
NS	1.12	1.44	1.06	**0.99**	1.06	1.05	1.03	1.01
NS-ES	1.09	1.36	**0.98**	**0.95**	1.02	1.01	1.01	**0.98**
NS-X	1.16	1.31	1.07	1.11	1.11	1.06	1.05	1.07
NS-MS	1.14	1.50	1.13	1.02	1.05	1.07	1.04	1.02
AS	1.07	1.36	1.05	**0.95**	1.04	1.02	1.00	**0.99**
AS-ES	1.04	1.25	**0.98**	**0.93**	1.00	1.00	1.00	**0.99**
AS-X	1.11	1.25	1.08	1.10	1.09	1.05	1.04	1.06
AS-MS	1.04	1.27	1.00	**0.92**^{5}	1.01	1.01	**0.99**	**0.98**
Panel B: 3 months								
RW	127.26	39.02	39.27	38.78	40.31	42.60	42.54	40.76
NS	1.10	1.24	1.08	1.06	1.10	1.09	1.06	1.05
NS-ES	1.00	1.10	**0.96**	**0.94**	**0.98**	**0.97**	**0.97**	**0.96**
NS-X	1.15	1.16	1.06	1.09	1.13	1.14	1.14	1.16
NS-MS	1.14	1.33	1.17	1.12	1.13	1.12	1.08	1.06
AS	1.08	1.23	1.09	1.05	1.09	1.05	1.03	1.02
AS-ES	**0.98**	1.06	**0.96**	**0.94**	**0.98**	**0.97**	**0.97**	**0.96**
AS-X	1.15	1.18	1.10	1.12	1.15	1.13	1.13	1.15
AS-MS	1.04	1.12	1.00	**0.98**	1.04	1.03	1.01	1.01
Panel C: 6 months								
RW	189.87	66.17	66.32	63.84	61.28	59.84	58.81	56.12
NS	1.13	1.16	1.09	1.10	1.15	1.15	1.12	1.11
NS-ES	**0.95**	1.00	**0.94**	**0.93**	**0.95**	**0.93**	**0.93**	**0.93**^{5}
NS-X	1.18	1.09	1.03	1.07	1.16	1.26	1.26	1.27
NS-MS	1.14	1.21	1.13	1.12	1.15	1.16	1.12	1.11
AS	1.12	1.17	1.10	1.10	1.15	1.12	1.09	1.08
AS-ES	**0.94**	**0.98**	**0.93**	**0.93**	**0.95**	**0.93**	**0.93**	**0.92**
AS-X	1.22	1.13	1.08	1.12	1.21	1.27	1.27	1.29
AS-MS	1.04	1.02	**0.97**	**0.98**	1.04	1.07	1.06	1.06
Panel D: 12 months								
RW	281.95	119.31	117.21	109.18	95.11	75.46	71.37	66.39
NS	1.19	1.11	1.08	1.11	1.22	1.34	1.31	1.31
NS-ES	**0.98**	1.00	**0.97**	**0.96**	**0.99**	**0.96**	**0.94**	**0.94**^{5}
NS-X	1.25	1.03	1.00	1.06	1.23	1.52	1.54	1.58
NS-MS	1.18	1.08	1.06	1.09	1.20	1.35	1.31	1.31
AS	1.19	1.12	1.10	1.13	1.22	1.32	1.28	1.27
AS-ES	**0.97**	1.00	**0.97**	**0.97**	1.00	**0.96**	**0.93**^{5}	**0.90**^{1}
AS-X	1.32	1.08	1.06	1.13	1.31	1.57	1.60	1.66
AS-MS	1.09	**0.98**	**0.96**	1.00	1.11	1.25	1.23	1.24

NOTE: This table reports the [Trace] Root Mean Squared Prediction Error ([T]RMSPE) for the random walk and individual models for selected yields. The first row contains the random walk statistics presented in basis points and the remaining rows contain the (T)RMSPE relative to the random walk. We highlight outperformance by bold numbers whereas $(\cdot)^5$ and $(\cdot)^1$ indicate significant different forecasts at 95% and 99% significance level according to the Diebold and Mariano (1995) test.

When we isolate the single maturity yields of Table 3, we notice the short-end of the yield curve to be particularly difficult to predict as we find the largest relative RMSPE at a 3-month yield. These results are, in a sense, contradicting to previous literature, for example compared to the results of De Pooter et al. (2010), however from the time series plot of Figure 7 we observe that our sample contains a period with short-rates close to zero for a prolonged period. This period especially affects the forecasting performance at the short-end of the yield curve. Further, most models tend to perform poor relative to no-change forecasts for longer maturities from the 2-year yield onward, where the poor performance becomes more pronounced for longer horizons. This is likely due to that longer maturity yields are more stable and therefore, for longer horizons, the no-change forecast approximates the true yield better than forecasts of one of our models which we can also see in the decrease in RMSPE for longer maturities and longer horizons for the no-change forecast. When we analyse the different model specifications, in contrast to in-sample results, we find quite some differences in forecasting results. Overall, in terms of TRMSPE, we find the weakest performing models to be the ones including macroeconomic factors, where it becomes more pronounced at longer maturities. Recall the time series plots of the factors in Figure 2, the second factor barely shows any persistence and as we forecast the factors as well in constructing



(a) $h = 1$

(b) $h = 3$

(c) $h = 6$

(d) $h = 12$

Figure 8: Plots of the Trace Cumulative Squared Prediction Error (TCSPE) for the individual models relative to the random walk forecasts over the complete forecasting sample October 2003 - April 2019. The shaded areas denote the NBER recession period.

h-month ahead forecasts, we accumulate forecasting errors which results in poor forecasts. Next, the shifting endpoints and Markov-switching extensions generally outperform the plain vanilla Nelson-Siegel and adjusted Svensson counterparts suggesting that these extensions describe the movement of yields better over time. The differences are more visible for longer maturities and the adjusted Svensson variant. The shifting endpoints extension allows the factors to better adjust to recent changes through the time-varying unconditional mean, whereas the Markov-switching forecasts are, in a sense, a combination of forecasts resulting in a more diversified forecast.

The DM test results in Table 3 are rather disappointing, in that they are more in favour of the random walk forecasts at the very least. Even though several models outperform the random walk, we barely find any significant outperformance for our models with a few exceptions. This suggests that, over the complete forecasting sample, we generally perform statistically similar or worse than random walk forecasts.

To gain a better understanding of a model's performance over time, we now move on to the TCSPE plots in Figure 8. First of all, we observe a substantial decrease during the global financial crisis of 2008 for all horizons and all models. Again, this period is characterised by extremely low short rates and large spreads between the long and short rates which held on for a longer than period than the NBER recession indicator as can be seen in Figure 7. Our models are unable to adequately adapt to these sudden changes in dynamics except for the shifting endpoints model, at least for larger horizons. In the period prior to the crisis, we generally find all models to be performing better relative to the random walk forecast with the exception of the 1-month horizon. It might also be worth noting how the performance of each model stabilises around mid 2015 for each horizon. Perhaps unsurprising as, from this period onward, the short rate starts to rise and the spread between the long rate and short rate decreases again to relatively normal levels. Next, these plots also highlight how the adjusted Svensson variant generally performs better than its Nelson-Siegel counterpart over all horizons. This implies that additional flexibility adds value not only in-sample but out-of-sample as well.

Focusing on the different dynamics, we observe a substantial increase in forecasting performance of models with macro factors during the recession period in contrast to the other extensions which is especially visible for horizons of 1, 3, and 6-months. However, this is only temporary as these models quickly resume their trend prior to the increase. For the 12-month horizon we observe a relatively stable performance of the factor augmented models. Further, the Markov-switching models are, prior to and during the recessions, among the best performing models which we observe especially for the adjusted Svensson variant. A plausible explanation for the substantial difference between the Nelson-Siegel and adjusted Svensson variants can be found in the persistence of transition probabilities. From the in-sample results we found that the Nelson-Siegel variant is less able to identify the current state. As we also forecast the future state, differences in persistence have an impact as well. Overall, this implies that, at least to some extent, the underlying dynamics of these models are able to capture the movements of yields over time with respect to our model set and the random walk. After the recession period these performances

Table 4: Out-of-sample forecasting results, October 2003 - April 2019

	TRMSPE	**RMSPE**						
Maturity	All	3m	6m	1y	2y	5y	7y	10y
Panel A: 1 month								
RW	67.91	19.04	17.85	17.55	20.05	23.75	24.26	23.74
Mean	1.04	1.22	**0.97**	**0.91**^{5}	**0.99**	1.00	1.00	**0.99**
Trimmed mean	1.04	1.24	**0.97**	**0.91**^{5}	**0.99**	1.01	1.00	**0.99**
TVW(0.01)	1.03	1.19	**0.97**	**0.91**^{5}	**0.99**	**0.99**	**0.99**	**0.98**
TVW(0.05)	1.01	1.08	**0.95**	**0.90**^{5}	**0.97**	**0.98**	1.00	**0.99**
TVW(0.10)	1.01	1.04	**0.92**	**0.89**^{5}	**0.96**	**0.99**	1.00	1.00
Con. TVW	1.04	1.23	**0.97**	**0.92**^{5}	**0.99**	1.01	1.00	1.00
Con. TVW($\tilde{\alpha}$)	1.04	1.21	**0.97**	**0.91**^{5}	1.00	1.02	**0.99**	**0.98**
Panel B: 3 months								
RW	127.26	39.02	39.27	38.78	40.31	42.60	42.54	40.76
Mean	1.02	1.06	**0.96**	**0.94**	1.00	1.01	1.01	1.01
Trimmed mean	1.02	1.07	**0.97**	**0.95**	1.00	1.02	1.01	1.01
TVW(0.01)	1.00	1.04	**0.95**	**0.94**	**0.98**	**0.99**	**0.99**	**0.98**
TVW(0.05)	**0.94**	**0.91**	**0.87**	**0.87**	**0.91**	**0.93**	**0.95**	**0.95**
TVW(0.10)	**0.90**	**0.81**^{5}	**0.79**^{5}	**0.80**	**0.86**	**0.91**	**0.94**	**0.95**
Con. TVW	1.02	1.07	**0.96**	**0.94**	**0.99**	1.01	1.02	1.01
Con. TVW($\tilde{\alpha}$)	1.02	1.07	**0.96**	**0.94**	**0.99**	1.02	1.02	1.01
Panel C: 6 months								
RW	189.87	66.17	66.32	63.84	61.28	59.84	58.81	56.12
Mean	1.01	1.00	**0.95**	**0.95**	1.00	1.04	1.04	1.04
Trimmed mean	1.02	1.00	**0.95**	**0.96**	1.01	1.05	1.04	1.05
TVW(0.01)	**0.97**	**0.97**	**0.93**	**0.93**	**0.98**	**0.99**	**0.98**	**0.98**
TVW(0.05)	**0.85**	**0.80**^{5}	**0.79**^{5}	**0.80**^{5}	**0.82**^{5}	**0.86**^{5}	**0.89**^{5}	**0.90**^{5}
TVW(0.10)	**0.78**	**0.66**^{5}	**0.67**^{5}	**0.68**^{5}	**0.71**^{5}	**0.82**^{5}	**0.86**^{5}	**0.88**^{5}
Con. TVW	1.02	1.00	**0.95**	**0.95**	1.01	1.05	1.05	1.04
Con. TVW($\tilde{\alpha}$)	1.02	1.00	**0.95**	**0.95**	1.01	1.05	1.04	1.05
Panel D: 12 months								
RW	281.95	119.31	117.21	109.18	95.11	75.46	71.37	66.39
Mean	1.05	**0.97**	**0.95**	**0.97**	1.06	1.17	1.17	1.18
Trimmed mean	1.06	**0.98**	**0.96**	**0.98**	1.06	1.18	1.18	1.19
TVW(0.01)	**0.98**	**0.93**	**0.91**	**0.93**	1.00	1.06	1.04	1.03
TVW(0.05)	**0.76**	**0.71**	**0.72**	**0.72**	**0.74**	**0.78**^{5}	**0.81**^{1}	**0.83**^{1}
TVW(0.10)	4.16	**0.59**	**0.60**	**0.60**	**0.62**	6.40	14.77	1.10
Con. TVW	1.05	**0.97**	**0.95**	**0.97**	1.05	1.17	1.16	1.19
Con. TVW($\tilde{\alpha}$)	1.06	**0.98**	**0.96**	**0.97**	1.08	1.20	1.17	1.20

NOTE: This table reports the [Trace] Root Mean Squared Prediction Error ([T]RMSPE) for the random walk and forecast combinations for selected yields. The first row contains the random walk statistics presented in basis points and the remaining rows contain the (T)RMSPE relative to the random walk. We highlight outperformance by bold numbers whereas $(\cdot)^5$ and $(\cdot)^1$ indicate significant different forecasts at 95% and 99% significance level according to the Diebold and Mariano (1995) test.

diminish which is due to the abrupt change in yield structure as explained earlier. Lastly, we want to highlight the performance of the shifting endpoints model. This model is able to quickly adapt to sudden changes to the yield curve which translates back in very stable performance with the exception of the 1-month horizon.

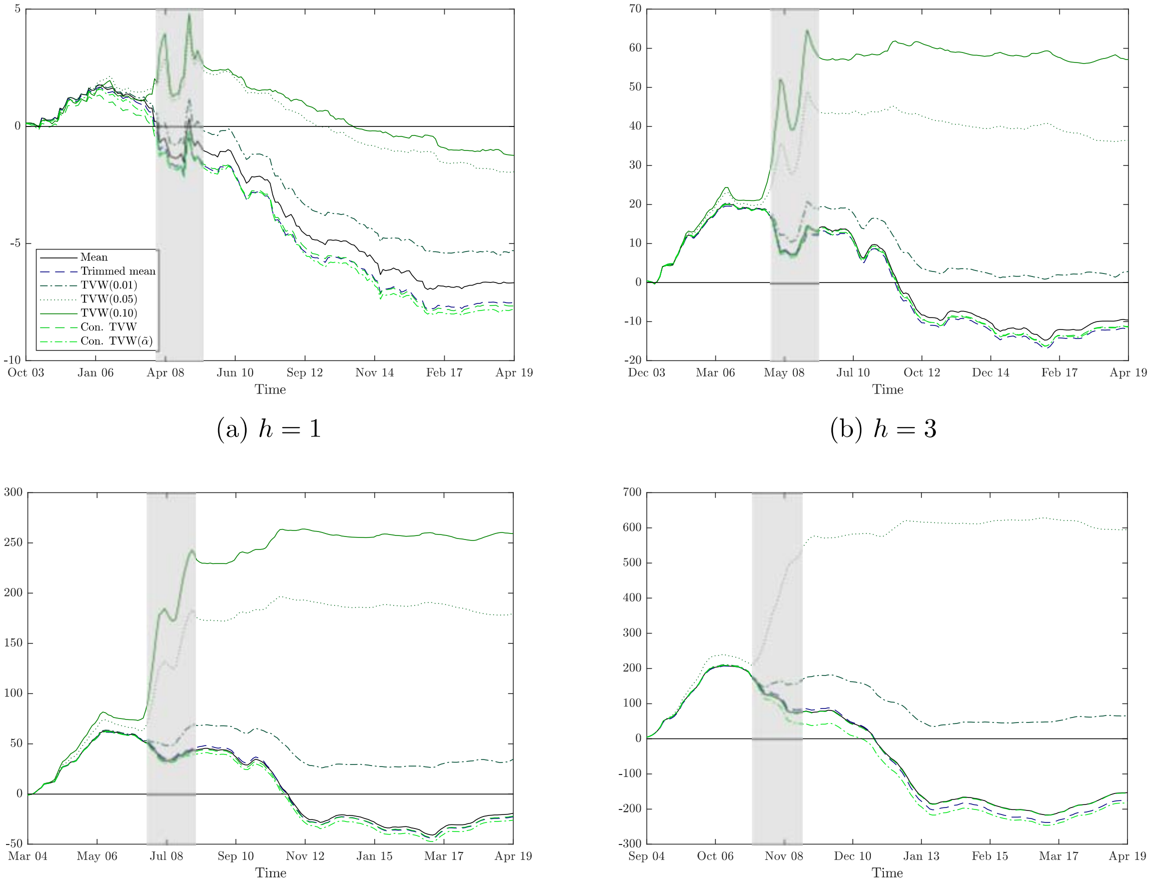
On a final note, these plots show us how no single model is able to outperform the random walk on a consistent basis and especially motivate to account for the uncertainty in the forecasts made by individual models. We analyse the merits of such an approach in the next section.

6.4.2 Forecast Combinations Results

From comparing the results of the individual models and forecast combinations, we can draw several conclusions. First and foremost, accounting for model uncertainty drastically improves forecasting performance relative to random walk forecasts, as well as individual model forecasts. We observe decreases of around 80% for all horizons of the lower bound of the TCSPE plots which translate back to what extent we underperform relative to the random walk. Similarly, we find increases in upper bound of the TCSPE plots of up to 200% for a 6-month horizon. For the remaining horizons we find more humbling results. Namely, no difference for a 1-month horizon, and increases up to 80% for the 6 and 12-month horizons. These results are also visible in Table 4 in contrast to Table 3 which shows more outperformance relative to individual models and, more importantly, stronger outperformance relative to the random walk forecasts. However, in terms of statistical significance, results remain rather disappointing, in the sense that we do not produce better forecasts than the random walk forecasts on a consistent basis.

Next, we want to emphasise the stability of the forecast performance under forecast combinations. Consider the TCSPE plots of the individual models and forecast combinations, we observe much smaller cumulative prediction errors, as well as less fluctuations over time. Where most individual models show decreasing series from the recession onward, we find that forecast combinations result in constant to even increasing series during the recession. On the one hand, the forecast combination series start to decrease in the period after the recession. On the other hand, these series stabilise much earlier than for individual models. Namely, we observe relatively stable series for forecast combination around fall 2012 whereas individual models start to stabilise around mid 2015. These results hold for all horizons except for the 1-month horizon which remains particularly difficult to forecast.

In examining the different methods of assigning weights, we find that all combination schemes have very similar movements in their TCSPE series. To ease discussion, we divide the forecast combination methods in two groups. The first group restricts the weights on the unit interval and the second group contains unrestricted weights. That is, the first group contains the mean, trimmed mean, and both variants of constrained time-varying weights. Next, the second group consists of the unrestricted time-varying weights. Up until around the recession we find all models to perform essentially identical. Just before the recession, the TCSPE series for both groups start to diverge. It is also around the period where the individual models start to diverge notably. The second group seems to be more capable in adapting to changes in dynamics than the first group. This can be attributed to the degree of time-



(a) $h = 1$

(b) $h = 3$

(c) $h = 6$

(d) $h = 12$

Figure 9: Plots of the Trace Cumulative Squared Prediction Error (TCSPE) for forecast combinations relative to the random walk forecasts over the complete forecasting sample October 2003 - April 2019. The shaded areas denote the NBER recession period.

variation allowed in the combination scheme as this allows the combination scheme to diverge quicker from the previous set of weights. In a sense, this is desirable as sudden changes can be accounted for, however, such freedom can quickly grow out of hand. An example of this is the TVW(0.10) scheme for a 12-month horizon. By not imposing restrictions or allowing for too much time-variation, we find that this particular scheme takes on too extreme weights. However, imposing restrictions does not necessarily improve on performance. For example, the constrained time-varying weighting schemes are not flexible enough to account for sudden changes as it performs virtually identical to the mean and trimmed mean. Overall, the better performance of the second group is particularly observed from around the recession until around fall 2012. In the subsequent period we barely observe any differences between the two groups as their TCSPE are relatively constant suggesting that both groups are performing similarly.

To further analyse the benefits of combining forecasts compared to individual model forecasts, we show plots of realised yields combined with model-implied forecasts in the left panels of Figure 10, and in the right panels we show the combined forecasts. More specifically, we show the forecasts and realised yields of March 2008 for forecast horizons $h = 1, 3, 6, 12$. Figure 10a shows very accurate forecasts over