Dynamic Calibration and Combination of Models Predictions*

Roberto Casarin† Dario Palumbo‡ Francesco Ravazzolo*

† Ca' Foscari University of Venice
‡ University Ca' Foscari of Venice and University of Cambridge
* BI Norwegian Business School, Free University of Bozen-Bolzano and RCEA

Abstract

This paper proposes a calibration and combination model that combines and dynamically calibrates predictive densities. While the weights are statically estimated the time-varying calibration is introduced giving an observation driven dynamics to the parameters of the calibrating function which is driven by the score of the assumed conditional likelihood of the data generating process. The model is very flexible and can handle different shapes, instability and model uncertainty in data generating process density. We show its effectiveness on various simulated datasets. Two empirical applications are also introduced, one on financial index density forecasts and one on short-term wind speed predictions. Both the simulations and the empirical applications documents the large instability of individual model performance in comparison to the properties of the combined and calibrated forecasts, favouring our model in terms of predictive accuracy.

Keywords: Heavy Tailed Distributions, Financial Returns Predictive Densities, Tail Risk, Climate Econometrics, Wind Speed Predictions, Model combinations, Calibration, Score-Driven Models, Bayesian Estimation.
JEL: C12, C18, C51, C52, C46, C58, G12.

*We are grateful for fruitful discussions, comments and suggestions at participants at "2021 EMCC-V: Econometric Models of Climate Change Conference". This paper is part of the research activities at the Centre for Applied Macroeconomics and commodity Prices (CAMP) at BI Norwegian Business School.

1. Introduction

Forecasting is a crucial task in time series, which becomes harder as the degree of uncertainty around the forecasted quantity increases. In certain contexts, such as risk management and finance, the selection *a priori* of the relevant information necessary to produce forecasts is not an easy task and, as a consequence, often several models with different assumptions are used to predict the same quantity. The recent challenges brought by more frequently occurring financial crisis, the appearance of more risky and unregulated assets, climate change have increased the level of uncertainty in financial and environmental variables while, on the other hand, have increased the demand for accuracy from practitioners, scientists and policy makers.

As showed by several authors, when multiple forecasts are available, it is possible to combine them to benefit simultaneously from the different approaches of processing the relevant information from the observations with the purpose of improving the prediction.

This paper builds up on this literature and proposes a dynamic model for the case when the optimal level of calibration, and the combination structure, varies across the time period forecasted. The model is then applied to produce accurate predictions of financial and climate variables.

Early papers on forecasting with models combinations are Barnard (1963), who considered airline passenger data, and Roberts (1965) who introduced a distribution which includes the predictions from two experts (or models). The combined distribution is a weighted average of the posterior distributions of two models and is similar to the result of a Bayesian Model Averaging (BMA) procedure, see Raftery et al. (1997). Raftery et al. (2005) extend the BMA framework by introducing a method for obtaining probabilistic forecasts from ensembles in the form of predictive densities and McAlinn and West (2018) extend it to Bayesian predictive synthesis.

Bates and Granger (1969) propose combination of predictions from different forecasting models using descriptive regression. Granger and Ramanathan (1984) extend this to combine forecasts with unrestricted regression coefficients as weights. Terui and van Dijk (2002) generalize the problem to a state space with weights that are assumed to follow a random walk process. Hoogerheide et al. (2010) propose robust time-varying weights and account for both model and parameter uncertainty in model averaging. Raftery et al. (2010) derive time-varying weights in dynamic model averaging, and speed up computations by applying forgetting factors in the recursive Kalman filter updating.

Combination weights that depend on (optimal) score functions have also been studied. Hall and Mitchell (2007) introduce the Kullback-Leibler divergence as a unified measure for the evaluation and suggest weights that maximize such a distance, see also Geweke and Amisano (2011) for a comprehensive discussion on how such weights are robust to model incompleteness, that is

the true model is not included in the model set. Gneiting and Raftery (2007) recommend strictly proper scoring rules. Billio et al. (2013) develops a general method that can deal with most of issues discussed above, including time-variation in combination weights, learning from past performance, model incompleteness, correlations among weights and joint combined predictions of several variables.

Finally, the last aspect relates to calibration and combinations. Ranjan and Gneiting (2010) and Gneiting and Ranjan (2013) introduces the idea of recalibration density forecasts when the density is not well-calibrated. They introduce a monotone non-decreasing map via a Beta distribution to achieve it. Bassetti et al. (2018) generalize to infinite Beta mixtures, allowing for more flexibility in calibrating and combinations in presence of fat tails, skewness and multiple-modes.

This paper extends the density calibration literature and proposes to apply combination with dynamic calibration models. We extend the static model in Ranjan and Gneiting (2010) to a score-driven dynamic model. The time varying calibration is obtaining giving dynamics to the parameters of the beta calibrating function with dynamics driven by the score of the conditional likelihood of the calibration-combination, see Creal et al. (2013) and Harvey (2013). Score-driven models for beta distribution have been already proposed by Gorgi and Koopman (2021), while in our case the beta distribution is used as a calibrating function. As we show in simulations, the model is very flexible and can handle different shapes, instability and model uncertainty. The parameters of the model are estimated through a Bayesian inference procedure and an efficient Monte Carlo Markov Chain sampler for posterior approximation as described by Łatuszyński et al. (2013).

Through a simulation results is possible to appreciate the flexibility of the model in recovering the dynamically the true density of the data generating process. The model is tested in four different and heterogeneous settings for potential time variation in the true density of the data generating process which include smooth transition in the mean, structural breaks and transition towards heavy tails. The model is parsimonious in the sense that each of the individual beta parameters tends to stretch or down-weight either the upper or lower quantiles of the *naive* static combination of predictive densities.

The empirical performance of the proposed model is then assessed on two different datasets. The first is a datasets of density forecars based on an S&P 500 Index returns already used in Amisano and Geweke (2010), Geweke and Amisano (2011), Kapetanios et al. (2015). The forecasts are obtained by standard GARCH models already used in Bassetti et al. (2018). The second dataset is of half hourly wind speed forecasts produced by various statistical state-of-the-art models on a dataset from a wind farm in Cairns Australia. In both the cases it is shown that the combined

dynamic Calibration and Combination approach provides superior density forecasts, both in terms of Log score and CRPS criteria.

The structure of the paper is organized as follows. Section Section 2 presents the novel dynamic calibration and combination model framework. Section Section 3 describes the Bayesian estimation and introduces model selection strategies to identify the presence of dynamics in the calibration part of the model. Section **??** presents four simulation studies. Section Section 5 shows two empirical applications while Section Section 6 sets up the conclusion.

2. Statistical Framework: Combination and Dynamic Calibration

Let $F_{1t}, ..., F_{Mt}$ be a sequence of predictive cumulative distribution functions (cdfs), for a real-valued variable of interest, y_t available at time t. Following the forecast combination and calibration literature, we assume the cdfs are externally provided. We consider a combination function which combines the sequence of cdfs into an aggregated predictive cdf, $F(\cdot|\boldsymbol{\gamma}) = G_{\boldsymbol{\gamma}}(\cdot|F_{1t}, \ldots, F_{Mt})$. Given a sequence of observations, $y_1, ..., y_T$, the cdf evaluated on one observation $F(y_t|\boldsymbol{\gamma})$, is referred as probability integral transform (PIT). Following Ranjan and Gneiting (2010) and Gneiting and Ranjan (2013), we say that the PITs, $F(y_1|\boldsymbol{\gamma}), ..., F(y_T|\boldsymbol{\gamma})$, are well calibrated (or probabilistically calibrated) if their distribution is uniform. The uniformity of the PITs are essential for density forecast evaluation.

Following the literature we consider a beta calibration function

$$F_t(y|\boldsymbol{\gamma}) = B_{\alpha,\beta}\left(\sum_{i=1}^{M} \omega_i F_{it}(y)\right) = B_{\alpha,\beta}\left(H_t(y|\boldsymbol{\omega})\right), \qquad y \in \mathcal{Y} \subset \mathbb{R} \tag{1}$$

where $B_{\alpha,\beta}(\cdot)$ is the cdf of a Beta distribution. The parameter vector is composed by $\boldsymbol{\gamma} = (\alpha, \beta, \boldsymbol{\omega})$, where $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_M)$. If the F_{it} admit probability density function (pdf) f_{it} then the pdf of the model in Equation (1) is

$$f_t(y|\boldsymbol{\gamma}) = b_{\alpha,\beta}\left(H_t(y|\boldsymbol{\omega})\right) h_t(y|\boldsymbol{\omega}), \tag{2}$$

where $b_{\alpha,\beta}(\cdot)$ is the pdf of the Beta distribution and linear combination density and distribution functions

$$h_t(y|\boldsymbol{\omega}) = \sum_{i=1}^{M} \omega_i f_{it}(y) \qquad H_t(y|\boldsymbol{\omega}) = \sum_{i=1}^{M} \omega_i F_{it}(y) \tag{3}$$

We propose a score-driven dynamic model for calibration and combination of predictive distributions. Dynamic Calibration in the model can be introduced giving dynamics to individual parameters of the calibration function α and β. The time-varying parameters can be fitted by an

observation driven model with dynamics driven by the score of the assumed conditional likelihood of the data generating process y_t (Creal et al. (2013), Harvey (2013)).

These class of models are called *score-driven* models, or otherwise called dynamic conditional score (DCS) models. This is a specific class of observation-driven models which assume that the variable y_t, conditionally on all the past observations up to $t-1$, is distributed with a distribution G as follows

$$y_t \mid \mathcal{F}_{t-1} \sim G(y_t|\boldsymbol{\theta}_{t|t-1}, \boldsymbol{\varphi}), \qquad t = 1, \ldots, T$$

where $\boldsymbol{\theta}_{t|t-1}$ is a sequence of dynamic parameters and the dynamics of each $\boldsymbol{\theta}_{t|t-1}$ at time t is conditional on the information at time $t-1$

$$\boldsymbol{\theta}_{t+1|t} = g(\boldsymbol{\theta}_{t|t-1}, \mathbf{u}_t; \boldsymbol{\varphi})$$

where $\boldsymbol{\varphi}$ is a time-invariant parameter and $\mathbf{u}_t$ is defined as the conditional score vector with respect to the dynamic parameters evaluated at y_t, that is

$$\mathbf{u}_t = \nabla_{\boldsymbol{\theta}} \ln f_t(y_t|\boldsymbol{\theta},\ \boldsymbol{\varphi}), \qquad t = 1, \ldots, T$$

The important feature of this framework is that it not only allows us to construct the likelihood for any form of the distribution F and it is robust even if the moments of the distributions don't exists. Moreover, it also allows us to model the dynamics of different parameters of the conditional distribution at the same time. At last, since the score is robust to the presence of outliers, it is capable in case of heavy-tailed observations to efficiently identify residual dynamics in the ACF of the fitted scores.

The parameters $\boldsymbol{\theta} = \{\alpha, \beta\}$ belong to a subspace $\Theta = \mathbb{R}^2_+$, while the time invariant parameters $\boldsymbol{\varphi} = \boldsymbol{\omega}$ belong to $\Phi = \mathbb{S}^{M-1}$ where $\mathbb{S}^{M-1} = \{\boldsymbol{\omega} \in \mathbb{R}^M,\ s.t.\ \boldsymbol{\omega}'\boldsymbol{\iota} = 1\}$ and in order to allows for an unrestricted DCS dynamics. Under this assumption the score vector is

$$\mathbf{u}_t = \begin{pmatrix} \left[\ln\left(H_{t|t-1}\left(y_t|\boldsymbol{\omega}_{\boldsymbol{t|t-1}}\right)\right) - \psi\left(\alpha_{t|t-1}\right) + \psi\left(\alpha_{t|t-1} + \beta_{t|t-1}\right)\right]\alpha_{t|t-1} \\ \left[\ln\left(1 - H_{t|t-1}\left(y_t|\boldsymbol{\omega}_{\boldsymbol{t|t-1}}\right)\right) - \psi\left(\beta_{t|t-1}\right) + \psi\left(\alpha_{t|t-1} + \beta_{t|t-1}\right)\right]\beta_{t|t-1} \end{pmatrix} \tag{4}$$

where $\psi(\cdot)$ is the digamma function.

In the score-driven models literature, the function $\mathbf{u}_t$ is scaled by a weighting factor S_t which if often considered as the inverse of the information matrix of the time varying parameters $\mathbf{I}_t\left(\boldsymbol{\theta}\right)^{-1}$, see Creal et al. (2013) and Harvey (2013).

In a parsimonious specification of the DCS model the conditional dynamics of the parameter considered is modelled by a first order Quasi-ARMA specification that can be represented as

$$\boldsymbol{\xi}_{t+1|t} = (I - \Phi)\,\boldsymbol{\nu} + \Phi\boldsymbol{\xi}_{t|t-1} + K\mathbf{u}_t, \qquad t = 1, \ldots, T$$

where $\Phi = \text{diagrv}(\phi_1, \phi_2)$, $\boldsymbol{\nu} = (\nu_1, \nu_2)'$, and $K = \text{diagrv}(\kappa_1, \kappa_2)$ are constant parameters to be estimated. Then $\boldsymbol{\varphi} = (\text{diag}(\Phi)', \boldsymbol{\nu}', \text{diag}(K)', \boldsymbol{\omega}')'$

2.1. Time-Varying Calibration Parameters

Giving the calibrating function in Equation (1), we can introduce dynamic calibration assuming that $\boldsymbol{\xi}_{t|t-1} = \left(\alpha_{t|t-1}, \beta_{t|t-1}\right)'$ while we have static combination, therefore $\boldsymbol{\omega} \in \boldsymbol{\varphi}$. The two shape parameters of the beta function will be allowed to vary over time and correct the quantiles of the combinated forecasted CDF $H_{t|t-1}\left(y_t|\boldsymbol{\omega}\right)$. The dynamics of the parameters will be driven by the score of the conditional distribution. Differently from the approach of Gorgi and Koopman (2021) we leave the general parameterisation of the Beta distribution setting a dynamics for each of the parameter independently.

Harvey and Luati (2014) and Harvey (2013) have showed that models with conditional distribution with heavier tails results in score-driven filters more robust to the impact of outliers. Score drive models applied to generalisation of the beta distribution have been used by Harvey and Palumbo (2022) for modelling realized volatility and Harvey (2013) for modelling range[1]. The dynamics of the parameters is driven by the conditional scores of the density in Equation (2).

$$\mathbf{u}_t = \frac{\partial \ln f_t\left(y_t|\boldsymbol{\theta}\right)}{\partial\left(\alpha_{t|t-1}, \beta_{t|t-1}\right)'} = \begin{cases} \ln\left(H_{t|t-1}\left(y_t|\boldsymbol{\omega}\right)\right) - \psi\left(\alpha_{t|t-1}\right) + \psi\left(\alpha_{t|t-1} + \beta_{t|t-1}\right) \\ \ln\left(1 - H_{t|t-1}\left(y_t|\boldsymbol{\omega}\right)\right) - \psi\left(\beta_{t|t-1}\right) + \psi\left(\alpha_{t|t-1} + \beta_{t|t-1}\right) \end{cases},$$

where $\psi\left(\cdot\right)$ is a digamma function.

Given that perfect calibration is achieved when $\alpha = \beta = 1$, as we can see from Section 2.1 while $H_{t|t-1}\left(y_t|\boldsymbol{\omega}_{t|t-1}\right)$ approaches 0 $u_{\alpha t}$ pushes *alpha* down β is pushed up so that the left quantiles gets more calibrated, while the opposite happens as $H_{t|t-1}\left(y_t|\boldsymbol{\omega}_{t|t-1}\right)$ approaches 1. The amount of calibration provided by the beta distribution is dictated by the magnitude of the two parameters thus, while the parameters are closer to 0 the response of the score is the largest, while it decreases as the value of the parameters approaches 1. When the parameters are larger than 1 the score curve flattens providing almost an identical response for each of the values of α and β.

[1]For a complete treatment of score-driven models applied to strictly positive location-scale distributions and generalised Beta distributions see Harvey (2013).

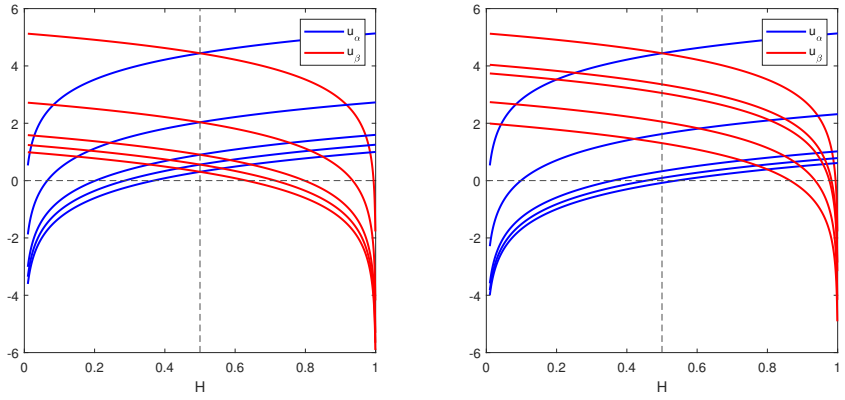

Figure 1: Value of the two scores, red line for the score wrt α and blue line for the score wrt β, for different values of $H_{t|t-1}(y_t|\boldsymbol{\omega})$. In the left plot $\alpha, \beta \in \{0.1, \ldots, 1\}$. In the right plot $\alpha \in \{0.1, \ldots, 1\}$ while $\beta \in \{0.1, \ldots, 0.5\}$

After imposing an exponential link function to constrain the parameters to be strictly positive the score becomes as in Equation (4).

3. Estimation and Model Selection

On the same line as Harvey et al. (2007), we propose a Bayesian approach to inference for our Calibration model which allows us to include extra-sample information about the model and the parameter through the prior choice.

3.1. Prior Distributions

For the vast majority of the parameters included in the full parameter vector $\boldsymbol{\varphi} = (\phi_1, \phi_2, \nu_1, \nu_2, \kappa_1, , \kappa_2, \omega_1, \ldots, \omega_{N-1})'$[2] of the model we use a flat prior which for the parameters $(\phi_1, \ldots, \phi_N) \in \boldsymbol{\varphi}$ is truncated for values greater than 1 and lower than -1. This is to guarantee stationarity of the filter following Harvey (2013). In addition, for the parameters $(\nu_3, \ldots, \nu_N) \in \boldsymbol{\varphi}$ we propose a normal distribution prior, with density function

$$\pi(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \tag{5}$$

[2]In the case of the combination we estimate $N-1$ weights in the interval $(0,1)$ while $\omega_N = \sum_{i=1}^{N-1} \omega_i$.

where the hyper-parameters are $\mu = 2$ and $\sigma^2 = 1$, while for $(\kappa_3, \ldots, \kappa_N) \in \boldsymbol{\varphi}$ we propose a Normal distribution prior with density function

$$\pi(x) = \frac{1}{b\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-a}{b}\right)^2\right) \tag{6}$$

with hyper-parameters $a = -1$ and $b = 2.5$, which are the intercepts and scales for the dynamic parameters driving the dynamics of the combination weights.

3.2. Posterior Approximation

Let $\mathbf{y} = (y_1, \ldots, y_T)'$ be the collection of observations and $\boldsymbol{\varphi}$ is the parameter vector associated to the full model, the likelihood associated to the model is

$$L(\boldsymbol{y}|\boldsymbol{\varphi}) = \prod_{t=1}^{T} b_{\alpha_{t|t},\beta_{t|t}} \left(\sum_{i=1}^{M} \omega_i F_{it}(y_t)\right) \sum_{i=1}^{M} \omega_i f_{it}(y_t) \tag{7}$$

Let $\pi(\boldsymbol{\varphi})$ be the joint prior distribution defined in the previous section, the joint posterior distribution $\pi(\boldsymbol{\varphi}, \boldsymbol{y}) \propto f(\boldsymbol{y}|\boldsymbol{\varphi})\pi(\boldsymbol{\varphi})$ is not tractable, however we can apply efficient Monte Carlo Markov Chain (MCMC) approximation methods. The estimation approach proposed is the adaptive random scan adaptive Metropolis-within-Gibbs discussed in Łatuszyński et al. (2013). Let $(\Psi, \mathcal{B}(\Psi))$ be a d-dimensional space, such that $\Psi = \Psi_1 \times \ldots \times \Psi_d$ and write $\boldsymbol{\varphi}^{(n)} \in \Psi$ as $\boldsymbol{\varphi}^{(n)} = (\boldsymbol{\varphi}_1^{(n)}, \ldots, \boldsymbol{\varphi}_d^{(n)})$. Let $\boldsymbol{\varphi}_{-i}^{(n)} = (\boldsymbol{\varphi}_1^{(n)}, \ldots, \boldsymbol{\varphi}_{i-1}^{(n)}, \boldsymbol{\varphi}_{i+1}^{(n)}, \ldots, \boldsymbol{\varphi}_d^{(n)})$ be the parameter vector obtained dropping the i-th component from $\boldsymbol{\varphi}^{(n)}$ with values in $\Psi_{-i} = \Psi_1 \times \ldots \times \Psi_{i-1} \times \Psi_{i+1} \times \ldots \times \Psi_d$ and with $\pi(\cdot|\boldsymbol{\varphi}_{-i})$ the conditional distribution of $\boldsymbol{\varphi}_i$ given $\boldsymbol{\varphi}_{-i}$. The adaptive random scan adaptive Metropolis within Gibbs, draws $\boldsymbol{\varphi}^{(n)}$ given $\boldsymbol{\varphi}^{(n-1)}$ performing a Metropolis Hastings step, by first choosing coordinates at random according to some selection probabilities $\alpha = (\alpha_1, \ldots, \alpha_d)$. Therefore, given $\boldsymbol{\varphi}_{-i}^{(n-1)}$ the i-th coordinate is selected with probability α_i and $\boldsymbol{\varphi}_i^{(n-1)}$ is updated by drawing $\boldsymbol{\varphi}^*$ from the proposal distribution $Q_{\boldsymbol{\varphi}_{-i}^{(n-1)}, v_i^{(n-1)}}(\boldsymbol{\varphi}_i^{(n-1)}, \cdot)$. The proposal is then chosen adaptively from the distribution family $Q_{\Psi_{-1}, v}$ by setting the parameters to $\gamma_i^{(n)}$. The sampler iterates the following steps

1. Set $\alpha^{(n)} = R_n(\alpha^{(0)}, \ldots, \alpha^{(n-1)}, \boldsymbol{\varphi}^{(n-1)}, \ldots, \boldsymbol{\varphi}^{(0)}, v^{(n-1)}, \ldots, v^{(0)}) \in A$.
2. Set $v^{(n)} = R'_n(\alpha^{(0)}, \ldots, \alpha^{(n-1)}, \boldsymbol{\varphi}^{(n-1)}, \ldots, \boldsymbol{\psi}^{(0)}, v^{(n-1)}, \ldots, v^{(0)}) \in G_1 \times \cdots \times G_n$.
3. Choose coordinate $i \in 1, \ldots, d$ according to selection probabilities α_n
4. Draw $\boldsymbol{\varphi}_i^* \sim Q_{\boldsymbol{\varphi}_{-i}^{(n-1)}, v_i^{(n-1)}}(\boldsymbol{\varphi}_i^{(n-1)}, \cdot)$.

5. With probability

$$\rho_i^{(n)} = \min\left(1, \frac{\pi(\boldsymbol{\varphi}_i^*|\boldsymbol{\varphi}_{-i}^{(n-1)}) q_{\boldsymbol{\varphi}_{-i}^{(n-1)}, v_i^{(n-1)}}(\boldsymbol{\varphi}^*|\boldsymbol{\varphi}_i^{(n-1)})}{\pi(\boldsymbol{\varphi}^{(n-1)}|\boldsymbol{\varphi}_{-i}^{(n-1)}) q_{\boldsymbol{\varphi}_{-i}^{(n-1)}, v_i^{(n-1)}}(\boldsymbol{\varphi}_i^{(n-1)}, \boldsymbol{\varphi}^*)}\right) \tag{8}$$

accept the proposal and set $X^{(n)} = (\boldsymbol{\varphi}_1^{(n-1)}, \ldots, \boldsymbol{\varphi}_{i-1}^{(n-1)}, \boldsymbol{\varphi}^*, \boldsymbol{\varphi}_{i+1}^{(n-1)}, \ldots, \boldsymbol{\varphi}_d^{(n-1)})$ otherwise reject and set $\boldsymbol{\varphi}^{(n)} = \boldsymbol{\varphi}^{(n-1)}$.

The adaptive proposal distribution is chosen following Andrieu and Thoms (2008). We assume the distribution $Q_{\boldsymbol{\varphi}_{-i}, v_i}(\boldsymbol{\varphi}_i, \cdot)$ has parameters $v_i = \{g_i, l_i, \mathbf{m}_i, S_i\}$ which update as follows:

$$\log(l_i^{(n+1)}) = \log(l_i^{(n)}) + g_i^{(n+1)}(\rho_i^{(n+1)} - \bar{\rho}) \tag{9}$$

$$\mathbf{m}_i^{(n+1)} = \mathbf{m}_i^{(n)} + g_i^{(n+1)}(\boldsymbol{\psi}_i^{(n+1)} - \mathbf{m}_i^{(n)}) \tag{10}$$

$$S_i^{(n+1)} = S_i^{(n)} + g_i^{(n+1)}((\boldsymbol{\psi}_i^{(n+1)} - \mathbf{m}_i^{(n)})(\boldsymbol{\psi}_i^{(n+1)} - \mathbf{m}_i^{(n)})' - S_i^{(n)}), \tag{11}$$

where $g_i^{(n+1)} = (n+1)^{-a_i}$ is the adaptive scale for the i-th parameter, $\bar{\rho}$ is the expected acceptance probability. Following the suggestions in Roberts et al. (1997) and Andrieu and Thoms (2008) we choose $\bar{\rho} = 0.42$.

3.3. Model Selection and Diagnostics

Another advantage of the score-driven structure is that we can use the fitted scores of the dynamic parameters to test for the presence of dynamics. As discussed by Harvey (2013), residual correlation in fitted scores under the null of no dynamics can be used to detect the presence of dynamics in time-varying parameters of score-driven models.

Calvori et al. (2017) sets up a general testing framework based on **?** which exploits residual correlation in fitted scores, while Harvey and Thiele (2016) and Palumbo (2021) sets up formal Lagrange Multiplier (LM) tests in score-driven models for testing the dynamics of correlation and tail index parameters. In the context of mixture models, **?** sets out tests for Markov switching models while **?** finds the Lagrange multiplier (LM) test statistic to have the best size and power properties. Harvey and Palumbo (2022) used the LM approach for detecting dynamics in the weights of DAMM model.

In this paper we extend to DCS score residuals the testing procedure developed by Zellner (1975) for linear regression residuals in Bayesian estimation. When a static mixture and calibration has been fitted, in the DCS framework LM tests can be approximated with the portmanteau tests and they enable the researcher to separate out transition dynamics from location and/or scale dynamics

Sim	CRPS (i)	CRPS (ii)	CRPS (iii)	CRPS (iv)	LS (i)	LS (ii)	LS (iii)	LS (iv)	W (i)	W (ii)	W (iii)	W (iv)
CalibrComb	1.4814	1.7063	1.5156	1.4788	2.4455	2.6088	2.4411	2.4446	0.6898	0.8367	0.7438	0.6837
Comb	2.0844	2.4247	3.1413	2.2847	2.8452	2.9867	3.0931	2.8743	2.6738	2.8029	4.0000	3.0210
Model 1	3.3666	3.4695	4.0146	3.5257	4.0418	4.2668	4.8957	4.2518	4.0020	4.0050	4.0000	4.0000
Model 2	3.3646	3.4302	4.0148	3.5275	4.0390	3.5025	4.8956	4.2540	3.9980	3.9989	4.0000	4.0000

Table 1: Average values for the CRPS, LS and W metrics in all the four scenarios for each of the models across all the simulations. Here LS is reported as the negative Logarithmic Score.

and concentration. The pattern of the correlograms may be informative as to possible models and so this initial step has the potential for playing an important role in model specification for the dynamic weights of the combination and the parameters of the beta for the calibration.

When the test is against dynamics in the $i-th$ location only, and $\mu_j, j \neq i$ is fixed, the LM statistic is equivalent to the Q-statistic formed from sample autocorrelations, $r_i(\tau) = c_i(\tau)/c_i(0)$, where $c_i(\tau) = \sum_{t=\tau+1}^{T} u_{i,t}u_{i,t-\tau}/T$, and $u_{i,t}$ are the fitted scores with respect to the time varying parameters for $i = 1, ...K$ and for lags $\tau = 1, ..., P,$, that is

$$Q_i(P) = T\sum_{\tau=1}^{P} r_i^2(\tau), \quad i = 1, 2, .. \tag{12}$$

As noted by Harvey and Thiele (2016), estimation of fixed parameters makes no difference to the distribution of $Q_i(P)$, which is asymptotically distributed as χ^2_P under the null hypothesis.

4. Simulation Study

In this section we investigate the performance of our combination with dynamic calibration model in finite samples. In order to so we present the following four simulation studies.

(i) In the first study the data generating process is a Gaussian distribution with time varying location between $\mu_0 = -4$ to $\mu_T = 4$, where

$$y_t = \mu_t + \varepsilon_t, \qquad \varepsilon_t \sim N(0, 2)$$
$$\mu_t = \mu_{t-1} + 8/T, \qquad t = 1, \ldots, T.$$

(ii) in the second study the data generating process is a Gaussian distribution with time varying

	(i)		(ii)		(iii)		(iv)	
	Avg Par	95% Est Int	Avg Par	95% Est Int	Avg Par	95% Est Int	Avg Par	95% Est Int
ν_1	0.086 (0.030)	(0.070, 0.100)	0.199 (0.058)	(0.169, 0.228)	0.307 (0.050)	(0.281, 0.332)	0.243 (0.049)	(0.219, 0.267)
ϕ_1	0.995 (0.007)	(0.991, 0.999)	0.964 (0.022)	(0.952, 0.975)	0.992 (0.003)	(0.990, 0.993)	0.998 (0.003)	(0.997, 0.999)
κ_1	0.019 (0.002)	(0.018, 0.019)	0.018 (0.001)	(0.017, 0.018)	0.049 (0.004)	(0.046, 0.050)	0.023 (0.001)	(0.022, 0.023)
ν_2	0.194 (0.041)	(0.173, 0.214)	0.668 (0.102)	(0.616, 0.718)	1.046 (0.093)	(0.999, 1.092)	0.271 (0.046)	(0.247, 0.293)
ϕ_2	0.999 (0.002)	(0.997, 1.000)	0.997 (0.004)	(0.994, 0.998)	0.990 (0.003)	(0.988, 0.991)	0.997 (0.007)	(0.993, 1.000)
κ_2	0.025 (0.001)	(0.024, 0.025)	0.030 (0.001)	(0.029, 0.031)	0.040 (0.003)	(0.038, 0.042)	0.020 (0.001)	(0.019, 0.020)
ω_1	0.511 (0.034)	(0.507, 0.515)	0.302 (0.092)	(0.292, 0.311)	0.540 (0.174)	(0.518, 0.561)	0.467 (0.060)	(0.459, 0.474)

Table 2: Average of the plug-in estimators from each of the simulations in each of the scenarios. In parenthesis we have the standard deviation of the estimated parameters and its 95% estimation interval across the simulations.

location between $\mu_0 = -4$ to $\mu_T = 4$, where

$$y_t = \mu_t + \eta_t, \qquad \eta_t \sim \begin{cases} N(0,2), & t < T/2 \\ T_4(0,2), & t \geq T/2 \end{cases}$$

$$\mu_t = \mu_{t-1} + 8/T, \qquad t = 1, \ldots, T.$$

(iii) in the second study the data generating process is a Gaussian distribution with time varying location between $\mu_0 = -4$ to $\mu_T = 4$, where

$$y_t = \mu_t + \varepsilon_t, \qquad \varepsilon_t \sim N(0,2)$$

$$\mu_t = \begin{cases} -4, & t \in (-\inf, T/4] \cup (T/2, T3/4] \\ 4, & t \in (T/4, T/2] \cup (T3/4, T]. \end{cases}$$

(iv) in the second study the data generating process is a Gaussian distribution with time varying location between $\mu_0 = 0$ to $\mu_T = 2\pi$, where

$$y_t = 4\sin(\mu_t) + \varepsilon_t, \qquad \varepsilon_t \sim N(0,2)$$

$$\mu_t = \mu_{t-1} + 2\pi/T, \qquad t = 1, \ldots, T.$$

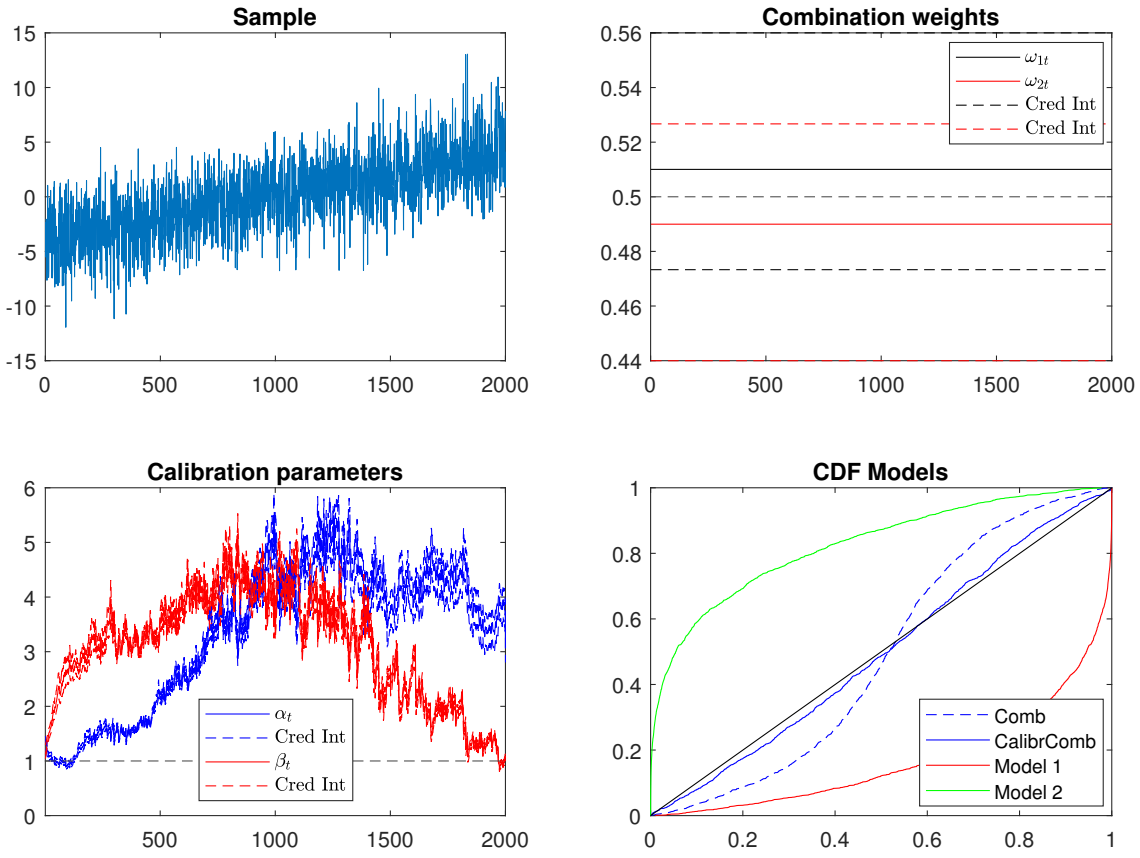

Figure 2: The figure displays a representative series generated in the simulation setting (i) (Top Left), the resulting estimated average combination weights with credible intervals (Top Right), the average dynamics of the time varying parameters with their credible intervals (Bottom Left) and the resulting average estimated PITs for all the models (Bottom Right).

The samples generated are of size $T = 1,000$ while the two models combined are a $N(-4, 2)$ and $N(4, 2)$. The number of simulations in all the studies is $M = 1,000$. These heterogeneous examples capture time-varying location, structural breaks and smooth transition towards heavy tails.

In all the settings we evaluate in-sample the goodness of fit measure of the combined and calibrated predictive densities with the average Logarithmic Score (LS)[3], the average continuous ranked probability score (CRPS) and the average Wasserstein Distance (W) which, given two cdf functions at t one obtained from the true data generating process, $F_t(x)$, and one for model k,

[3]Here for consistency across the measures we have reported the negative of the logarithmic score. Therefore a smaller values would imply a better fit of the data generating process density.

	CRPS	Logscore
CalibrComb	0.6762	1.1653
Comb	0.6764	1.2216
GARCH	0.6836	1.2522
GARCH-t	0.6763	1.2002

Table 3: Average values for the CRPS, LS across all the iterations for each of the models. Here LS is reported as the negative Logarithmic Score.

	η_1	ϕ_1	κ_1	η_2	ϕ_2	κ_2	ω_1
Avg Par	0.424 (0.003)	0.182 (0.012)	0.089 (0.002)	0.415 (0.003)	0.357 (0.033)	0.022 (0.002)	0.506 (0.029)
95% Cred Int	(0.422, 0.425)	(0.175, 0.187)	(0.088, 0.089)	(0.412, 0.416)	(0.341, 0.373)	(0.021, 0.023)	(0.505, 0.505)

Table 4: The table reports the average plug-ins estimator across the iterations, their standard deviation in parenthesis and their 95% credible intervals.

$F_{tk}(x)$, can be obtained as[4]

$$W_k = \frac{1}{T}\sum_{t=1}^{T}\int_{\mathbb{R}} |F_t(x) - F_{tk}(x)|\, dx$$

In Table 1 and Table 2 is possible to see the results from the simulations. In all the cases the calibration provides a significant improvement over the static combination across all the metrics. As it is possible to see in Figure 2 as well as and in Figure 9, Figure 10 and Figure 11 in the Appendix, in all the cases the average PIT of the combined and calibrated model across the simulations is the closest to the 45 degree line, which represents the PIT plot for the unknown true/ideal model. Given that in all the cases the shift between the two distribution has been time varying, the inclusion of a dynamic calibration giver a strong advantage to the model. Looking at the filtered dynamics of the two parameters $\alpha_{t|t-1}$ and $\beta_{t|t-1}$ they tend to compensate for the shift in the location giving different weights to either upper or lower quantiles of the combined distribution. Higher $\alpha_{t|t-1}$ implies that the calibration function weights more the left tail quantiles making them heavier, while $\beta_{t|t-1}$ affect in the same way the right tail. This feature is quite clear in Figure 2 where the combined distribution almost equally weights the predictive densities from the two models, resulting in a location close to 0. Is also worth noticing that almost in all the cases the static combination tends to equally weight the predictive distributions from the models, a part from (ii) as can be seen in Figure 9. This suggests that the combination part plays a very marginal role in obtaining the resulting combines and calibrated distribution. It is possible to show that similar results can be obtained also from statically fix the combination weighting parameters across the models[5].

[4]Computationally the integral in the measures is discretised and the cdfs are computed at every t over a finite grid.
[5]Further results upon request.

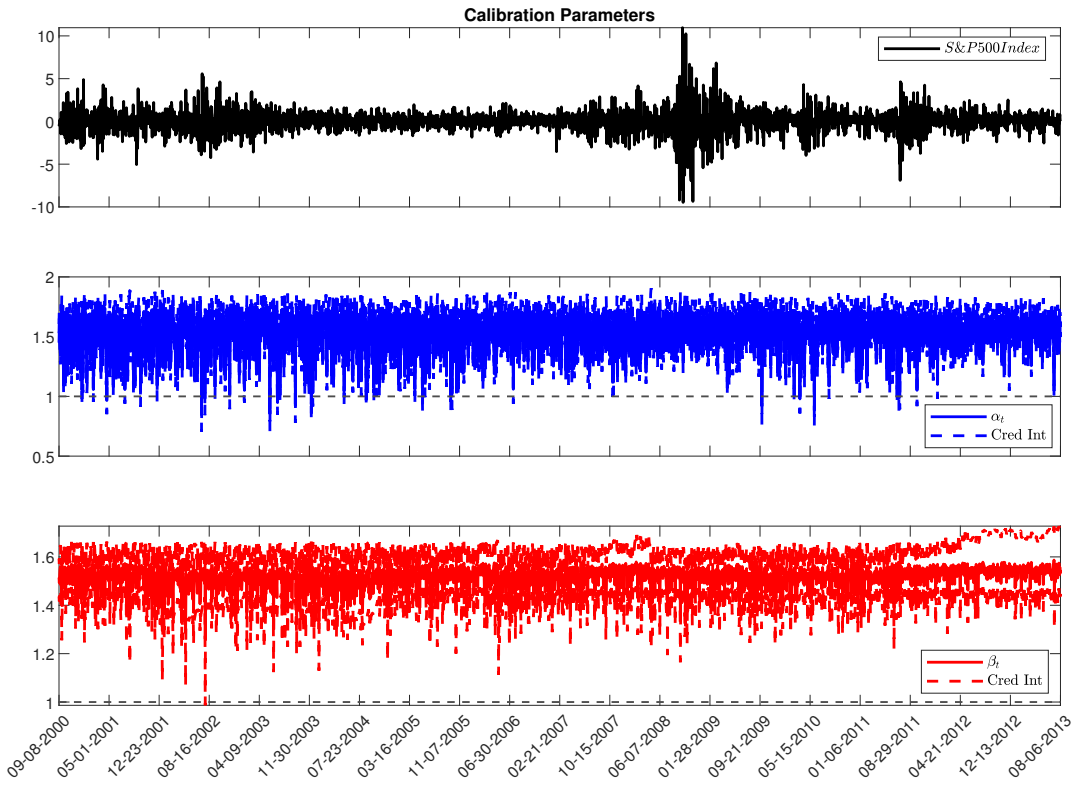

Figure 3: Data series (Top), estimated time varying $\alpha_{t|t-1}$ with credible intervals (Mid), estimated time varying $\beta_{t|t-1}$ with credible intervals (Bottom).

5. Empirical application

We investigate the relative predictability accuracy of the novel approach on a well-studied datasets from finance (Amisano and Geweke (2010), Geweke and Amisano (2011), Kapetanios et al. (2015) already used in Bassetti et al. (2018) and a novel dataset from meteorology.We evaluate the predictive densities using the average Logarithmic Score (LS) and the continuous ranked probability score (CRPS) conditional to the information available up to time $t-1$.

5.1. Forecasting Financial Returns

The first application considers S&P 500 daily percent log returns data from the 3rd of January 1972 to the 6th of August 2013. We used the forecasts from Bassetti et al. (2018) which are from a Normal GARCH(1,1) model and a GARCH-t(1,1) obtained through maximum likelihood (ML) using rolling samples of 1250 trading days (about five years) and produce one day ahead density forecasts for the period from the 8th of September 2000 to the 6th of August 2013. The predictive densities are formed by substituting the ML estimates for the unknown parameters. Then the novel combination-calibration model is applied on the resulting predictive cdfs and pdfs.

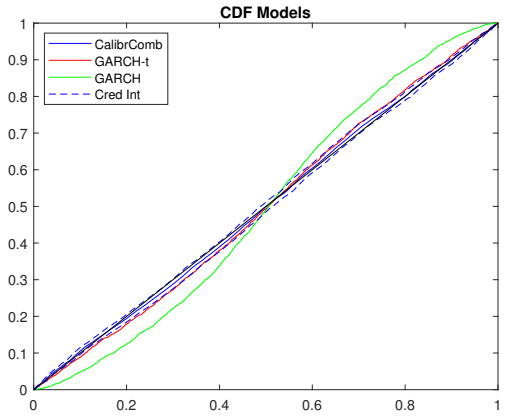

Figure 4: The resulting average estimated PITs for all the models including the credible intervals.

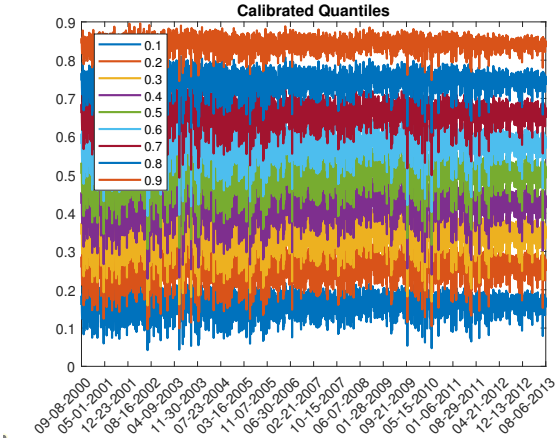

Figure 5: Estimated quantile weights on the beta calibrating function.

The parameter estimation results and model performance are summarised in Table 3 and Table 4. As can be seen the dynamics of the two calibration parameters is not too persistent while the weights are estimated around 0.5 having very little effect at improving the performance of the combined forecasts. However, as can be seen in Figure 4, the PIT of the combined and calibrated model is the closer to the 45 degree line, which also in this case represents the PIT plot for the unknown true/ideal model. This 45 degree line always belongs to the confidence interval of the Combined and Calibrated model, while is not the case for the PIT of the GARCH and occasionally not even for the GARCH-t model. This results is confirmed from the LS and CRPS results in Table 3. The t-model provides better scores than the Gaussian one. The accuracy of the Gaussian GARCH model is very low during the selected period, in particular due to the extreme events provided by the uncertainty in the 2008 period.

The impact of the dynamic calibration can be seen in Figure 5 and Figure 3 where we can see that both α and $\beta_{t|t-1}$ tend to stay away from 1, the perfect calibrated values, particularly in periods of highest volatility.

5.2. Forecasting Wind Speed

Here we present the results for the Cairns wind wind speed half-hourly dataset from the Australian Bureau of Meteorology. The dataset includes data of wind direction and velocity over a period of 2 month (60 days, for a total of 2880 observations), recorded at a wind farm in Cairns, Australia. Some of the wind speed observations were missing and, according to the Bureau, this is because there was no wind at these hours. These missing observations accounted for roughly 7% of the data.

We have produced forecasts on a rolling window basis using half of the dataset for estimation

	CRPS	Logscore
CalibrComb	2.7438	0.7960
Comb	3.8626	1.7182
ARMA	3.2791	1.0660
ARMA-GARCH	4.2753	2.3220
ARFIMA	3.0451	0.8379
GAR-GARCH	7.0129	2.9701

Table 5: Average values for the CRPS, LS across all the iterations for each of the models. Here LS is reported as the negative Logarithmic Score.

	η_1	ϕ_1	κ_1	η_2	ϕ_2	κ_2	ω_1	ω_2	ω_3
Est Par	1.285 (1.728)	1.000 (0.000)	0.026 (0.019)	2.250 (0.961)	0.986 (0.033)	0.014 (0.008)	0.065 (1.736)	0.694 (0.289)	0.034 (0.606)
95% Cred Int	(1.214, 1.356)	(0.999, 0.999)	(0.025, 0.026)	(2.210, 2.289)	(0.984, 0.987)	(0.013, 0.013)	(0.061, 0.069)	(0.694, 0.692)	(0.034, 0.034)

Table 6: The table reports the average plug-ins estimator across the iterations, their standard deviation in parenthesis and their 95% credible intervals.

(1440 observations) producing a dataset of 1440 forecasts. The model tested were an ARMA, an ARMA-GARCH, an ARFIMA and the GAR-GARCH of Caporin and Preś (2012)

The parameter estimation results and model performance can be seen in Table 5 and Table 6. As can be seen the dynamics of the two calibration parameters this time is much more persistent. The estimated combination weights tend to down-weight the ARMA and ARFIMA model in favour of the ARMA-GARCH and GAR-GARCH model. However the static combination improves only little on the overall density forecasts. As can be seen in Figure 7 the PIT of the combined and calibrated model is also in this case the closer to the 45 degree line and it lies in the credible intervals while this time all the other model are much further away, with the exception for the GAR-GARCH. This is consistent with the LS and CRPS results in Table 6. In this case from Figure 8 we can see that the quantiles gets calibrated much more, in particular towards the end of the sample when speed falls while α rises and $\beta_{t|t-1}$ falls.

6. Conclusion

This paper extends the predictive density calibration literature and proposes to apply combination with dynamic calibration. We extend the static model in Ranjan and Gneiting (2010) to a score-driven dynamic model for calibration and combination of predictive distributions. The predictive densities from various models are first collected and combined and a beta transformation is applied. As in Ranjan and Gneiting (2010) and Bassetti et al. (2018) the beta distribution is used as a calibrating function which calibrates dynamically at every t the combined predictive density. The weights of the combination are statically estimated while the time varying calibration is obtaining giving a score-driven dynamics to the parameters of the beta calibrating function. We provide a Bayesian inference procedure and an efficient Monte Carlo Markov Chain sampler for

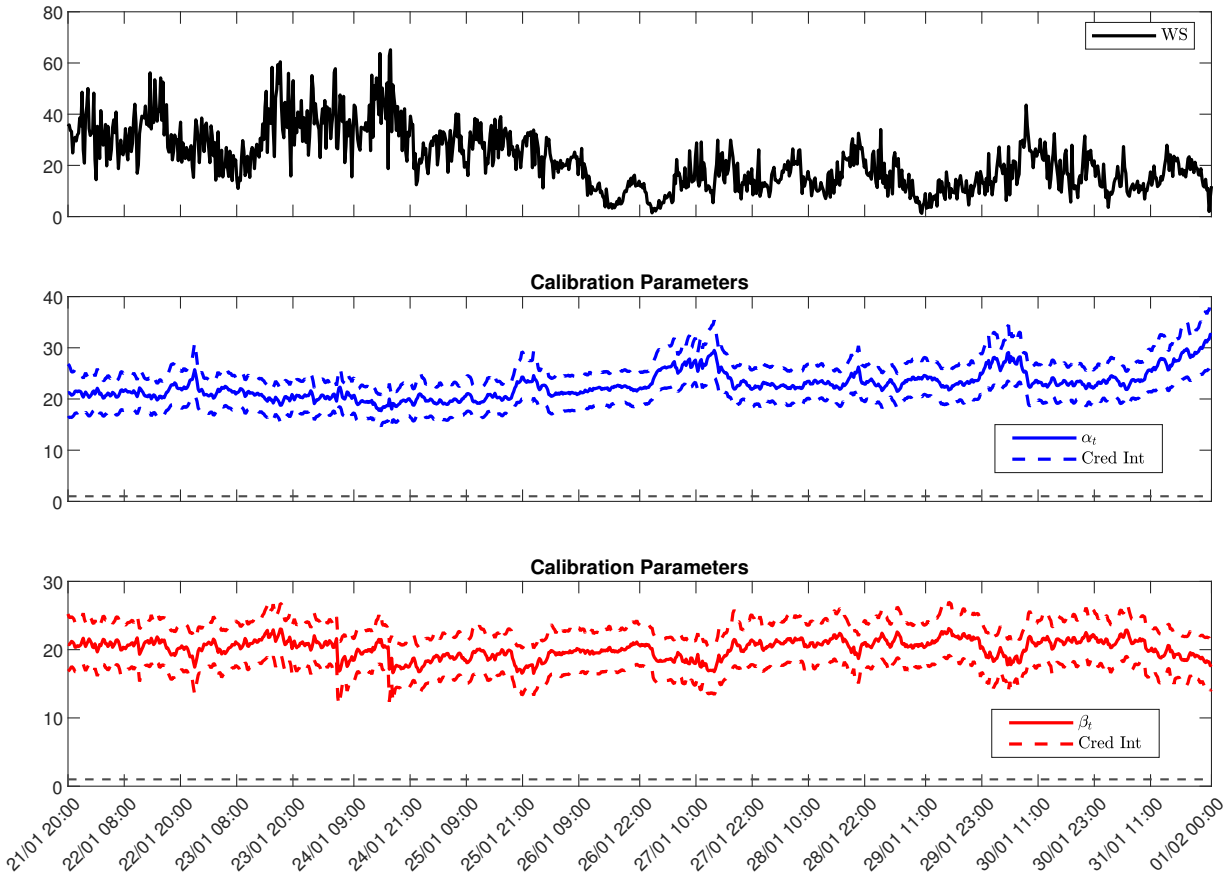

Figure 6: Data series (Top), estimated time varying $\alpha_{t|t-1}$ with credible intervals (Mid), estimated time varying $\beta_{t|t-1}$ with credible intervals (Bottom).

posterior approximation.

Through simulations, we show that the model is very flexible and can handle different shapes, instability and model uncertainty. The simulations provides examples of data generating processes with time varying location, structural breaks and smooth transition towards heavy tails. In all the cases the model calibrates correctly the ensemble of predictive densities showing that the calibration step is almost independent from the initial combination. This suggests that the model flexibility could allow to rescale effectively the quantiles of the combined predictive density even under a fixed *naive* selection of the weights.

The empirical performance of the proposed model is then assessed on two different datasets. The first is a dataset of GARCH predictive densities based on S&P 500 Index returns which includes the 2008 crisis, and the second is of half-hourly wind speed forecasts produced by various statistical state-of-the-art models on a dataset from a wind farm in Cairns Australia. In both the cases

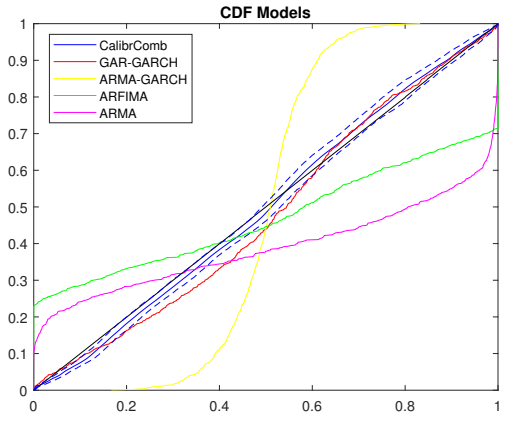

Figure 7: The resulting average estimated PITs for all the models including the credible intervals.

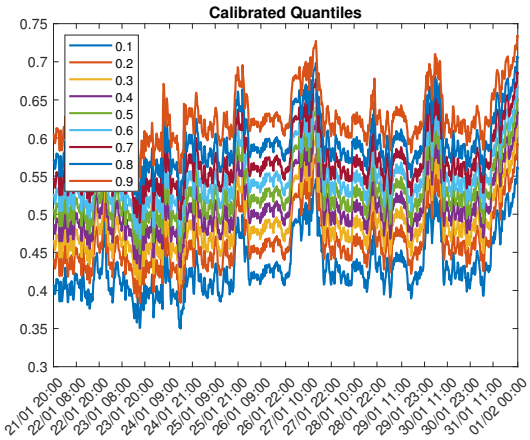

Figure 8: Estimated quantile weights on the beta calibrating function.

it is shown that the combined dynamic calibration and combination approach provides superior density forecasts, both in terms of Log score and CRPS criteria. The application shows that as the number of models increase the static combination down-weights the prediction from the worse performing models, while the dynamic calibration adapts the quantiles of the combined density to fit the observed data structure. Ultimately in both the applications the credible intervals of the resulting combined and calibrated PITs from the model contain the 45 degree line of correctly calibrated uniform quantiles.

The introduced model is a powerful tool in many fields that use predictive density to make predictive risk assessments, such as finance or environmental studies. Can be essential to asset manager to predict portfolios risk-metrics and expected losses in presence of total uncertainty over the market condition. It can be also very useful for climate predictions where is essential to reconcile physical predictive models with the statistical structure of the climate data observed, in particular in presence of increasing uncertainty given by climate change.

References

Amisano, G. and J. Geweke (2010, April). Comparing and evaluating bayesian predictive distributions of asset returns. *International Journal of Forecasting 26*(2), 216–230.

Andrieu, C. and J. Thoms (2008). A tutorial on adaptive MCMC. *Statistics and computing 18*(4), 343–373.

Barnard, G. A. (1963). New methods of quality control. *Journal of the Royal Statistical Society, Series A 126*, 255–259.

Bassetti, F., R. Casarin, and F. Ravazzolo (2018). Bayesian nonparametric calibration and combination of predictive distributions. *Journal of the American Statistical Association 113*(522), 675–685.

Bates, J. and C. Granger (1969). The combination of forecasts. *Operations Research Quarterly 20*(4), 451–468.

Billio, M., R. Casarin, F. Ravazzolo, and H. K. van Dijk (2013). Time-varying combinations of predictive densities using nonlinear filtering. *Journal of Econometrics 177*, 213–232.

Calvori, F., D. Creal, S. Koopman, and A. Lucas (2017). Testing for parameter instability in competing modeling frameworks. *Journal of Financial Econometrics* (15), 223–246.

Caporin, M. and J. Preś (2012). Modelling and forecasting wind speed intensity for weather risk management. *Computational Statistics and Data Analysis 56*, 3459–3476.

Creal, D., S. J. Koopman, and A. Lucas (2013). Generalized autoregressive score models with applications. *Journal of Applied Econometrics 28*, 777–795.

Geweke, J. and G. Amisano (2011). Optimal prediction pools. *Journal of Econometrics 164*(1), 130 – 141.

Gneiting, T. and A. E. Raftery (2007, March). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association 102*, 359–378.

Gneiting, T. and R. Ranjan (2013). Combining predictive distributions. *Electronic Journal of Statistics 7*, 1747–1782.

Gorgi, P. and S. Koopman (2021). Beta observation-driven models with exogenous regressors: A joint analysis of realized correlation and leverage effects. *Journal of Econometrics*.

Granger, C. W. J. and R. Ramanathan (1984). Improved Methods of Combining Forecasts. *Journal of Forecasting 3*, 197–204.

Hall, S. G. and J. Mitchell (2007). Combining density forecasts. *International Journal of Forecasting 23*(1), 1–13.

Harvey, A. (2013). *Dynamic models for volatility and heavy tails: with applications to financial and economic time series*.

Harvey, A. and A. Luati (2014). Filtering with heavy tails. *Journal of the American Statistical Association 109*(507), 1112–1122.

Harvey, A. and D. Palumbo (2022). Score-driven models for realized volatility. *Journal of Econometrics*.

Harvey, A. and S. Thiele (2016). Testing against changing correlation. *Journal of Empirical Finance 38*, 575–589.

Harvey, A., T. M. Trimbur, and H. K. Van Dijk (2007). Trends and cycles in economic time series: A bayesian approach. *Journal of Econometrics 140*(2), 618–649.

Hoogerheide, L., R. Kleijn, R. Ravazzolo, H. K. van Dijk, and M. Verbeek (2010). Forecast Accuracy and Economic Gains from Bayesian Model Averaging using Time Varying Weights. *Journal of Forecasting 29*(1-2), 251–269.

Kapetanios, G., J. Mitchell, S. Price, and N. Fawcett (2015). Generalised density forecast combinations. *Journal of Econometrics 188*, 150–165.

Łatuszyński, K., G. O. Roberts, J. S. Rosenthal, et al. (2013). Adaptive Gibbs samplers and related MCMC methods. *The Annals of Applied Probability 23*(1), 66–98.

McAlinn, K. and M. West (2018). Dynamic bayesian predictive synthesis in time series forecasting. *Journal of Econometrics forthcoming*.

Palumbo, D. (2021). Testing and modelling time series with time varying tails. *Cambridge Working Paper Series*.

Raftery, A., M. Karny, and P. Ettler (2010). Online prediction under model uncertainty via dynamic model averaging: Application to a cold rolling mill. *Technometrics 52*, 52–66.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski (2005). Using Bayesian Model Averaging to Calibrate Forecast Ensembles. *Monthly Weather Review 133*, 1155–1174.

Raftery, A. E., D. Madigan, and J. A. Hoeting (1997, March). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association 92*(437), 179–91.

Ranjan, R. and T. Gneiting (2010). Combining probability forecasts. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72*(1), 71–91.

Roberts, G. O., A. Gelman, and W. R. Gilks (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability 7*(1), 110–120.

Roberts, H. V. (1965). Probabilistic prediction. *Journal of American Statistical Association 60*, 50–62.

Terui, N. and H. K. van Dijk (2002). Combined forecasts from linear and nonlinear time series models. *International Journal of Forecasting 18*, 421–438.

Zellner, A. (1975). Bayesian analysis of regression error terms. *Journal of the American Statistical Association 70*(349), 138–144.

7. Appendix

7.1. Derivation of the Conditional Score for the Calibration Parameters

To obtain the result in Equation (4) we start from stating that the log likelihood function of a single observation y_t is

$$\begin{aligned}\ln f\left(y|\boldsymbol{\theta}\right) &= \log b_{\alpha,\beta}\left(H\left(y|\boldsymbol{\omega}\right)\right) + \log h\left(y|\boldsymbol{\omega}\right) \\ &= \left(\alpha-1\right)\ln\left(H\left(y|\boldsymbol{\omega},\right)\right) + \left(\beta-1\right)\ln\left(1-H\left(y|\boldsymbol{\omega}\right)\right) - \log B\left(\alpha,\beta\right) + \log h\left(y|\boldsymbol{\omega}\right)\end{aligned}$$

where $B\left(\alpha,\beta\right)$ is the Euler's beta function. Under our assumptions on the link function

$$\nabla_{\xi}\mathbf{g}(\boldsymbol{\xi}) = \text{diagrv}(\alpha,\beta) \tag{13}$$

where diagrv denotes the diagonal matrix with on the main diagonal elements α and β. Moreover,

$$\nabla_{\theta}\log f(y_t|\boldsymbol{\theta},\boldsymbol{\omega}) = \begin{pmatrix} \ln\left(H\left(y_t|\boldsymbol{\omega}\right)\right) - \psi\left(\alpha\right) + \psi\left(\alpha+\beta\right) \\ \ln\left(1-H\left(y_t|\boldsymbol{\omega}\right)\right) - \psi\left(\beta\right) + \psi\left(\alpha+\beta\right) \end{pmatrix}$$

Then

$$\mathbf{u}_t = \nabla_\xi \log f(y_t|\boldsymbol{\xi}, \boldsymbol{\omega}) = \nabla_\theta \log f(y_t|\boldsymbol{\theta}, \boldsymbol{\omega}) \nabla_\xi \mathbf{g}(\boldsymbol{\xi}) \tag{14}$$

7.2. Figures

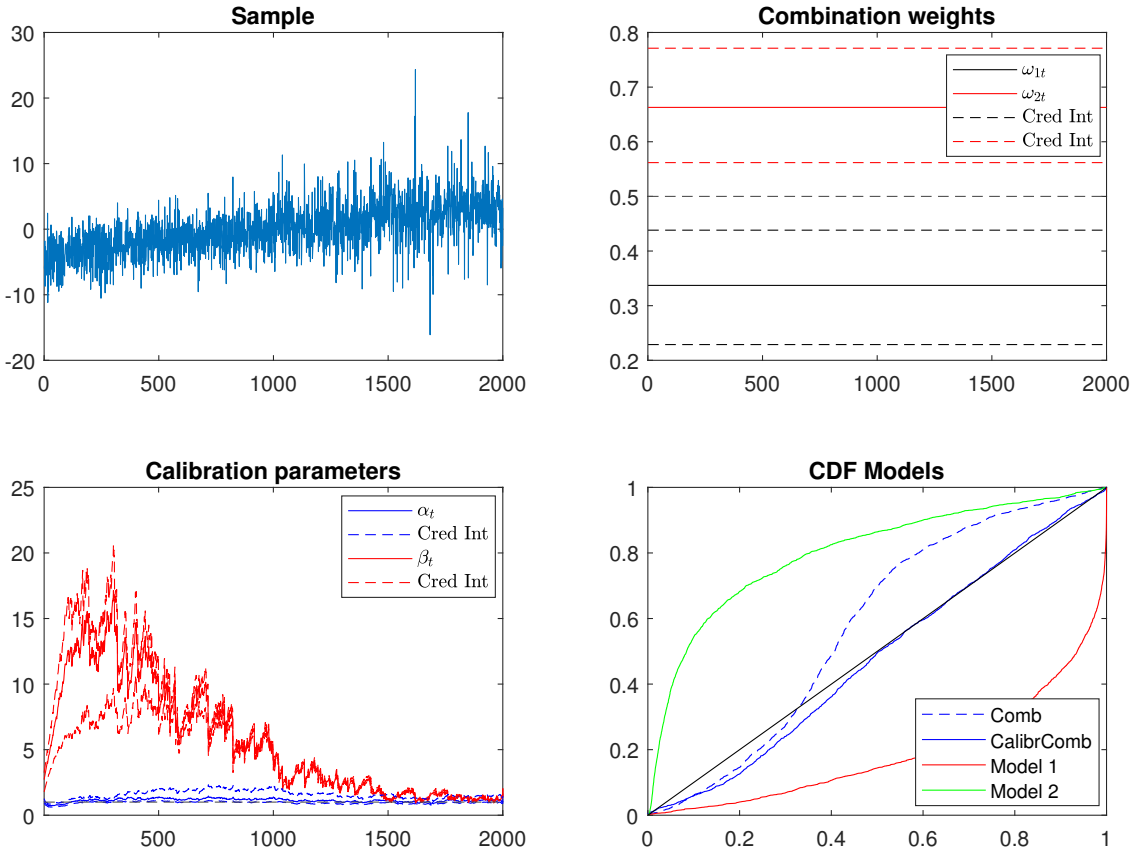

Figure 9: The figure displays a representative series generated in the simulation setting (ii) (Top Left), the resulting estimated average combination weights with credible intervals (Top Right), the average dynamics of the time varying parameters with their credible intervals (Bottom Left) and the resulting average estimated PITs for all the models (Bottom Right).

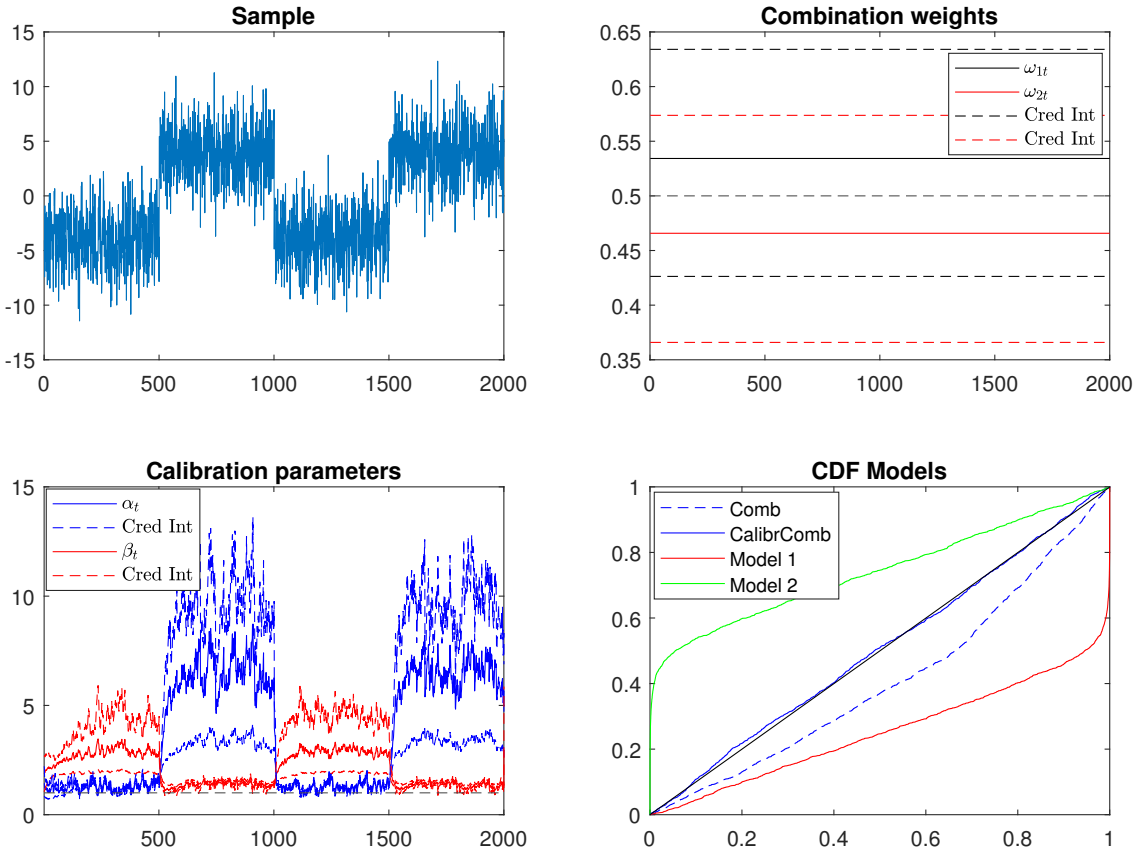

Figure 10: The figure displays a representative series generated in the simulation setting (iii) (Top Left), the resulting estimated average combination weights with credible intervals (Top Right), the average dynamics of the time varying parameters with their credible intervals (Bottom Left) and the resulting average estimated PITs for all the models (Bottom Right).

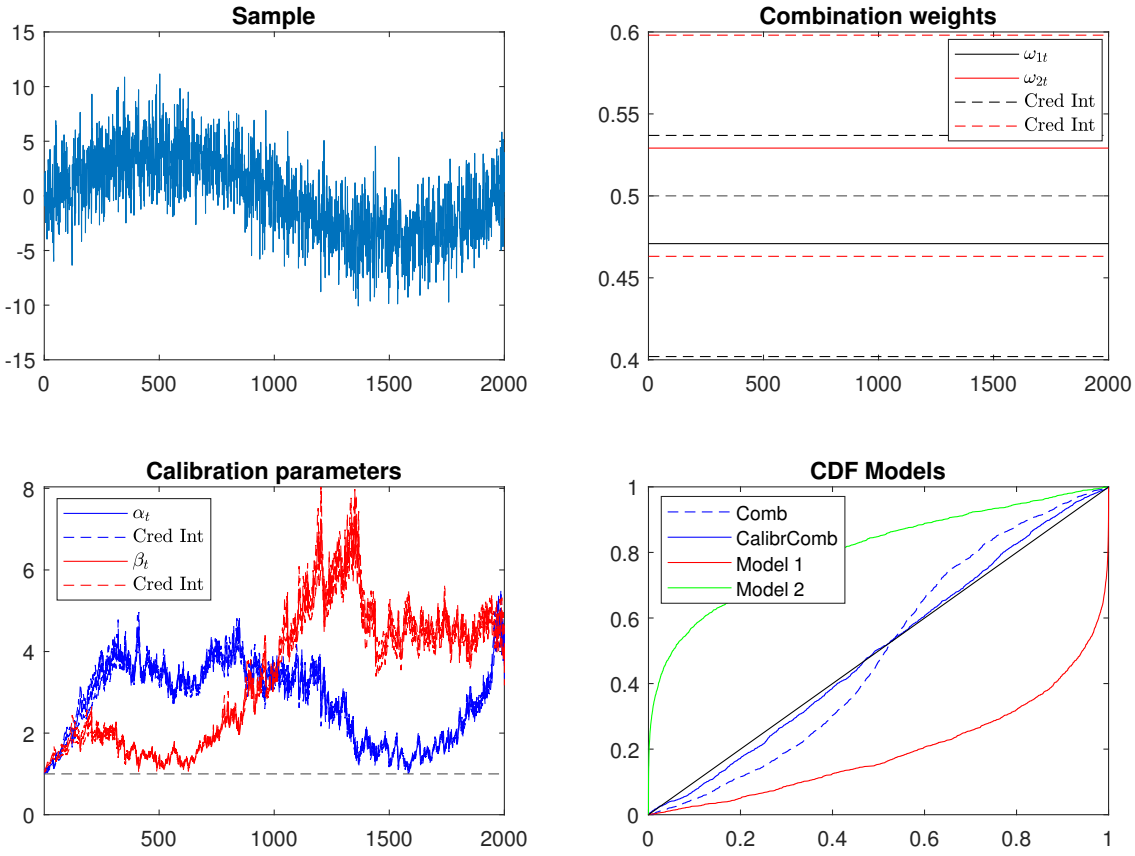

Figure 11: The figure displays a representative series generated in the simulation setting (iv) (Top Left), the resulting estimated average combination weights with credible intervals (Top Right), the average dynamics of the time varying parameters with their credible intervals (Bottom Left) and the resulting average estimated PITs for all the models (Bottom Right).

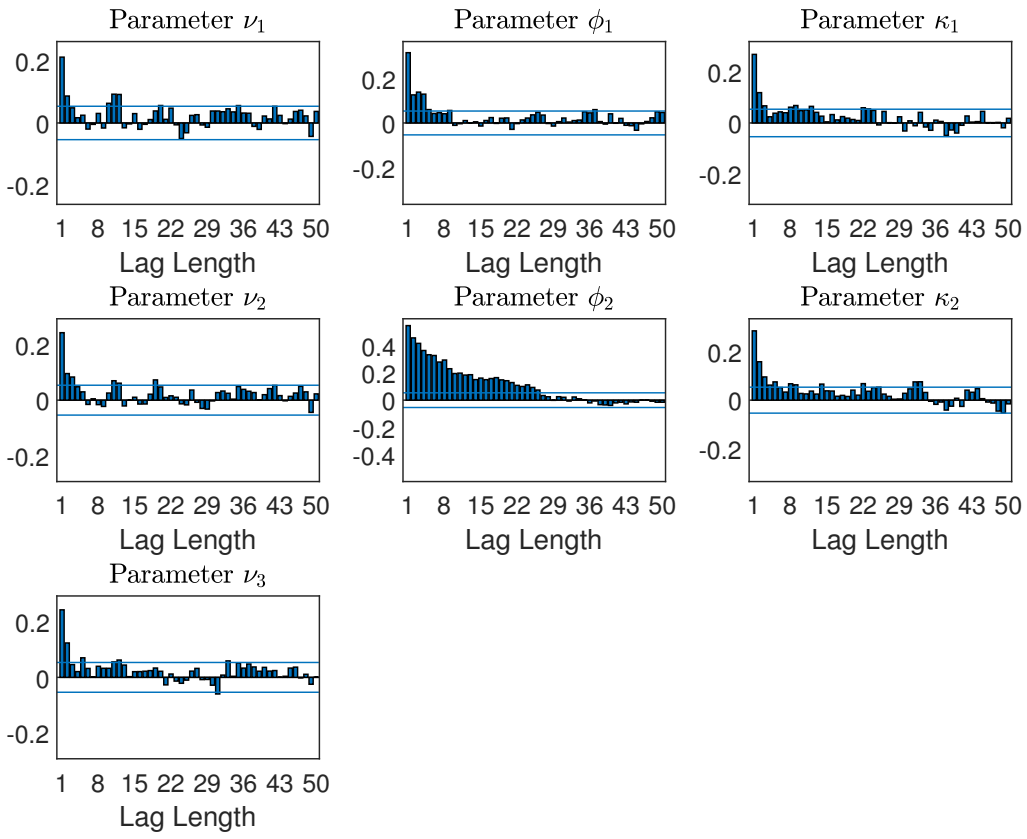

Figure 12: ACF of the MCMC iterations for each of the parameters in fitting the model on the predicted densities of the S&P 500 dataset.