We modeled long memory with just one lag!*

Luc Bauwens[a], Guillaume Chevillon[b], and Sébastien Laurent[c]

[a]LIDAM/CORE, UCLouvain, Voie du Roman Pays 34, 1348 Louvain-La-Neuve, Belgium;
[b]ESSEC Business School;
[c]Aix-Marseille University (Aix-Marseille School of Economics), CNRS, EHESS,
Aix-Marseille Graduate School of Management – IAE

April 29, 2022

Abstract

We build on two contributions that have found conditions for large dimensional networks or systems to generate long memory in their individual components, and provide a methodology for modeling and forecasting series displaying long range dependence. We model long memory properties within a vector autoregressive system of order 1 and consider Bayesian estimation or ridge regression. For these, we derive a theory-driven parametric setting that informs a prior distribution or a shrinkage target. Our proposal significantly outperforms univariate time series long memory models when forecasting a daily volatility measure for 250 U.S. company stocks, as well as seasonally adjusted monthly streamflow series recorded at 97 locations of the Columbia river basin.

Keywords: Bayesian estimation, Ridge regression, Vector autoregressive model, Forecasting.
JEL: C10, C32, C58.

*Guillaume Chevillon acknowledges financial support from Institut Europlace de Finance & Labex Louis Bachelier, ESSEC Foundation and CERESSEC, and Labex MME-DII. Sébastien Laurent acknowledges the research support of the French National Research Agency Grants ANR-17-EURE-0020 and ANR-21-CE26-0007-01 and the Excellence Initiative of Aix-Marseille University - A*MIDEX. The authors are grateful for the helpful discussions with seminar participants at Aix-Marseille, the Bank of Japan, the Joint Research Centers of the European Commission at Ispra, Lancaster, Maastricht, Macquarie, Nottingham, Oxford, Paris School of Economics, Queen Mary University, Sydney University, the University of New South Wales, as well as the QFFE 2022 conference in Marseille, the 2019 EC^2 in Oxford, the 4th Vienna workshop on Time Series, and the workshop on Financial Econometrics at Orebro University and on Long Memory in Hannover for helpful discussions. We thank in particular Matei Demetrescu, Jurgen Doornik, Domenico Giannone, Uwe Hassler, Alain Hecq, Liudas Giraitis, Roselyne Joyeux, Sebastiano Manzan, Sophocles Mavroeidis, Ulrich Müller, Bent Nielsen, Morten Nielsen, Susanne Schennach, Toshitaka Sekine, Shuping Shi, Philipp Sibbertsen, and Lorenzo Trapani for useful comments.

1 Introduction

Ever since Granger (1966) and Nelson and Plosser (1982), the question of the degree of persistence in macroeconomic and financial variables has exhibited regular puzzles. Long memory (i.e., a dependence between observations decaying hyperbolically with their distance in time, see Beran, 1992) is often encountered in economic and financial time series (at least since Smith, 1938, and Cox and Townsend, 1947; see, e.g., Baillie, 1996,, Robinson, 2003, and Hualde and Nielsen, 2022, for introductions and overviews of the field, including recent contributions thereof) and long memory models are found to provide a good empirical representation of persistence that is stronger than ARMA models but weaker than unit-root processes. The econometric literature has found that its origin can take several forms, such as aggregation (Granger, 1980, Abadir and Talmain, 2002, Leipus and Surgailis, 2003), linear modeling of a nonlinear process (e.g., Robinson and Zaffaroni, 1998, Davidson and Sibbertsen, 2005, Miller and Park, 2010, Chen, Hansen and Carrasco, 2010), structural changes (e.g., Diebold and Inoue, 2001, Gourieroux and Jasiak, 2001, Perron and Qu, 2010, Haldrup and Vera Valdés, 2017) as well as resulting from agents' self-referential learning behaviors and forward expectations (Chevillon and Mavroeidis, 2017, 2018).

Long memory has also been shown to arise in individual series that are linked within an infinite dimensional network or system. In the context of modeling space-time covariance functions as, e.g., when considering geophysical time series, Stein (2005) noticed that some forms of spatial dependence could yield, when marginalized within a Markov (in time) process, long-range dependence in time. In the context of multivariate modeling, two recent contributions have also found such a result. Chevillon, Hecq, and Laurent (2018) have proved that long memory can result from the marginalization of a large dimensional system. More specifically, they provide a parametric framework under which the variables of an n-dimensional vector autoregressive model of order 1, i.e., a VAR(1), can be individually modelled as a fractional white noise (see Granger and Joyeux, 1980) as n tends to infinity. Long memory may therefore be a feature of univariate or low dimensional models that vanishes when considering larger systems in their entirety: while the infinite dimensional system is Markovian, modeling the series individually requires infinite lags. Working with network dynamics, Schennach (2018) has found a related result of hyperbolic response of outputs to distant input shocks. These sources of long memory differ from other sources mentioned in the literature, in particular the aggregation mechanism of Granger (1980). Also, focusing on how long memory arises in individual series through interactions within a system, the results above differ from work on interactions between fractionally integrated and, possibly, cointegrated variables, as in, inter alia, Robinson and Hualde (2003), Marmol and Velasco (2004), and Johansen and Nielsen (2012), see Hualde and Nielsen (2022) for a comprehensive overview.

In this paper, we address the question of whether and how the asymptotic theoretical results of Chevillon, Hecq and Laurent (2018, CHL henceforth) and Schennach (2018, Schennach henceforth) can be put to use in empirical work. Given the large dimensional nature of their models, inference in empirical works is likely to be imprecise. Hence, rather than attempting to test the specification of a large scale model using a finite data set, we provide

an assessment of the proximity of the models to the data generating process by means of forecasting exercises. We provide in particular a set of techniques using classical and Bayesian inference which allow an empirical modeler to benefit from the asymptotic theoretical results of CHL and Schennach. The success of these modeling techniques can be interpreted as an empirical validation of the theoretical results of CHL and Schennach about long memory originating from interactions within large dimensional systems.

Given their asymptotic nature (in the cross-sectional dimension n, not in the sample size T), the results of CHL and Schennach involve systems so large that inference may be infeasible or highly imprecise in finite samples. Our approach hence relies on a parsimonious dynamic modeling using a large number of variables, so our benchmark system is the vector autoregressive model of order 1. It is well known that such a VAR(1) can be estimated equation by equation, each equation being an AR(1) (autoregressive of order 1) model augmented by the first lag of all the other variables in the system. We refer to these equations as AR(1)-X models. Our objective is to evaluate whether in large dimensions, such AR(1)-X models may constitute a viable alternative to pure long memory models like the autoregressive fractionally integrated moving average (ARFIMA) model, or to models designed to approximate well the long memory feature of time series, like the heterogeneous autoregressive (HAR) model of Corsi (2009).

By careful estimation of the AR(1)-X models, we evaluate whether the source of long memory identified by CHL and Schennach is empirically relevant and is specifically useful for forecasting variables displaying long memory. We propose two methods to estimate an AR(1)-X model augmented with long-memory prone constraints. These shrink the parameters towards values implied by a set of characteristics derived from CHL and Schennach. The first shrinkage strategy relies on an L2 penalization of the AR(1)-X model (i.e., ridge regression) and is denoted RAR(1)-X (for Ridge AR(1)-X). The second one relies on an informative prior density in a Bayesian approach, denoted BAR(1)-X (for Bayesian AR(1)-X). The degree of shrinkage, which is governed by the L2 penalty weight or by the prior variances, is chosen by cross validation between the two extremes of no or dogmatic restrictions.

We perform empirical applications in two contexts where long memory is an established feature. We focus on (i) the logarithm of a robust-to-jumps estimate of the daily integrated variance (i.e., the MedRV of Andersen, Dobrev, and Schaumburg, 2012) computed from 5-minute returns for 250 US stocks over twelve years, and (ii), the logarithm of monthly seasonally adjusted series of river streamflows at 97 locations in the Columbia river basin over ninety years. We compare the forecasting properties of the AR(1)-X, estimated by the shrinkage strategies we propose or by standard OLS, to three univariate models for short and long range dependence: the AR(1) model, the ARFIMA model, and the HAR model of Corsi (2009). Since we compare models based on different information sets, and as these models are of reduced-form type and aimed at forecasting, it is reasonable to use measures of forecast accuracy as criteria for comparisons. Hence, we compare the forecasts produced by the different models using the mean square forecast error (MSFE) and mean absolute forecast error (MAFE) loss functions, and we rely on the Model Confidence Set procedure of Hansen, Lunde and Nason (2011) to discriminate between the models.

The rest of this paper is organized as follows. Section 2 provides a theoretical framework under which a VAR(1) model can generate long memory in its components when the dimension of the system is large. The theory implies restrictions on the VAR parameters that can improve estimation and forecasting. Section 3 then explains the theory-induced constraints we suggest when estimating the equations of the VAR(1) system. It shows how to introduce them, either through an informative prior density for conducting Bayesian estimation, or by ridge estimation. Section 4 contains the empirical results. Conclusions are offered in the last section. Proofs, technical details and additional figures are collected in the appendix.

2 Long memory in a VAR(1) model

This section reviews the elements of the theoretical frameworks of Chevillon et al. (2018) and Schennach (2018) that preside over our own modeling strategy. We provide a unifying treatment and derive constraints that are germane to our estimation procedures.

Both CHL and Schennach prove that long-memory observed in a univariate time-series can be the result of the marginalization of an infinitely large VAR(1) system that satisfies some specific assumptions. For this reason, we let the observable vector $\boldsymbol{y}_{n,t}$ of dimension n satisfy, for $t \geq 1$,

$$\left(\boldsymbol{I}_n - \boldsymbol{A}_n L\right)\left(\boldsymbol{y}_{n,t} - \boldsymbol{\mu}\right) = \boldsymbol{\epsilon}_{n,t}, \tag{1}$$

where $\boldsymbol{\epsilon}_{n,t}$ is a short memory process with zero expectation and variance-covariance matrix $\boldsymbol{\Sigma}_n$.

In order to reduce expositional complexity and simplify their derivations, both CHL and Schennach restrict themselves to matrices that belong to the Toeplitz family since these require only $O(n)$ parameters. While their high-level assumptions differ, all can be subsumed in

$$\boldsymbol{A}_n = \boldsymbol{T}_n + \eta_n \boldsymbol{D}_n,$$

where η_n is a vanishing scalar sequence, and $\{\boldsymbol{T}_n\}$ and $\{\boldsymbol{D}_n\}$ denote generic sequences of Toeplitz matrices that are, respectively, symmetric and antisymmetric. Both assume that $\{\boldsymbol{D}_n\}$ plays no role asymptotically, so large-system dynamics are governed by $\boldsymbol{T}_n$, the entries of which are labelled as

$$\boldsymbol{T}_n = \begin{bmatrix} t_0^{(n)} & t_1^{(n)} & \cdots & t_{n-1}^{(n)} \\ t_1^{(n)} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & t_1^{(n)} \\ t_{(n-1)}^{(n)} & \cdots & t_1^{(n)} & t_0^{(n)} \end{bmatrix}. \tag{2}$$

According to the exposition by Schennach, the process $\boldsymbol{y}_{n,t}$ can be seen as generated by a network that lies in a space of dimension one. She also considers higher dimensions (hence the Toeplitz assumption to control complexity), but for the purpose of the analysis using financial and hydrological data, we restrict ourselves to a one-dimensional linear network so each node lies in $\mathbb{Z}$. In the spirit of Diebold and Yılmaz (2009, 2014), who model connectedness within a network using a VAR model, this amounts to a system that consists of an infinite but

countable number of variables indexed by $j \in \mathbb{Z}$. We denote the limiting, infinite dimensional, vectors by $(\boldsymbol{y}_t, \boldsymbol{\epsilon}_t) = \lim_{n\to\infty} (\boldsymbol{y}_{n,t}, \boldsymbol{\epsilon}_{n,t})$, and the ith elements of $\boldsymbol{y}_t, \boldsymbol{\epsilon}_t$ by $y_t^{(i)}, \epsilon_t^{(i)}$, for $i \in \mathbb{Z}$ or $\mathbb{N}$. We next describe the two models that have been shown to generate long memory within an infinite dimensional VAR(1) model such as (1).

Chevillon et al. (2018) These authors make a set of parametric assumptions (their Assumption T) where they specify a mapping such that entries of $\boldsymbol{T}_n$ only depend on a scalar sequence $\delta_n \in (0, 1/2)$ satisfying $n^2(\delta_n - 1/2) = o(1)$. Their Assumption T implies in particular that, as $n \to \infty$, with $(n-1)/4 \in \mathbb{N}$,

$$t_0^{(n)} \to 1/2, \tag{3a}$$

$$t_k^{(n)} = O\left(n^{-1}\right), \quad \text{for } k \neq 0, \tag{3b}$$

$$\sum\nolimits_{k=0}^{n-1} t_k^{(n)} = 1. \tag{3c}$$

Under the additional assumption $\boldsymbol{\epsilon}_{n,t} \sim \mathsf{NID}(\boldsymbol{0}, \boldsymbol{\Sigma}_n)$, with $\boldsymbol{\Sigma}_n$ diagonal, they prove (in their Theorem 1) that, as $n \to \infty$, all components of $\boldsymbol{y}_{n,t}$ tend to independent fractional white noises with identical degrees of integration (all equal to 1/2):

$$\boldsymbol{y}_{n,t} \Rightarrow \boldsymbol{\mu} + \Delta^{-1/2}\boldsymbol{\epsilon}_t,$$

where $\Delta = 1 - L$ and $\Rightarrow$ denotes weak convergence of the associated probability measures. Since the entries of $\boldsymbol{A}_n - \frac{1}{2}\boldsymbol{I}_n$ tend to zero as $n \to \infty$, the cross-sectional dependence between the elements of $\boldsymbol{y}_{n,t}$ vanishes as $n \to \infty$. Yet, as in this setting $\sum_{k=0}^{n-1} t_k^{(n)} = 1$ remains nonzero, the dependence across individual series is sufficient to generate long memory in each of the components of the multivariate process.

Schennach (2018) She considers the limiting structure where $\boldsymbol{T} = \lim_{n\to\infty} \boldsymbol{A}_n = \lim_{n\to\infty} \boldsymbol{T}_n$, i.e., the case of an infinite dimensional network. She assumes that $\boldsymbol{\epsilon}_t$ constitutes a short memory MA(∞) process. The entries (t_k) of $\boldsymbol{T}$ are assumed to satisfy

$$t_0 > 0, \tag{4a}$$

$$\mathsf{card}\{k \in \mathbb{Z},\ t_k > 0\} < \infty, \tag{4b}$$

$$\sum\nolimits_{k=0}^{\infty} t_k = 1. \tag{4c}$$

She then proves (in her Theorem 4) that, for all i, j, there exists a $c_{ij} > 0$ such that, as $k \to \infty$

$$\frac{\partial y_{t+k}^{(i)}}{\partial \epsilon_t^{(j)}} = c_{ij} k^{-1/2} + O\left(k^{-3/2}\right),$$

i.e., the impulse response function of $y_{t+k}^{(i)}$ to a shock $\epsilon_t^{(j)}$ is hyperbolic and its speed of decay corresponds to that of a process that is integrated of order 1/2.

Both Schennach (2018) and Chevillon et al. (2018) find long memory of fractional degree one-half in infinite dimensional networks. They use different approaches and assumptions, but rely on the Toeplitz nature of dependence across the infinite – yet countable – number of variables in the system (or nodes in the network). Both of them consider so-called bistochastic matrices whose rows and columns sum to unity. Schennach focuses on the interactions within the limiting system while CHL consider the evolution in dynamics as the finite system grows larger. Both find that long memory arises only in the infinitely dimensional environment.

Schennach's assumptions on $\boldsymbol{\epsilon}_t$ are less restrictive. She also does not specify the values of the entries of $\boldsymbol{A}$ but assumes that only a finite number of t_k coefficients are nonzero, so that a rotation of $\boldsymbol{A}$ is *banded* (i.e., all subdiagonals are zero beyond a point). Hence, in the system she considers (i.e., the one dimensional version), each variable is only directly connected to a finite number of variables. By contrast, Chevillon et al. (2018) rely on *i.i.d.* shocks and make parametric assumptions on $\boldsymbol{T}_n$. In their setting, variables are directly connected to *all* other variables, but with a connection that becomes weaker as the dimension of the system increases.

Schennach's result is, then, that all response functions of all variables to all shocks exhibit hyperbolic decay, whereas CHL's applies only to the responses of variables to their idiosyncratic shocks in the VAR system.

The similarities between equations (3a)-(3c) and (4a)-(4c) are clear. The main differences relate to specifications of the Toeplitz assumptions, (4b) in particular. In empirical work, the Toeplitz assumption *unreasonably* requires a specific ordering of the variables so we cannot retain it. This implies that we cannot either assume without extra knowledge that specific variables are unconnected. Hence, denoting by $a_{ij}^{(n)}$ the entries of $\boldsymbol{A}_n$ in equation (1), the model can be said to be *long memory prone*, i.e., compatible with the theoretical results of CHL and Schennach, if there exist $\varepsilon, \varepsilon' > 0$ and 'small' such that for all (i, j),

$$a_{ii}^{(n)} \in (1/2 - \varepsilon, 1/2], \tag{5a}$$

$$n \left| a_{ij}^{(n)} \right| < \varepsilon', \tag{5b}$$

$$\sum\nolimits_{k=0}^{\infty} a_{ik}^{(n)} = \sum\nolimits_{k=0}^{\infty} a_{kj}^{(n)} = 1. \tag{5c}$$

The stylized "long memory prone" restrictions on the matrix $\boldsymbol{A}_n$ are that the diagonal elements are close to 0.5, the other elements are close to 0, and the sum of the elements of each row or column is equal to 1. Assumptions (5a)-(5c) participate to the empirical methodology we explore in the next section.

3 Methodology for long memory prone estimation

We turn to the question of estimating $\boldsymbol{A}_n$ to obtain forecasts of $\boldsymbol{y}_{n,t}$ when the latter may exhibit long range dependence. We present a methodology to shrink the estimates of $\boldsymbol{A}_n$ in a manner that is informed by the stylized assumptions (5a)-(5c) resulting from CHL and Schennach. Indeed, it does not seem efficient when the system has a large dimension to

ignore these stylized assumptions altogether, and estimate the VAR by ordinary least-squares (OLS).

An obvious approach to being informed by (5a)-(5c) consists in imposing them strictly such as, e.g., via parametrizing explicitly the elements of $\boldsymbol{A}_n$. For instance CHL use in their Assumption T a mapping that defines all the elements of $\boldsymbol{A}_n$ through a single scalar δ_n. The latter could be estimated by minimum distance or by maximum likelihood (ML). This is certainly too restrictive as explained in the previous section, and we may want to retain a certain degree of flexibility around these restrictions.

We therefore consider intermediate strategies. One of them is penalized regression (e.g., ridge or LASSO), where the least squares criterion is augmented with restrictions whose strength is modulated through penalty parameters. The resulting estimator is shrunk in the direction of the restrictions. Since the theoretical restrictions we consider do not imply the exclusion of specific variables, we prefer to treat all variables in the same way and therefore use ridge estimation.

Bayesian estimation provides another intermediate method, whereby the restrictions are embedded in a prior density, so they hold a priori on average (through the prior expectation of the parameters), but with some degree of uncertainty (through prior positive variances on the parameters or functions thereof). Depending on the degree of tightness of the prior, the prior information pulls the data information more or less strongly in the direction of the restrictions.

We first detail the model in the next subsection, where we resort to an "equation by equation" estimation of the VAR system. Our approaches to ridge regression and Bayesian estimation are exposed in Subsections 3.2 and 3.3, respectively. We denote the resulting model estimated by ridge or Bayesian methods by RAR(1)-X and BAR(1)-X. Many authors have contributed to the Bayesian estimation of VAR models, using different types of prior information, see Karlsson (2013) for a review. The types of restrictions considered in the literature, such as the so-called "Minnesota prior" for unit roots (see Doan, Litterman, and Sims, 1984), or the "long run" forecasting restrictions (Giannone, Lenza, and Primiceri, 2019) are relevant to modeling and forecasting short-memory macroeconomic time series. Our contribution differs in that we use a prior density that shrinks the parameters to values informed by long memory prone restrictions.

3.1 Framework

We consider the estimation of a VAR(1) system, written at date t (dropping the subscript n on $\boldsymbol{A}_n$ and on the processes) as

$$\boldsymbol{y}_t = \boldsymbol{\tau} + \boldsymbol{A}\boldsymbol{y}_{t-1} + \boldsymbol{\epsilon_t}, \tag{6}$$

for the vector $\boldsymbol{y}_t$ consisting of n variables. In this paper, we suggest to estimate parameters $\boldsymbol{\tau}$ and $\boldsymbol{A}$ "equation-by-equation", instead of globally for the entire system. Assuming $\boldsymbol{\epsilon_t}$ is multivariate Gaussian with zero expectation and constant covariance matrix $\boldsymbol{\Sigma}$, then estimating each equation separately by OLS is equivalent to estimating the system jointly

by Maximum Likelihood (ML), even if $\boldsymbol{\Sigma}$ is not diagonal. For Bayesian estimation, equation-by-equation estimation is not equivalent to the joint estimation of all equations, but the latter approach is much more demanding in computing time for the dimensions we are interested in (e.g., 250 in the first empirical application).

A typical equation of the VAR(1) system is an AR(1)-X regression equation that is written at date t as

$$y_t = \gamma_0 + \boldsymbol{\gamma}' \boldsymbol{x}_t + \epsilon_t, \tag{7}$$

where y_t denotes a variable of the system, γ_0 is the intercept parameter, $\boldsymbol{x}_t$ is the column vector containing the first lag of the n variables of the system (including the first lag of y_t), $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_n)'$ is the vector of n slope coefficients of $\boldsymbol{x}_t$, and ϵ_t is an error term assumed to be Gaussian with zero expectation and constant variance σ^2. By convention, for any variable of the VAR, $\boldsymbol{x}_t$ is ordered in such a way that its first element is the lagged dependent variable (y_{t-1}), and $\boldsymbol{\gamma}$ is ordered accordingly: its first element (γ_1) is the autoregressive coefficient of the dependent variable, and the remaining elements are the coefficients of the other lagged variables. For example, if y_t is the first element of $\boldsymbol{y}_t$, $\boldsymbol{\gamma}'$ is the first row of matrix $\boldsymbol{A}$, and γ_0 is the first element of $\boldsymbol{\tau}$.

Over a sample of T observations, write the AR(1)-X equation in the standard regression notation

$$\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{8}$$

where $\boldsymbol{Y} = (y_1, y_2, \ldots, y_T)'$, $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_T)' \sim \mathsf{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I_T})$, $\boldsymbol{Z}$ is a $T \times k$ matrix, with $k = 1 + n$ and t-th row equal to $(1, \boldsymbol{x}_t')$, and $\boldsymbol{\beta} = (\gamma_0, \boldsymbol{\gamma}')'$.

Estimation of $\boldsymbol{\beta}$ by OLS is likely to be imprecise when n is large compared to T, and this will affect the quality of forecasts negatively. To align with the stylized assumptions derived in equations (5a)-(5c), we recommend shrinking the elements of vector $\boldsymbol{\beta} = (\gamma_0, \boldsymbol{\gamma}')'$ in (8) to a target that satisfies:

C1: the autoregressive coefficient (γ_1) is close to 0.5,

C2: the other elements of $\boldsymbol{\gamma}$ are close to 0,

C3: the sum of the elements of $\boldsymbol{\gamma}$ is equal to 1.

In what follows, we explain how suggest introducing these conditions by ridge and Bayesian estimation.

3.2 Ridge estimation

To achieve **C1** and **C2**, we define as the shrinkage target of $\boldsymbol{\beta}$ the vector

$$\boldsymbol{\beta}_0 = (0, d_0, a_0, \ldots, a_0)', \tag{9}$$

where $a_0 = (1 - d_0)/(n - 1)$ is repeated $n - 1$ times. The scalar $d_0 \in (0, 1)$ is the target for the autoregressive coefficient and it determines the target a_0 of the other coefficients that

are shrunk to a value that is close to zero when n is large. We allow $d > 1/2$ despite our assumption (5a) to avoid boundary effects. We use two penalty parameters to control the shrinkage strength: λ_d^2 for the autoregressive parameter, and λ_a^2 for the other coefficients. The penalty function is defined as

$$\lambda_d^2(\gamma_1 - d_0)^2 + \lambda_a^2 \sum_{i=2}^{n} (\gamma_i - a_0)^2 = (\boldsymbol{\beta} - \boldsymbol{\beta}_0)' \boldsymbol{\Lambda}_k (\boldsymbol{\beta} - \boldsymbol{\beta}_0), \text{ where } \boldsymbol{\Lambda}_k = \operatorname{diag}(0, \lambda_{\mathrm{d}}^2, \lambda_{\mathrm{a}}^2, \ldots, \lambda_{\mathrm{a}}^2). \tag{10}$$

In this way, the last n elements of $\boldsymbol{\beta}$ are shrunk to the corresponding elements of $\boldsymbol{\beta}_0$, but the first element of $\boldsymbol{\beta}$ is not shrunk, the value (zero) of the first element of $\boldsymbol{\beta}_0$ being practically irrelevant.

The choice of $\boldsymbol{\beta}_0$ implies that the sum of the last n coefficients is equal to 1 in the target, but the penalty is distributed over the n coefficients. To better achieve **C3**, we add the penalty term $\lambda_s^2(\boldsymbol{\iota}'\boldsymbol{\beta} - 1)^2$, where λ_s^2 is a penalty parameter and $\boldsymbol{\iota} = (0, 1, 1, \ldots, 1)'$ is a vector of k elements. More generally, by writing the penalty as $\lambda_s^2(\boldsymbol{\iota}'\boldsymbol{\beta} - \boldsymbol{\iota}'\boldsymbol{\beta}_0)^2$, we cover the possibility that the target value be different from 1.

The extended ridge (ER) estimator is obtained by minimizing the objective function

$$(\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{\beta}) + (\boldsymbol{\beta} - \boldsymbol{\beta}_0)' \boldsymbol{\Lambda}_k (\boldsymbol{\beta} - \boldsymbol{\beta}_0) + \lambda_s^2(\boldsymbol{\iota}'\boldsymbol{\beta} - \boldsymbol{\iota}'\boldsymbol{\beta}_0)^2, \tag{11}$$

and can be shown to be (see SA, Section A)

$$\boldsymbol{\beta}_{ER} = \left(\boldsymbol{Z}'\boldsymbol{Z} + \boldsymbol{\Lambda}_k + \lambda_s^2 \boldsymbol{\iota}\boldsymbol{\iota}'\right)^{-1} (\boldsymbol{Z}'\boldsymbol{Y} + \boldsymbol{\Lambda}_k \boldsymbol{\beta}_0 + \lambda_s^2 \boldsymbol{\iota}\boldsymbol{\iota}'\boldsymbol{\beta}_0). \tag{12}$$

As usual, the ridge estimator simplifies to the OLS estimator if all the penalty parameters are set to zero.

The values of d_0, λ_d^2, λ_a^2, and λ_s^2 can be chosen by cross validation on a training sample. A grid of values is set a priori for each of them. For each point of the grid, the estimator is computed using 80 percent of the training sample, forecasts are computed for the last 20 percent, and a forecast loss function is computed. The chosen triplet is the value minimizing the loss function over the grid. After this step, estimation is performed on a subsequent sample, and forecasts are computed and evaluated over a post-estimation sample. Details are provided in Section 4.

3.3 Bayesian estimation

Bayesian estimation is based on a prior density for $\boldsymbol{\beta}$ and σ^2, and the likelihood function, the latter resulting from the assumption of normality of the error terms. Since the theory does not provide information on σ^2, its prior "density" $p(\sigma^2)$ is chosen to be the usual "non-informative" prior:

$$p(\sigma^2) \propto 1/\sigma^2. \tag{13}$$

The prior density of $\boldsymbol{\beta}$ is designed to include the theory restrictions **C1-C3**. We opt for a Gaussian density for three reasons: (*i*) it is convenient for computing the posterior density (see Section B of SA); (*ii*) implementation of the restrictions is easily done using four

scalar parameters, as explained below; and (*iii*) the restrictions do not explicitly require an asymmetric density. The prior density is proportional to

$$\exp[-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\boldsymbol{Q}_0(\boldsymbol{\beta}-\boldsymbol{\beta}_0)]\exp[-\frac{1}{2}h_0(\boldsymbol{\beta}'\boldsymbol{\iota}-\boldsymbol{\beta}_0'\boldsymbol{\iota})^2]. \tag{14}$$

The vector $\boldsymbol{\beta}_0$ is defined as in (9) and depends on the scalar hyperparameter d_0 (which is shown below to be the prior mean of γ_1). To explain the prior, let us first fix the scalar hyperparameter h_0 to zero, and discuss the first Gaussian kernel of (14), which corresponds to restrictions **C1** and **C2**. There, $\boldsymbol{\beta}_0$ is the prior expectation, and $\boldsymbol{Q}_0$ is the prior precision matrix. We specify this matrix to be diagonal:

$$\boldsymbol{Q}_0 = \mathrm{diag}(0, 1/\mathrm{s}_\mathrm{d}^2, 1/\mathrm{s}_\mathrm{a}^2, \ldots, 1/\mathrm{s}_\mathrm{a}^2), \tag{15}$$

so that s_d is the prior standard deviation of the autoregressive coefficient and s_a is the prior standard deviation of the other coefficients. The strength with which restrictions **C1** and **C2** are imposed depends on the values of s_d and s_a, respectively: values close to zero correspond to a strong prior belief in favor of the restrictions. For the intercept term, the prior precision is set to zero, so that data information dominates the prior information on this term.

Although the prior expectation $\boldsymbol{\beta}_0$ embeds restriction **C3** that the sum of the last n elements of $\boldsymbol{\beta}$ is equal to 1, the prior variance of this sum is equal to $s_d^2 + (n-1)s_a^2$. Hence, to fix the latter variance to a small value, s_a itself must be fixed to an even smaller value, thus impacting how restriction **C2** is introduced. The second Gaussian kernel of (14) is designed to avoid the potential trade-off between the two restrictions, by adding a prior parameter that controls the strength imposed on the unit sum, independently of the strength imposed on the individual coefficients. Notice that in the second exponential function of (14), we have written $\boldsymbol{\beta}_0'\boldsymbol{\iota}$ after the minus sign, instead of 1, to cover the case where one wants this target to be different from 1, that is, the case where one defines $\boldsymbol{\beta}_0$ differently from (9).

If $\boldsymbol{Q}_0$ in the first kernel is a null matrix, the second kernel specifies that the prior mean of the sum of the last n elements of $\boldsymbol{\beta}$ is equal to $\boldsymbol{\beta}_0'\boldsymbol{\iota}$ (i.e., equal to 1 if $\boldsymbol{\beta}_0$ is given by (9)), and that its prior precision is equal to h_0. Hence a large value of h_0 corresponds to a strongly informative prior on the target value for the sum of the coefficients.

It is well-known that the product of two Gaussian kernels is a kernel of a Gaussian density. Hence, (14) is the kernel of the Gaussian density (see Section A of SA)

$$\boldsymbol{\beta} \sim \mathsf{N}_k(\boldsymbol{\beta}_0, \boldsymbol{V}_0), \tag{16}$$

where

$$\boldsymbol{V}_0 = (\boldsymbol{Q}_0 + h_0\boldsymbol{\iota}\boldsymbol{\iota}')^{-1}. \tag{17}$$

Notice that the expectation of $\boldsymbol{\beta}$ is $\boldsymbol{\beta}_0$, the same as in the first kernel in (14). If $h_0 > 0$, the prior covariance matrix is not diagonal: in fact, the covariances are negative, which is what is needed to reduce the prior standard deviation of $\boldsymbol{\beta}'\boldsymbol{\iota}$ compared to its value when the prior covariance matrix is diagonal. Taking for example values that relate to our empirical exercises below, i.e., $d_0 = 0.5$, $s_d = s_a = 0.02$, $h_0 = 5000$, $n = 250$, then

$\boldsymbol{\beta}_0 = (0, 0.5, 0.002008(249 \text{ times}))$, $\boldsymbol{Q}_0^{-1} = \text{diag}(100, 0.02^2(250 \text{ times}))$, the diagonal of $\boldsymbol{V}_0$ is $(100, 0.01996^2(250 \text{ times}))$, the off-diagonal elements are equal to 0 in the first line (and column), and the other covariances are equal to $-1.59681/10^6$ (the corresponding correlation coefficient being equal to -0.004008). The prior standard deviation of $\boldsymbol{\beta}'\boldsymbol{\iota}$ is equal to 0.014128, i.e., a value much smaller than its value of 0.317 when $h_0 = 0$ so the prior is $\mathsf{N}_k(\boldsymbol{\beta}_0, \boldsymbol{Q}_0^{-1})$, where $\boldsymbol{Q}_0^{-1}$ is defined as $\text{diag}(0, s_d^2, s_a^2, \ldots, s_a^2)$.

To summarize, the prior density (16), when $\boldsymbol{\beta}_0$ is defined by (9) and $\boldsymbol{Q}_0$ by (15), is fully determined by the four scalar hyperparameters d_0, s_d, s_a, and h_0, whatever the dimension n of the VAR. These hyperparameters can be fixed to some values, as in the example above, or they can be chosen for each equation of the VAR by a cross validation procedure similar to the procedure defined in the last paragraph of the previous subsection.

The computation of the posterior mean of $\boldsymbol{\beta}$ for the prior (13)-(16) is performed by a simple Gibbs sampling algorithm defined in SA (Section B). The prior is not conjugate since $\boldsymbol{V}_0$ is not proportional to σ^2. It becomes conjugate if (16) is replaced by

$$\boldsymbol{\beta}|\sigma^2 \sim \mathsf{N}_k(\boldsymbol{\beta}_0, \sigma^2 \boldsymbol{V}_0). \tag{18}$$

The posterior mean corresponding to this conjugate prior is

$$\left(\boldsymbol{Z}'\boldsymbol{Z} + \boldsymbol{Q}_0 + h_0\boldsymbol{\iota}\boldsymbol{\iota}'\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{Y} + \boldsymbol{Q}_0\boldsymbol{\beta}_0 + h_0\boldsymbol{\iota}\boldsymbol{\iota}'\boldsymbol{\beta}_0\right), \tag{19}$$

where (17) has been used. If we set $\boldsymbol{Q}_0 = \boldsymbol{\Lambda}_k$ (by setting $\lambda_d^2 = 1/s_d^2$ and $\lambda_a^2 = 1/s_a^2$) and $h_0 = \lambda_s^2$, this posterior mean is exactly the ER estimator (12). With the non-conjugate prior, one can only derive the conditional (to σ^2) posterior mean of $\boldsymbol{\beta}$, which can be expressed (see SA, Section B) as

$$\boldsymbol{\beta}_*(\sigma^2) = \left(\frac{\boldsymbol{Z}'\boldsymbol{Z}}{\sigma^2} + \boldsymbol{Q}_0 + h_0\boldsymbol{\iota}\boldsymbol{\iota}'\right)^{-1}\left(\frac{\boldsymbol{Z}'\boldsymbol{Y}}{\sigma^2} + \boldsymbol{Q}_0\boldsymbol{\beta}_0 + h_0\boldsymbol{\iota}\boldsymbol{\iota}'\boldsymbol{\beta}_0\right). \tag{20}$$

This differs from (19) only by the presence of σ^2. The Gibbs sampler defined in SA (Section B) is a way to marginalize $\boldsymbol{\beta}_*(\sigma^2)$ with respect to σ^2. The resulting unconditional posterior mean of $\boldsymbol{\beta}$ then differs from the corresponding posterior mean/ER estimator when the prior is conjugate.

3.4 Forecasting

After obtaining a point estimate of $\boldsymbol{\beta}$ for an equation of the VAR system, such as the OLS estimator, the extended ridge estimator, or the posterior mean, a one-step ahead point forecast of y_{t+1} is simply obtained using a point estimate of (7) and the regressor $\boldsymbol{x}_{t+1}$ observable at time t. This is equivalent to using the point estimates of all equations to form the estimated $\boldsymbol{\tau}$ and $\boldsymbol{A}$ of the VAR system (6), and then computing one-step ahead point forecasts as $\hat{\boldsymbol{y}}_{t+1} = \hat{\boldsymbol{\tau}} + \hat{\boldsymbol{A}}\boldsymbol{y}_t$.

To compute h-step ahead forecasts, with $h > 1$, we can use either iterated multistep forecasting or direct multistep forecasting. An iterated h-step ahead forecast is based on the

estimated VAR and computed recursively as $\hat{\boldsymbol{y}}_{t+h} = \hat{\boldsymbol{\tau}} + \hat{\boldsymbol{A}}\hat{\boldsymbol{y}}_{t+h-1}$. This approach amounts to compute $\hat{\boldsymbol{A}}^h$, i.e., to forecast all variables even if one is interested in only a subset of them (even just a single one). Hence, the forecast of a variable of interest may be contaminated by erroneous and imprecise forecasts of the other variables (see, e.g., Schorfheide, 2005, or Chevillon and Hendry, 2005).

If the objective is to forecast a subset of the series, or if one wishes to avoid the drawback inherent in the iterated multistep method highlighted above, the direct multistep forecasting method is preferable. The method consists in directly projecting $\boldsymbol{y}_t$ on its lag $\boldsymbol{y}_{t-h}$, as in

$$\boldsymbol{y}_t = \boldsymbol{\tau}_h + \boldsymbol{A}_h \boldsymbol{y}_{t-h} + \boldsymbol{u_t}. \tag{21}$$

Ignoring that $\boldsymbol{A}_h = \boldsymbol{A}^h$, a typical equation of (21) can be cast in the form of (7) and (8), adapting the definitions of $\boldsymbol{Y}$, $\boldsymbol{x}_t$ and $\boldsymbol{Z}$, and ignoring the dependence in $\boldsymbol{u_t}$ induced by recursive substitution. For $h > 1$, we denote the equation corresponding to (8) by

$$\boldsymbol{Y_{(h)}} = \boldsymbol{Z}_{(h)}\boldsymbol{\beta_h} + \boldsymbol{u_{(h)}}. \tag{22}$$

Hence, the system (21) can be estimated equation by equation, by OLS, ridge and Bayesian estimation, as is the case when $h = 1$. By proceeding in this spirit, no direct use is made in estimation of the relation $\boldsymbol{A}_h = \boldsymbol{A}^h$, because this would imply that the regression coefficients of the different equations of (21) are nonlinear functions of the same parameters (those of $\boldsymbol{A}$), so that equation by equation estimation would be pointless. In brief, the parameter $\boldsymbol{\beta}_h$ is not treated as a function of the underlying parameters of $\boldsymbol{A}$.

Yet, for ridge and Bayesian estimations, we allow the target towards which $\boldsymbol{\beta}_h$ is shrunk to be function of h, and we denote it by $\boldsymbol{\beta}_{h,0}$. The target $\boldsymbol{\beta}_{h,0}$ relates to the first row of $\boldsymbol{A}_0^h$, as in the case for $h = 1$, where $\boldsymbol{\beta}_0$ is directly the first row of $\boldsymbol{A}_0 = d_0\boldsymbol{I}_n + a_0\left(\boldsymbol{J}_n - \boldsymbol{I}_n\right)$, with $\boldsymbol{J}_n$ being a matrix of ones, and $a_0 = (1 - d_0)/(n - 1)$. In practice, we choose the last n elements of $\boldsymbol{\beta}_{h,0}$ to be close to the first row of $\boldsymbol{A}_0^h$ when n is large relative to h: this is achieved by setting (see SA, Section C)

$$\boldsymbol{\beta}_{h,0} = \left(0, d_0^h, \frac{1 - d_0^h}{n-1}, \ldots, \frac{1 - d_0^h}{n-1}\right)'. \tag{23}$$

The extended ridge estimator for the corresponding $\boldsymbol{\beta}_h$ is defined as in (12), replacing $\boldsymbol{\beta}_0$ with $\boldsymbol{\beta}_{h,0}$, the penalty parameters and the value of d_0 being chosen by cross validation for each horizon h. For Bayesian estimation, we use the same type of prior as when $h = 1$ (i.e., (13) and (16)), also replacing $\boldsymbol{\beta}_0$ with $\boldsymbol{\beta}_{h,0}$. Forecasts for specific elements of $\boldsymbol{y}_t$ can readily be formed by estimating only specific rows of (21), so that forecasts are obtained from the corresponding individual equations, as in the case $h = 1$.

4 Empirical illustrations

In this section, we provide two applications to data where long memory has been documented in the literature and for which multiple series supposedly belonging to the same system are

available (so we do not explore additional exogenous variables to retain the autoregressive nature of the system dynamics). In both applications, and for a large number of variables, we compare out-of-sample forecasts obtained from the AR(1)-X equation (7) estimated using OLS, ridge, and Bayesian estimation, as defined in Section 3. We also include in the comparison forecasts obtained using three benchmark models, which are purely univariate time series models in the sense that they specify y_t as a function of the past of y_t only. The six models we consider and their estimation method are listed below:

1. AR(1): $y_t = \gamma_0 + \gamma_1 y_{t-1} + \epsilon_t$, estimated by OLS.

2. ARFIMA(1,d,0): $(1-L)^d(y_t - \gamma_0 - \gamma_1 y_{t-1}) = \epsilon_t$, estimated by Gaussian ML.

3. HAR (Corsi, 2009): $y_t = \gamma_0 + \gamma_1 y_{t-1} + \gamma_2 \frac{1}{5}\sum_{i=1}^{5} y_{t-i} + \gamma_3 \frac{1}{21}\sum_{i=1}^{21} y_{t-i} + \epsilon_t$, estimated by OLS.

4. AR(1)-X: $y_t = \gamma_0 + \gamma_1 y_{t-1} + \sum_{i=2}^{n} \gamma_i x_{i,t-1} + \epsilon_t$, estimated by OLS. This is the model defined in (7).

5. RAR(1)-X: this model is identical to the AR(1)-X. The estimator is the extended ridge estimator defined by (12), see Section 3.2. Recall that in this case we shrink γ_1 towards d_0 with penalty parameter λ_d^2, γ_i toward $(1-d_0)/(n-1)$ $(\forall i > 1)$ with penalty λ_a^2, and $\sum_{i=1}^{n} \gamma_i$ towards 1 with a penalty of λ_S^2. The penalty parameters (i.e., λ_d^2, λ_a^2 and λ_S^2) and d_0 are chosen by cross validation as explained at the end of Section 3.2; details are provided in SA, Section D.

6. BAR(1)-X: this specification is also identical to the AR(1)-X but it is estimated by the Bayesian method presented in Section 3.3. The prior for the variance of ϵ_t is non-informative, see (13), and the prior for the regression coefficients $\boldsymbol{\beta} = (\gamma_0, \gamma_1, \gamma_2, \ldots, \gamma_n)'$ is the Gaussian density defined by (16) together with (9), (17) and (15). More specifically, the prior on γ_0 is quasi-noninformative (with a mean of 0 and a variance of 100), the prior mean of γ_1 is set equal to d_0, and the prior mean of γ_i, for all $i > 1$, is set to $(1-d_0)/(n-1)$. The prior precision of γ_1 is $1/s_d^2 + h_0$, the prior precision of γ_i $(i > 1)$ is $1/s_a^2 + h_0$. The co-precisions (the off-diagonal elements of the inverse of V_0) are all set to h_0. The larger h_0, the smaller the prior variance for the difference between the sum of the last n elements of $\boldsymbol{\beta}$ and the corresponding sum in the prior mean (equal to 1 for (9)). The prior parameters d_0, s_d, s_a and h_0 are chosen by cross validation (see SA, Section D for details).

For higher forecast horizons, $h > 1$, we use iterated multistep forecasts (i.e., recursive substitution) for AR(1), ARFIMA and HAR, and direct multistep forecasts for AR(1)-X, RAR(1)-X and BAR(1)-X. As discussed in Subsection 3.4, this avoids contaminating forecasts across variables when additional (non autoregressive) regressors are present.

The out-of-sample forecasts (at differing horizons) are compared to the observed values using both the mean absolute forecast error (MAFE) and the mean squared forecast error

(MSFE) loss functions. These loss functions are defined for each model m as

$$MAFE_h^{(m)} = \frac{1}{T_h}\sum_{t=1}^{T_h}|\hat{y}_{t,h}^{(m)} - y_t| \quad \text{and} \quad MSFE_h^{(m)} = \frac{1}{T_h}\sum_{t=1}^{T_h}(\hat{y}_{t,h}^{(m)} - y_t)^2, \tag{24}$$

where h is the forecast horizon, T_h is the number of forecasts at horizon h, and $\hat{y}_{t,h}^{(m)}$ is the forecast of y_t at horizon h by model m. The comparison tool is the model confidence set (MCS) procedure of Hansen, Lunde, and Nason (2011), see SA, Section D for details about the implementation.

In the first application, y_t is the logarithm of a measure of daily volatility for a set of 250 U.S. company stocks. In the second application, it is the logarithm of the monthly seasonally adjusted river streamflows at 97 locations in the Columbia river basin.

4.1 Daily realized volatilities of U.S. stocks

The initial dataset (purchased from tickdatamarket) consists of transaction prices at the 1-second sampling frequency for 1,412 stocks from the NYSE, AMEX and NASDAQ markets, for the period ranging from January 1st, 1991 to October, 14, 2019 covering 7,510 trading days. We ordered the stocks by decreasing average daily transaction volume, and kept the 250 largest capitalization stocks for the period from 2005-01-03 to 2017-07-24 (3,276 trading days). These start and end dates were chosen to maximize the number of available series out of the larger dataset of 7,510 trading days.

We aggregated the data at the 5-minute frequency and computed the MedRV estimator of Andersen, Dobrev, and Schaumburg (2012), a non-parametric robust to jumps estimator of the integrated variance. If $r_{t,i}$ is the ith 5-minute return of a given stock on a day t containing $M = 78$ (since trading is from 9:30 to 16) such returns, $\log(MedRV_t)$ (denoted by y_t hereafter) is computed as the logarithm of

$$MedRV_t = \frac{\pi}{6 - 4\sqrt{3} + \pi}\frac{M}{M-2}\sum_{i=3}^{M}\text{med}(|r_{t,i}|, |r_{t,i-1}|, |r_{t,i-2}|)^2,$$

where med $(\cdot)$ denotes the median. Notice that VAR and Vector Heterogenous Autoregressive (VHAR) models have been used respectively by Anderson and Vahid (2007) Cubadda, Hecq, and Riccardo (2019) for forecasting realized volatility measures. The six competing models are estimated on rolling windows of $T = 1,000$ observations. They are estimated first on the sample spanning the period from 2005-01-03 to 2008-10-31, and h-step ahead forecasts of y_t are computed for ten horizons ($h = 1, 2, \ldots, 10$) leading to a total number of $2,277$ minus h forecasts. The parameters estimated on each window are kept constant to produce 25 consecutive forecasts and then re-estimated on the next window of T observations obtained by a translation of one period. To speed up the estimation, the four tuning parameters of the RAR(1)-X and BAR(1)-X models are only estimated once by cross validation on the first window of T observations and then kept constant (see Section D of the SA). Rolling forward

is continued until the last possible window of the full sample. The models are estimated for each of the 250 available series.

The presence of long memory in the volatility of the log-returns of financial assets is a well recognized stylized fact (see Baillie, Bollerslev and Mikkelsen, 1996, Breidt, Crato and de Lima, 1998, Comte and Renault, 1998, Andersen, Bollerslev, Diebold and Ebens, 2001, among others). For the sake of illustration, the average value (over the 250 series) of the estimated d parameters of the ARFIMA$(1, d, 0)$ obtained on the full sample is about 0.48 (with a standard deviation of 0.02). Despite the numerous models that generate long memory in volatility (from, e.g., Giraitis, Robinson and Surgailis, 2000, Hurvich, Moulines and Soulier, 2005, and Lieberman and Phillips, 2008, who focus specifically on realized volatility), there is however a debate on whether realized variance displays long memory or roughness. Recently, Shi and Yu (2021) shown that, when including an AR term in an ARFIMA model, both the semiparametric methods and maximum likelihood methods have trouble distinguishing a long memory process from a rough process.

To make sure that the empirical results are not specific to the chosen forecasting period, we compare the forecasting performance of the competing models on rolling windows. More specifically, the left panels of Figure 1 show, for three forecast horizons ($h = 1, 5$ and 5), the averages (over the 250 stocks) of the MAFE loss functions for a sequence of rolling samples of 250 forecasts. The right panels of the same figure shows the corresponding time evolution of the frequencies at which each model belongs to the MCS at the confidence level of 75% (denoted MCS75). A frequency of 50 (percent) for model m at date t means that the model m is in the MCS75 for fifty percent of the 250 series, the MCS75 in question being obtained using the loss function computed from the 250 forecasts ending at date t. Notice that to reduce the computing time, the MCS procedure is not applied to every consecutive window of 250 forecasts, but to every 25-th window, so that each line is drawn by joining 82 values. The corresponding figure based on the MSFE loss function is reported in the SA (Section E) and shows very similar results as Figure 1.

To complement the figures, Table 1 reports the average values (over the 82 windows) of the frequencies at which each model belongs to the MCS75, for all horizons $h = 1, \ldots, 10$. The following comments can be made, and they apply equally to both loss functions:

- AR(1) and AR(1)-X are strongly outperformed by the other models over the forecast period. Their average losses are larger (often strongly) than those of the other models. The frequencies of inclusion of these models in the MCS75 are very often smaller than 10 percent, and almost never above 20. This is confirmed in Table 1, where these two models are by far the least present on average in MCS75, whatever the forecast horizon.

- ARFIMA and HAR perform comparably, especially considering their average losses. Their frequencies of inclusion in the MCS75 are also similar, but sometimes more different than the corresponding losses. Overall, these frequencies fluctuate between 25 and 50 percent until mid-2012, and then between 50 and 70 percent. Table 1 shows that on average these two models belong to the MCS75 in about 50% of the cases, whatever the forecast horizon.

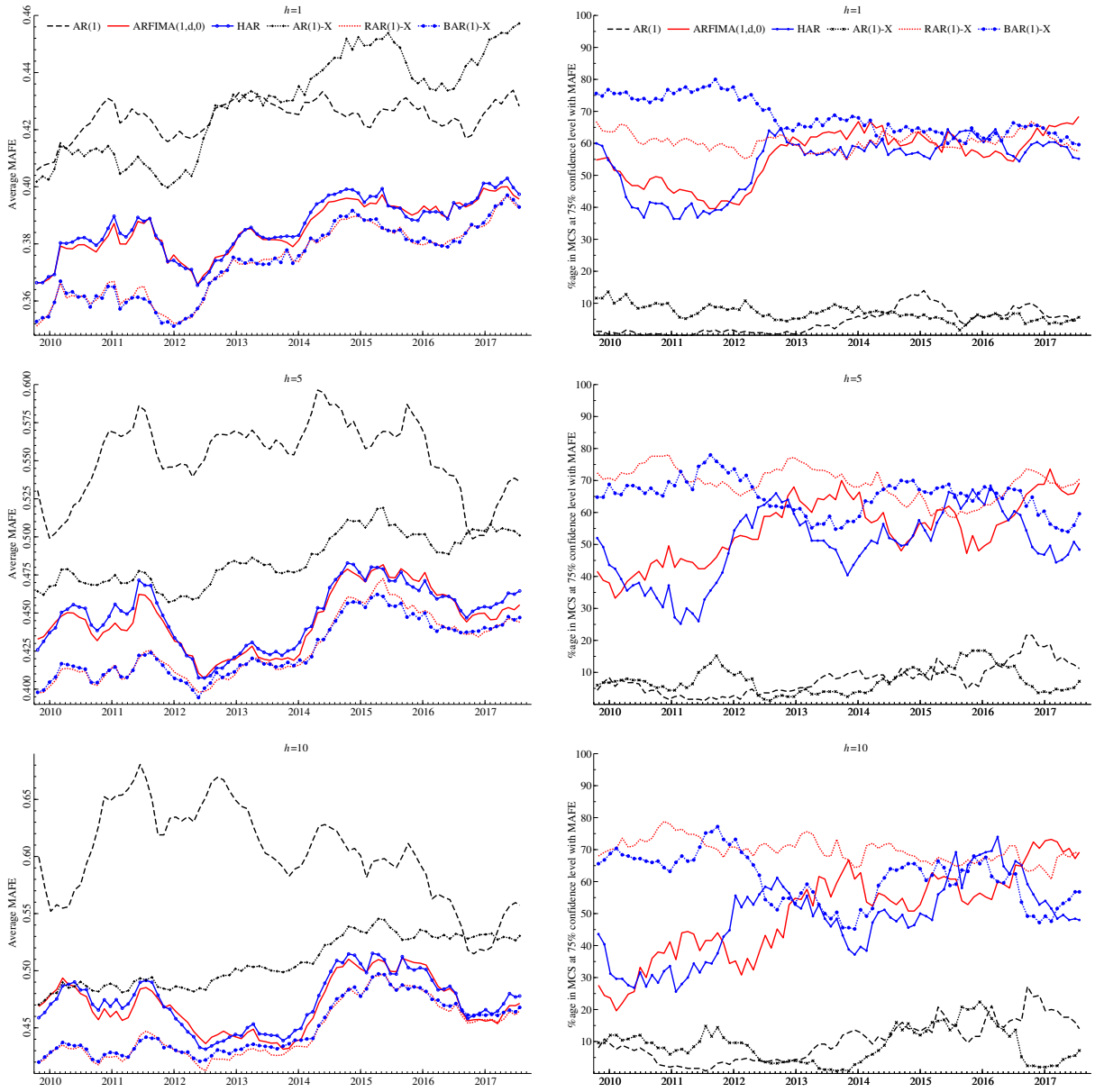

Figure 1: Left panels: Average (over the 250 series) MAFE loss computed on rolling windows of 250 observations. Right panels: Frequencies (over the 250 series), at each date, at which each model belongs to the MCS (at the 75% confidence level) for the MAFE loss function. The top block is for the forecast horizon $h = 1$, the middel one for $h = 5$, and the bottom one for $h = 10$.

Table 1: Average frequencies (over the 250 series and the 82 rolling windows), at which each model belongs to the MCS (at the 75% confidence level) for each forecast horizon h

h	AR(1)	ARFIMA(1,d,0)	HAR	AR(1)-X	RAR(1)-X	BAR(1)-X
			MSFE			
1	4.03	53.88	51.40	7.06	57.73	67.03
2	5.21	45.45	46.20	5.14	69.33	59.53
3	6.60	47.97	47.01	5.79	71.73	58.87
4	7.49	50.49	46.88	7.16	70.51	60.08
5	8.58	53.31	48.19	7.55	68.07	63.78
6	8.58	53.31	48.19	7.55	68.07	63.78
7	9.53	51.18	48.73	7.23	70.19	59.00
8	9.69	51.77	49.05	8.00	70.28	59.47
9	10.11	50.67	47.53	8.10	69.08	61.57
10	10.93	49.19	46.10	8.82	68.92	59.97
			MAFE			
1	4.17	56.01	53.87	6.96	60.64	68.28
2	4.92	46.25	47.69	5.34	70.95	60.06
3	6.21	48.79	47.89	6.15	72.52	60.28
4	7.03	52.02	48.44	7.48	71.51	61.64
5	7.90	54.79	49.91	7.55	69.11	65.03
6	7.90	54.79	49.91	7.55	69.11	65.03
7	8.78	52.40	50.40	7.78	71.66	60.86
8	9.12	53.09	51.61	8.24	71.65	61.72
9	9.56	51.44	49.78	8.05	69.95	62.70
10	10.27	50.44	48.16	8.93	69.65	60.75

- RAR(1)-X and BAR(1)-X perform comparably and better than ARFIMA and HAR, with smaller losses and higher frequencies. The latter are most of the time between 65 and 75%, though for horizons 5 and 10, the RAR frequencies are higher (by 10 to 20 points) than the BAR frequencies in 2012 and 2013, and again from March 2016. Table 1 indicates that for $h = 1$, BAR(1)-X is on average the most frequently in the MCS75 but for $h > 1$, RAR(1)-X is even better than BAR(1)-X, with an average frequency in the MCS75 around 70%, i.e., 20 points higher than ARFIMA and HAR.

In brief, the use of the theoretical constraints in the AR(1)-X model through the proposed Bayesian and ridge estimation methods strongly improves the model forecasting performance with respect to OLS. The bad performance of the latter is due to a lack of precision because 251 coefficients are estimated using 1,000 observations, whereas the shrinkage methods impose a relevant theoretical structure on the estimated coefficients. The performance of the shrinkage methods is also most of the time significantly superior to that of the ARFIMA

and HAR models; this difference can be attributed to the use of a larger, but relevant, information set.

4.2 Monthly river streamflows in the Columbia river basin

Natural streamflows play a significant role in shaping biological communities and they regulate ecological processes in local ecosystems. In most industrialized economies streamflows are modified as a result of human activity (agricultural, industrial, ...), and regulated as such. Forecasting future flows is essential for planning dam discharges and adaptation.

In the hydrology community, many studies have been carried out to the test the long memory of streamflows, following the seminal paper of Hurst (1951) on dimensioning dams for the Nile river, and which pioneered the literature on long memory. For instance, Montanari, Rosso and Taqqu (1997) applied ARFIMA modelling to the monthly and daily inflows of Lake Maggiore, Italy. Depending on the modelling strategy, their confidence interval for the degree of long memory varies with a $[.35, 45]$ range. This is a feature that has often been documented in the hydrology and streamflow forecasting literatures, and we provide a short overview of this literature in the SA (Section E).

To illustrate our modelling approach, we assess its forecasting accuracy using the Modified Streamflow dataset (M series) of the Columbia river basin provided by the Bonneville Power Administration (BPA), the United States Army Corps of Engineers and the U.S. Bureau of Reclamations. In their 2020 Level Modified Streamflow Report, Dakhlalla, *et al.* (2020, Section 1, page 1) explain that "Since irrigation practices have changed since the historical streamflows were observed, the historical streamflows have been adjusted to account for current levels of irrigation depletions." Hence "Modified streamflows are historical streamflows that would have been observed if current irrigation depletions (as of year 2018) existed in the past and if the effects of river regulation were removed." These modified flows allow for intertemporal comparisons of the natural inflows since they are adjusted to a common level of irrigation development and evaporation in upstream reservoirs and lakes, and they reflect no regulation by dams. They are recorded and computed at 97 locations in the Columbia river basin over 90 years (October 1928-December 2018, i.e., $1,083$ monthly observations). We model and forecast the logarithm of the monthly series and we adjust them for seasonal variations using X12arima in Oxmetrics version 8.10. Interactions between locations are not purely hierarchical (up or down stream) but reflect inter alia the dependence in climate, geography and weather variations.

To confirm the presence of long memory in the data, we estimated on the 97 series and on the full sample an ARFIMA$(0, d, 0)$ and an ARFIMA$(1, d, 0)$ model by Gaussian ML. The average $\widehat{d}$ is equal to 0.45 with a standard deviation of 0.06 for the former and 0.21 with a standard deviation of 0.19 for the latter.

We report the results of a forecasting comparison of the six models listed at the beginning of this section. The estimation and forecasting are organized as described in the previous subsection, with rolling windows of 400 observations for estimation but because both the number of series and the number of observations are smaller than in the previous application, all models are re-estimated each time a new observation becomes available. The first window

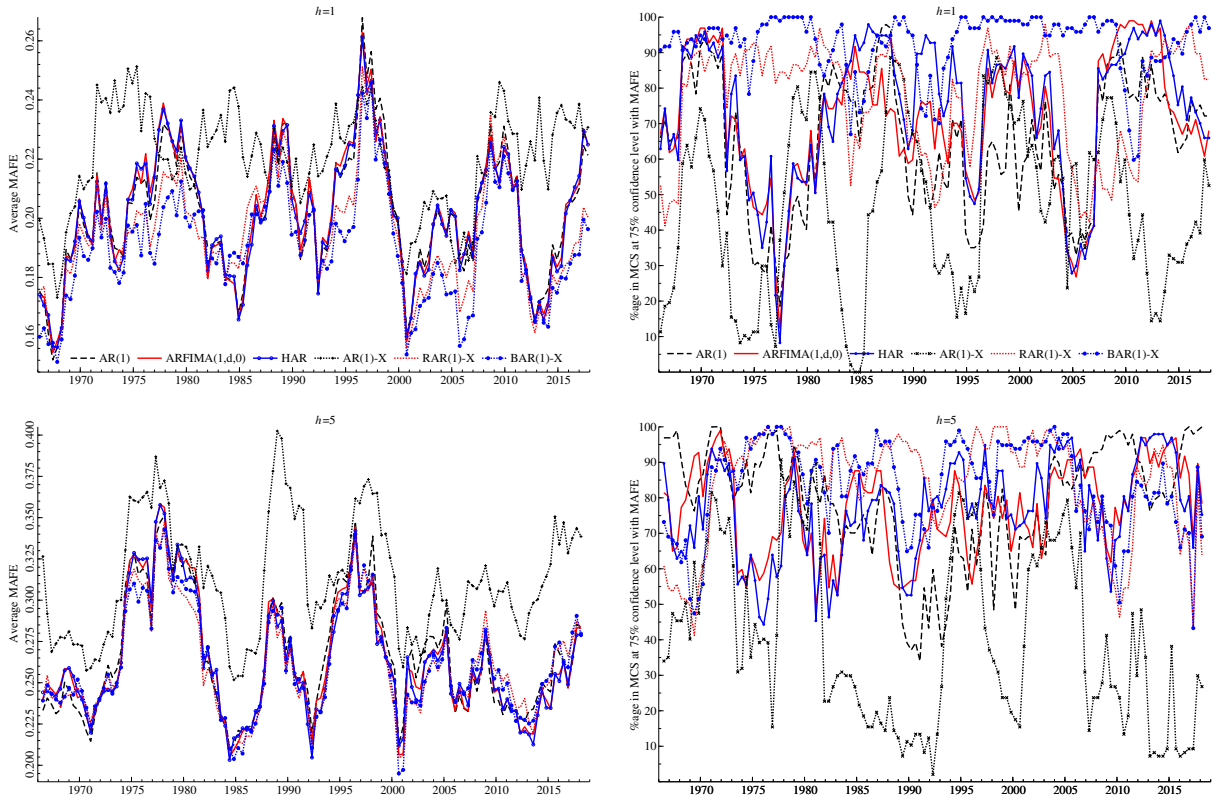

Figure 2: Left panels: Average (over the 97 series) MAFE loss computed on rolling windows of 50 observations. Right panels: Frequencies (over the 97 series), at each date, at which each model belongs to the MCS (at the 75% confidence level) for the MAFE loss function. The top block is for the forecast horizon $h = 1$, the bottom one for $h = 5$.

corresponds to the period October 1928-January 1962. We obtain a total of 683 minus h forecasts for the 97 series, where $h = 1, 2, \ldots, 6$.

Figure 2 reports the forecasting results in the same way as Figure 1, but for horizons 1 and 5; the corresponding figure for MSFE loss (see SA, Section F). The MCS procedure is applied every 5-th window of 50 forecasts leading to a total of 125 tests. Table 2 reports the average frequencies for the six horizons.

Table 2: Average frequencies (over the 97 series and the 125 rolling windows), at which each model belongs to the MCS (at 75% confidence level)

h	AR(1)	ARFIMA(1,d,0)	HAR	AR(1)-X	RAR(1)-X	BAR(1)-X
			MSFE			
1	68.65	73.65	76.39	48.32	81.43	93.85
2	69.10	72.21	76.60	42.02	80.78	94.16
3	71.15	72.58	75.79	50.68	78.85	91.72
4	71.74	71.43	74.00	46.57	83.50	88.79
5	77.60	74.32	74.49	40.10	83.50	83.10
6	76.21	70.80	71.51	35.82	83.24	82.28
			MAFE			
1	64.49	70.68	73.96	46.40	77.62	92.63
2	67.81	70.77	75.34	39.17	78.97	93.28
3	70.55	71.98	75.59	50.11	77.15	90.98
4	72.84	72.48	75.08	46.64	82.22	89.00
5	79.07	76.76	76.58	40.46	84.20	83.57
6	78.08	72.65	73.23	36.41	84.09	82.80

Theses results lead to the following comments:

- The AR(1)-X model (estimated by OLS) has the worst forecasting performance, whatever the forecasting horizon. Nevertheless, it is included in the MCS75 for 35 to 50% (depending on h) of the series on average, but through time, these frequencies fluctuate considerably, as can be seen on Figure 2.

- AR(1), ARFIMA, HAR have lower average losses and higher frequencies of inclusion in the MCS75 than AR(1)-X. The average frequencies of these models (see in Table 2) vary between 65 and 75% (depending on h).

- RAR(1)-X and BAR(1)-X are performing better than the other models, with average frequencies between 77 and 94%. BAR(1)-X belongs to the MCS75 in more than 90% of the cases for $h = 1, 2$ and 3, which is about 10 to 15 points higher than RAR(1)-X. For $h = 4$, the difference is about 5 points in favor of BAR, and for $h = 5$ and 6, the two models perform similarly.

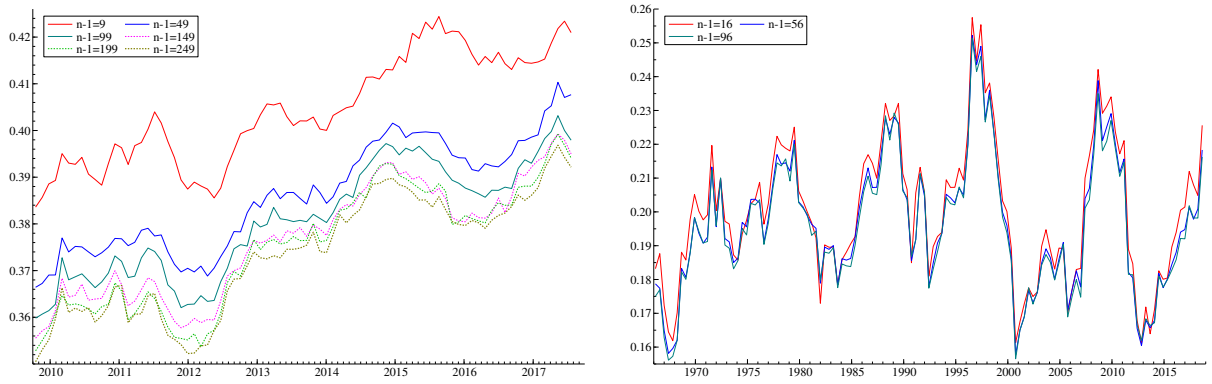

Figure 3: Average MAFE loss (for horizon 1) computed on rolling windows for an increasing number of X variables. Left panel: RAR(1)-X model for realized volatilities; the line for $n-1=249$ is the same as in the top left graph of Figure 1. Right panel: BAR(1)-X model for modified river streamflows; the line for $n-1=96$ is the same as in the top left graph of Figure 2.

- AR(1)-X and BAR(1)-X have inclusion frequencies that are more stable through time than the other models.

4.3 Impact of the cross-sectional dimension

In both applications, we used above the maximum possible number of "X" variables in the AR(1)-X equation, which is the number of available series (n) minus one. We now consider using a smaller number of additional series. We focus on the forecast results of the best forecasting model for each application: RAR(1)-X for the first application, and BAR(1)-X for the second. In each case, we perform estimations and forecasts for all series (250 and 97, respectively), with a number of non-autoregressive regressors set to $n-1=9, 49, 99, 149, 199$ and 249 in the first application (249 corresponds to the results reported in Subsection 4.1), and $n-1=16, 56$ and 96 in the second (with results for 96 as in Subsection 4.2). Figure 3, which reports the MAFE the loss functions for several values of $n-1$, shows clearly that increasing n reduces the values of loss functions (with some rare exceptions at some dates), and that the incremental decrease in loss becomes less important when n gets closer to the largest possible value.

5 Conclusions

This paper considers a novel approach in empirical work for modeling variables exhibiting long memory, using one lag of a large cross-section of related variables instead of the usual technique that models variables using a long history of their own lags. This approach is based

on combining two theoretical contributions that prove that long memory in a variable can be caused by its dependences within a large system or network. We provide two estimation methods that harness the informativeness of the theoretical models and use them to drive the estimation, either via an extended ridge regression that shrinks the estimates toward a structure derived from the theory, or by using the latter to design an informative prior in a Bayesian setup.

In applications to realized volatilities of stocks and river streamflows, we show that the proposed modeling and estimation strategy improves upon standard univariate models (ARFIMA and HAR models) in terms of predicting series characterized by the presence of long memory. Such results suggest that it may be fruitful to model variables that exhibit long range dependence by using one lag of a set of related variables, provided that the cross-sectional dimension is large.

The success of the proposed approach opens the door to more exploration about the impact that dependencies within a large network or system may have on each variable's idiosyncractic long range persistence. It could readily be extended to include richer short-term dynamics (e.g., through VAR(p) modeling, with $p > 1$) or higher dimensional networks.

Appendix

A: Proof of (12) and of (16)

Proof of (12): notice that $(\boldsymbol{\beta}'\boldsymbol{\iota}-\boldsymbol{\beta}_0'\boldsymbol{\iota})^2 = (\boldsymbol{\beta}'\boldsymbol{\iota}-\boldsymbol{\beta}_0'\boldsymbol{\iota})(\boldsymbol{\beta}'\boldsymbol{\iota}-\boldsymbol{\beta}_0'\boldsymbol{\iota})' = \boldsymbol{\beta}'\boldsymbol{\iota}\boldsymbol{\iota}'\boldsymbol{\beta}-2\boldsymbol{\beta}'\boldsymbol{\iota}\boldsymbol{\iota}'\boldsymbol{\beta}_0+\boldsymbol{\beta}_0'\boldsymbol{\iota}\boldsymbol{\iota}'\boldsymbol{\beta}_0$. By developing the quadratic forms, the ER objective function (11) is equal to $\boldsymbol{\beta}'\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\beta} - 2\boldsymbol{\beta}'\boldsymbol{Z}'\boldsymbol{Y}+\boldsymbol{\beta}'\boldsymbol{\Lambda}_k\boldsymbol{\beta}-2\boldsymbol{\beta}'\boldsymbol{\Lambda}_k\boldsymbol{\beta}_0+\lambda_s^2\boldsymbol{\beta}'\boldsymbol{\iota}\boldsymbol{\iota}'\boldsymbol{\beta}-2\lambda_s^2\boldsymbol{\beta}'\boldsymbol{\iota}\boldsymbol{\iota}'\boldsymbol{\beta}_0+\boldsymbol{Y}'\boldsymbol{Y}+\lambda\boldsymbol{\beta}_0'\boldsymbol{\Lambda}_k\boldsymbol{\beta}_0+\lambda_s^2\boldsymbol{\beta}_0'\boldsymbol{\iota}\boldsymbol{\iota}'\boldsymbol{\beta}_0$. Solving the first-order condition yields the solution (12).

Proof of (16): to show that the kernel (14) corresponds to (16), we can write that (14) is equal to

$$\exp\{-\frac{1}{2}[(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\boldsymbol{Q}_0(\boldsymbol{\beta}-\boldsymbol{\beta}_0)+h_0(\boldsymbol{\beta}'\boldsymbol{\iota}-\boldsymbol{\beta}_0'\boldsymbol{\iota})(\boldsymbol{\beta}'\boldsymbol{\iota}-\boldsymbol{\beta}_0'\boldsymbol{\iota})']\} = K_0\exp[-\frac{1}{2}f(\boldsymbol{\beta})],$$

where K_0 does not depend on $\boldsymbol{\beta}$ and

$$f(\boldsymbol{\beta}) = \boldsymbol{\beta}'(\boldsymbol{Q}_0+h_0\boldsymbol{\iota}\boldsymbol{\iota}')\boldsymbol{\beta}-2\boldsymbol{\beta}'(\boldsymbol{Q}_0\boldsymbol{\beta}_0+h_0\boldsymbol{\iota}\boldsymbol{\beta}_0'\boldsymbol{\iota}) = (\boldsymbol{\beta}-\bar{\boldsymbol{\beta}}_0)'\boldsymbol{V}_0^{-1}(\boldsymbol{\beta}-\bar{\boldsymbol{\beta}}_0)+C_0,$$

where $\boldsymbol{V}_0^{-1} = \boldsymbol{Q}_0+h_0\boldsymbol{\iota}\boldsymbol{\iota}'$, $\bar{\boldsymbol{\beta}}_0 = \boldsymbol{V}_0(\boldsymbol{Q}_0\boldsymbol{\beta}_0+h_0\boldsymbol{\iota}\boldsymbol{\beta}_0'\boldsymbol{\iota})$, and $C_0 = \bar{\boldsymbol{\beta}}_0'\boldsymbol{V}_0^{-1}\bar{\boldsymbol{\beta}}_0$ does not depend on $\boldsymbol{\beta}$. Hence, the prior density depends on $\boldsymbol{\beta}$ only through $\exp[-\frac{1}{2}(\boldsymbol{\beta}-\bar{\boldsymbol{\beta}}_0)'\boldsymbol{V}_0^{-1}(\boldsymbol{\beta}-\bar{\boldsymbol{\beta}}_0)]$, which is the kernel of the Gaussian density $N_k(\bar{\boldsymbol{\beta}}_0,\boldsymbol{V}_0)$. To show that this Gaussian density

is the same as (16), we show that $\bar{\boldsymbol{\beta}}_0 = \boldsymbol{\beta}_0$:

$$\begin{aligned}
\bar{\boldsymbol{\beta}}_0 &= (\boldsymbol{Q}_0 + h_0\boldsymbol{\iota}\boldsymbol{\iota}'^{-1}(\boldsymbol{Q}_0\boldsymbol{\beta}_0 + h_0\boldsymbol{\iota}\boldsymbol{\beta}_0'\boldsymbol{\iota}) = (\boldsymbol{Q}_0^{-1} - \frac{h_0\boldsymbol{Q}_0^{-1}\boldsymbol{\iota}\boldsymbol{\iota}'\boldsymbol{Q}_0^{-1}}{1 + h_0\boldsymbol{\iota}'\boldsymbol{Q}_0^{-1}\boldsymbol{\iota}})(\boldsymbol{Q}_0\boldsymbol{\beta}_0 + h_0\boldsymbol{\iota}\boldsymbol{\beta}_0'\boldsymbol{\iota}) \\
&= \boldsymbol{\beta}_0 + h_0\boldsymbol{Q}_0^{-1}\boldsymbol{\iota}\boldsymbol{\beta}_0'\boldsymbol{\iota} - \frac{1}{1 + h_0\boldsymbol{\iota}'\boldsymbol{Q}_0^{-1}\boldsymbol{\iota}}(h_0\boldsymbol{Q}_0^{-1}\boldsymbol{\iota}\underbrace{\boldsymbol{\iota}'\boldsymbol{Q}_0^{-1}\boldsymbol{Q}_0\boldsymbol{\beta}_0}_{=\boldsymbol{\beta}_0'\boldsymbol{\iota}} + h_0\boldsymbol{Q}_0^{-1}\boldsymbol{\iota}\boldsymbol{\iota}'\boldsymbol{Q}_0^{-1}h_0\boldsymbol{\iota}\boldsymbol{\beta}_0'\boldsymbol{\iota}) \\
&= \boldsymbol{\beta}_0 + h_0\boldsymbol{Q}_0^{-1}\boldsymbol{\iota}\boldsymbol{\beta}_0'\boldsymbol{\iota}\Big(1 - \frac{1}{1 + h_0\boldsymbol{\iota}'\boldsymbol{Q}_0^{-1}\boldsymbol{\iota}} - \frac{h_0\boldsymbol{\iota}'\boldsymbol{Q}_0^{-1}\boldsymbol{\iota}}{1 + h_0\boldsymbol{\iota}'\boldsymbol{Q}_0^{-1}\boldsymbol{\iota}}\Big) = \boldsymbol{\beta}_0.
\end{aligned}$$

In the first line, the explicit form of the inverse of $\boldsymbol{Q}_0 + h_0\boldsymbol{\iota}\boldsymbol{\iota}'$ is obtained by applying the Sherman-Morrison formula.

B: Bayesian estimation of the AR-X(1) model

The results exposed in this appendix are included for ease of reference. They are well known, see e.g., Bauwens, Lubrano, and Richard (1999) for details.

For the regression Equation (8), with the assumption of normality of the error term, the prior (13) and (16), the posterior density of $\boldsymbol{\beta}$ and σ^2 is proportional to

$$(\sigma^2)^{-(T+2)/2}\exp\{-\frac{\hat{s}}{2\sigma^2}\}\exp\{-\frac{1}{2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'\frac{\boldsymbol{Z}'\boldsymbol{Z}}{\sigma^2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\}\exp\{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\boldsymbol{V}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\}, \quad (25)$$

where $\hat{\boldsymbol{\beta}}$ is the OLS estimator $(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y}$, and $\hat{s}$ is the sum of squared OLS residuals.

Because the prior density is not conjugate, the posterior marginal density of $\boldsymbol{\beta}$ is not available analytically. However, the posterior density of $(\boldsymbol{\beta}, \sigma^2)$ can be simulated by applying a Gibbs sampler iterating between $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}^2$. Indeed, the posterior density of $\boldsymbol{\beta}$ conditional on σ^2 is Gaussian:

$$\boldsymbol{\beta}|\sigma^2, \boldsymbol{Y}, \boldsymbol{Z} \sim \mathsf{N}_k(\boldsymbol{\beta}_*, \boldsymbol{V}_*), \quad (26)$$

where

$$\boldsymbol{V}_* = \left(\frac{\boldsymbol{Z}'\boldsymbol{Z}}{\sigma^2} + \boldsymbol{V}_0^{-1}\right)^{-1}, \quad (27)$$

$$\boldsymbol{\beta}_* = \boldsymbol{V}_*\left(\frac{\boldsymbol{Z}'\boldsymbol{Y}}{\sigma^2} + \boldsymbol{V}_0^{-1}\boldsymbol{\beta}_0\right) := \boldsymbol{\beta}_*(\sigma^2). \quad (28)$$

and the complementary conditional density of σ^2 is inverted-gamma:

$$\sigma^2|\boldsymbol{\beta} \sim \mathsf{IG}(T, (\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{\beta})). \quad (29)$$

The Gibbs sampling algorithm to generate S draws $(\boldsymbol{\beta}^{(s)}, (\sigma^2)^{(s)})$, for $s = 1, 2, \ldots, S$, from the posterior of the parameters (after S_0 warming-up draws) is organized as follows:

1. *Choose an initial value* $(\sigma^2)^{(0)}$ *(e.g.* $\hat{s}/(T-k-2)$*).*

2. *Set* $s = 1$.

3. *Draw successively* $\boldsymbol{\beta}^{(s)}$ *from* the normal density (26) where β_* and Q_* are computed with $\sigma^2 = (\sigma^2)^{(s-1)}$, and $(\sigma^2)^{(s)}$ *from* $\mathsf{IG}(T, (\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{\beta}^{(s)})'(\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{\beta}^{(s)}))$.

4. *Set* $s = s + 1$ *and go to step 3 unless* $s > S_0 + S$.

5. *Discard the first* S_0 *values of* $\boldsymbol{\beta}^{(s)}$ *and* $(\sigma^2)^{(s)}$.

The posterior expectation of $\boldsymbol{\beta}$ is approximated by the mean of the S draws $\boldsymbol{\beta}^{(s)}$, or by the mean of the S conditional expectations $\boldsymbol{\beta}_*[(\sigma^2)^{(s)}]$.

C: Explanation of (23)

Using $a_0 = (1-d_0)/(n-1)$, $\boldsymbol{A}_0 = d_0\boldsymbol{I}_n + a_0(\boldsymbol{J}_n - \boldsymbol{I}_n) = \frac{nd_0-1}{n-1}\boldsymbol{I}_n + \frac{1-d_0}{n-1}\boldsymbol{J}_n$. Using $\boldsymbol{J}_n^h = n^{h-1}\boldsymbol{J}_n$ for $h \geq 1$ and denoting $\boldsymbol{I}_n$ as $\boldsymbol{J}_n^0$,

$$\begin{aligned}\boldsymbol{A}_0^h &= \sum_{j=0}^{h} \frac{h!}{j!\,(h-j)!} \left[\left(\frac{nd_0 - 1}{n-1} \right)^{h-j} \left(\frac{1-d_0}{n-1} \right)^{j} \right] \boldsymbol{J}_n^j \\ &= \left(d_0 + \frac{d_0 - 1}{n-1} \right)^{h} \boldsymbol{I}_n + \frac{1}{n} \left[1 - \left(d_0 + \frac{d_0 - 1}{n-1} \right)^{h} \right] \boldsymbol{J}_n\end{aligned}$$

and hence $\boldsymbol{A}_0^h = \left(d_0^h + o\left(n^{-1}\right)\right)\boldsymbol{I}_n + \left(\frac{1-d_0^h}{n} + o\left(n^{-1}\right)\right)\boldsymbol{J}_n$, for $n >> h$, so that the first row is then close to $\left(d_0^h, \frac{1-d_0^h}{n}, ..., \frac{1-d_0^h}{n}\right)'$. The target $\boldsymbol{\beta}_{h,0}$ in (23) is obtained by putting 0 as first element and dividing the last $n-1$ elements by $n-1$ (instead of n) to ensure that the sum of the target is exactly equal to 1.

D: Technical details

Model confidence set

The procedure of Hansen et al. (2011) is applied using the MAFE and MSFE loss functions defined in (24) to perform the hypothesis tests of equal predictive accuracy needed to obtain each model confidence set. These tests are performed at the 25% significance level, so that the resulting MCS is at the confidence level of 75%. The test statistic is the range statistic that requires a bootstrap procedure.

For the application to daily realized volatilities, 10,000 bootstrap samples are used, with a block length of 5 observations to account for potential serial correlation and conditional heteroscedasticity in the losses. For the application to monthly river streamflows, the number of bootstrap samples is 10,000 and the block length is 3.

Cross validation

Table 3 reports the grids of the cross validations performed to choose the values of the tuning parameters that determine the shrinkage of the RAR(1)-X and BAR(1)-X models. The grids are the same for both applications. The cross validations are performed only on the first estimation window. It might be more to the advantage of both methods to renew the cross validation for each new window of estimation, but this would increase the computation time considerably.

Table 3: Grids for the cross validations

RAR(1)-X	d_0	0.2 to 0.55 by steps of 0.025
	λ_d^{-1}	0.01 to 0.05 by steps of 0.01
	λ_a^{-1}	0.01 to 0.05 by steps of 0.01
	λ_S^2	0 to 5,000 by steps of 1,000
BAR(1)-X	d_0	0.2 to 0.55 by by steps of 0.05
	s_d	0.01 to 0.05 by steps of 0.01
	s_a	0.01 to 0.05 by steps of 0.01
	h_0	0 to 5,000 by steps of 1,000

For the sake of illustration, Figures 4 and 5 provide the histograms of the values obtained by the cross validations, for RAR(1)-X and BAR(1)-X and $h = 1$. The ordinates show the number of series, for example d_0 is equal to 0.55 for a little less than 150 series (out of 250) for RAR and a little more than 150 for BAR in the first application. In the second application, the cross validation procedure chooses $d_0 = 0.55$ for about half of the series, and the value 0.2 for about 25 percent in the case of RAR (40 in the case of BAR).

The parameters $1/\lambda_d$ of RAR and s_d in BAR are selected at the lowest values of the grid (0.01 or 0.02) for about two-thirds of the series in the first application. In the second application, $1/\lambda_d$ is selected in equal proportions at the boundaries of the grid range (0.01 and 0.05), whereas s_d is selected mainly at the end of the range. The parameters $1/\lambda_a$ of RAR and s_a in BAR are selected differently between RAR and BAR and between applications 1 and 2.

The additional shrinkage of the sum of the coefficients toward 1 by the parameter λ_S^2 (RAR) or the equivalent parameter h_0 (BAR) is effective for abound 120 series (about 48 percent) in the first application, but for very few series in the second application. The impact of the shrinkage of the sum towards unity is, however, effective through the other constraints. In the first application, the OLS estimated sum ranges from -0.06 to 1.50 (over the 150 series), the mean being 0.92 and the standard deviation 0.23; RAR estimation results in the range (0.66, 1.44), with mean 0.97 and standard deviation 0.11; the BAR range is (0.65, 1.48) with the same mean and standard deviation as RAR.

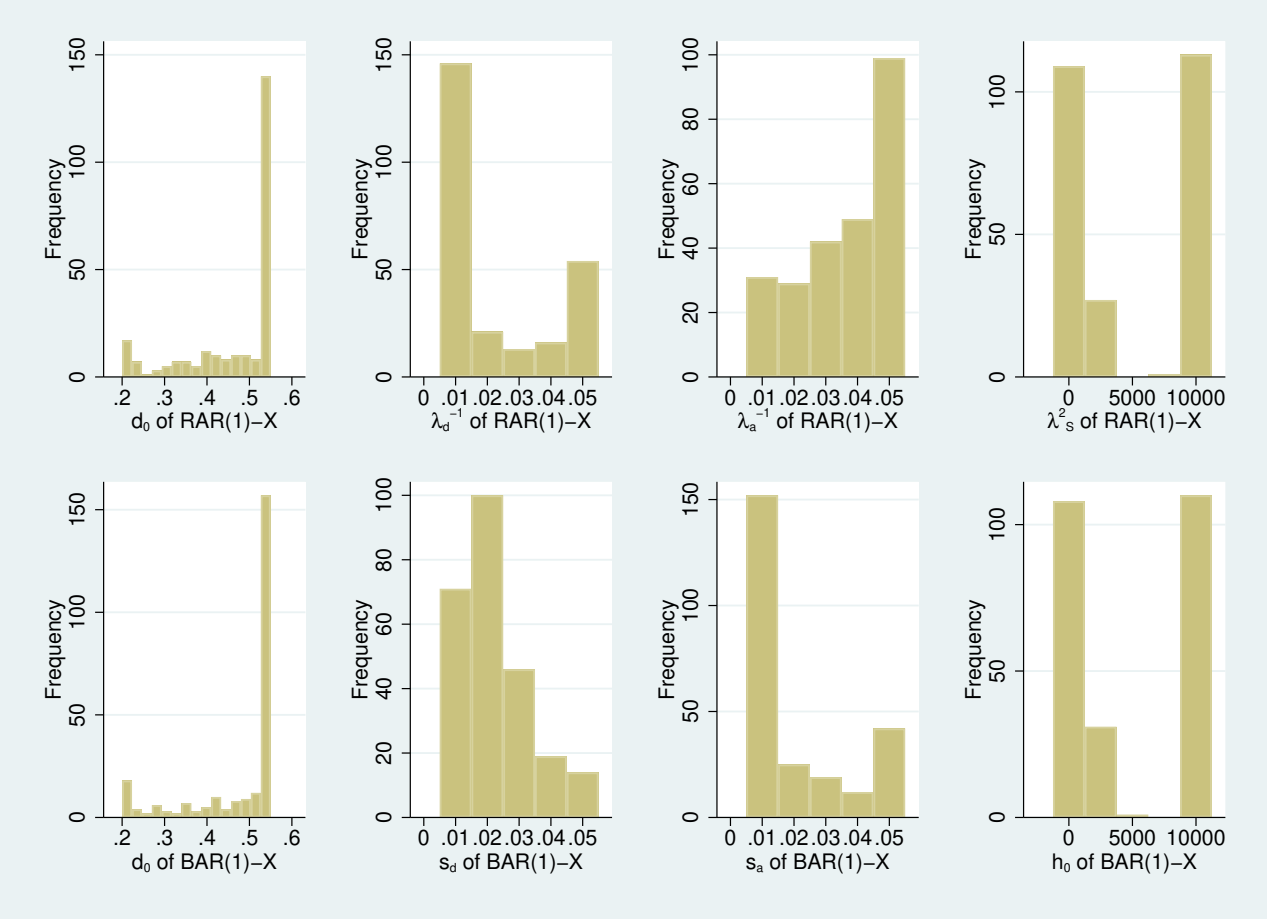

Figure 4: Histogram of the four tuning parameters estimated by cross validation on the first sample of 1,000 observations for the application to realized volatilities .

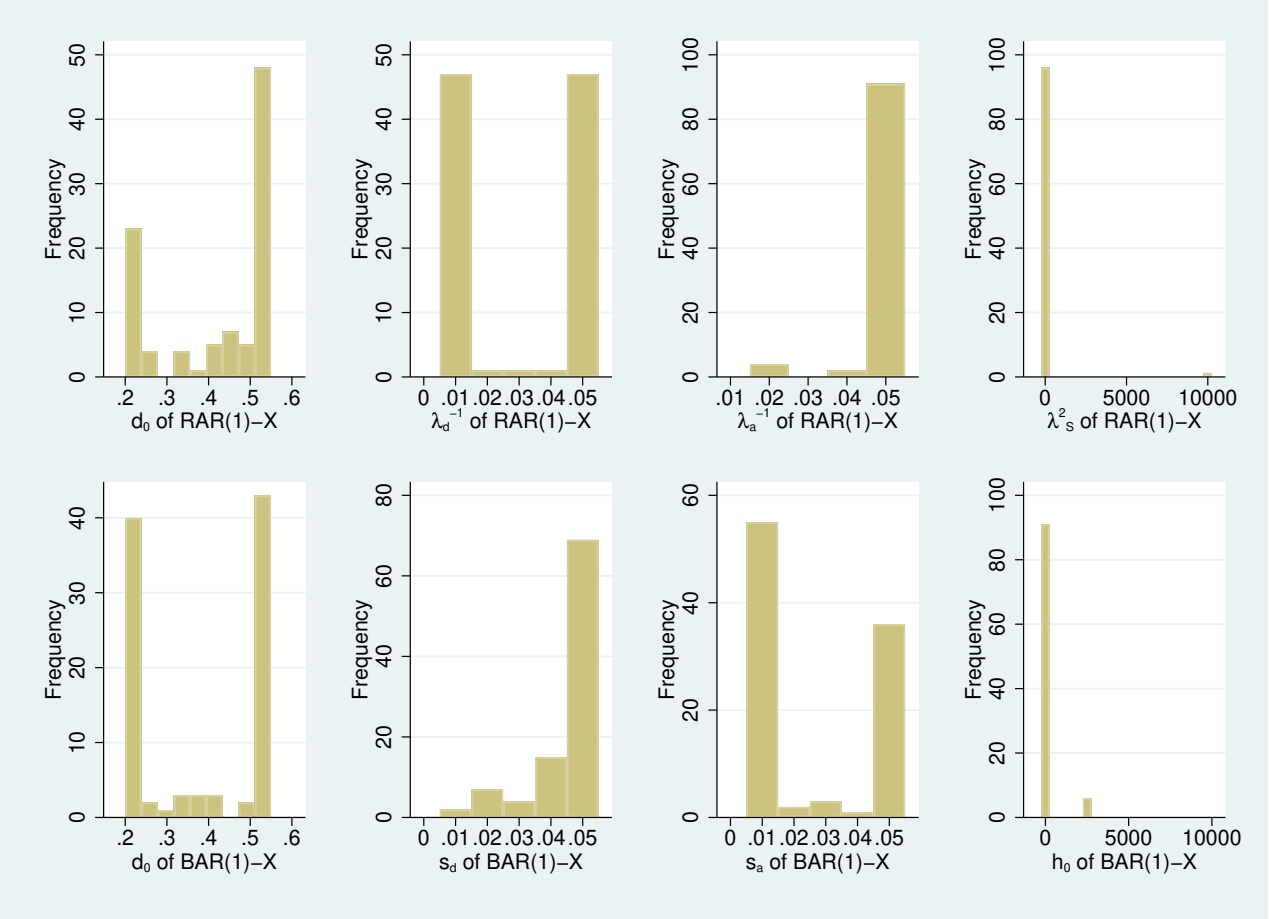

Figure 5: Histograms of the four tuning parameters estimated by cross validation on the first sample of 400 observations for the application to the modified river streamflows.

E: Data source and references for the second application

Short literature review

Ooms and Franses (2001) documented that monthly river flow data displays long memory, in addition to pronounced seasonality, based on simple time series plots and periodic sample autocorrelations. Wang, *et al.* (2005) and Koscielny-Bunde, *et al.* (2006) obtained similar results for daily data for a sample of river basins. It must be noted that long memory is not found in all hydrological datasets, depending on the data considered, the frequency and length of observation (see, for instance, Rao and Bhattacharya, 1999, and Montanari et al., 2000), but as mentioned in the doctoral thesis of Wen Wang (2006) at the Technological University of Delft, ARFIMA models remained at the time the main contenders for forecasting streamflows (though some neural network based techniques may help capturing some nonlinearities). Over the last 15 years, the literature has explored machine learning techniques (artificial neural networks, support vector machines, ...) for forecasting hydrological series and have found mixed evidence depending on the situations. To assess these results Papacharalampous et al. (2019) perform an extensive comparison of 20 prototypical multi-step forecasting models (11 'stochastic', i.e., extensions of ARMA models, and 9 'machine learning' models) over hundreds of simulated and empirical datasets and using 18 accuracy metrics. Their findings are that (*i*) most empirical series exhibit a degree of long memory between 0 and 0.45, with a median close to 0.2 (see their Figure 1), (*ii*) the most accurate 'stochastic' and machine learning techniques perform similarly, and (*iii*) ARFIMA models belong to the class of most accurate 'stochastic' techniques (see their Figure 18).

Data source

The data for the modified river streamflows of the Columbia river basin are available at "https://www.bpa.gov/-/media/Aep/power/historical-streamflow-reports/historic-streamflow-all-monthly-data.zip". The BPA report states that "at certain locations, modified flow values can be negative during instances when the evaporation and/or irrigation adjustments are larger than the calculated inflows or routed flows. In these cases, it is likely that current levels of irrigation require more water than was historically observed". As we work with the logarithm of modified flows, we disregarded the four locations with negative values.

F: Figures using the MSFE loss function

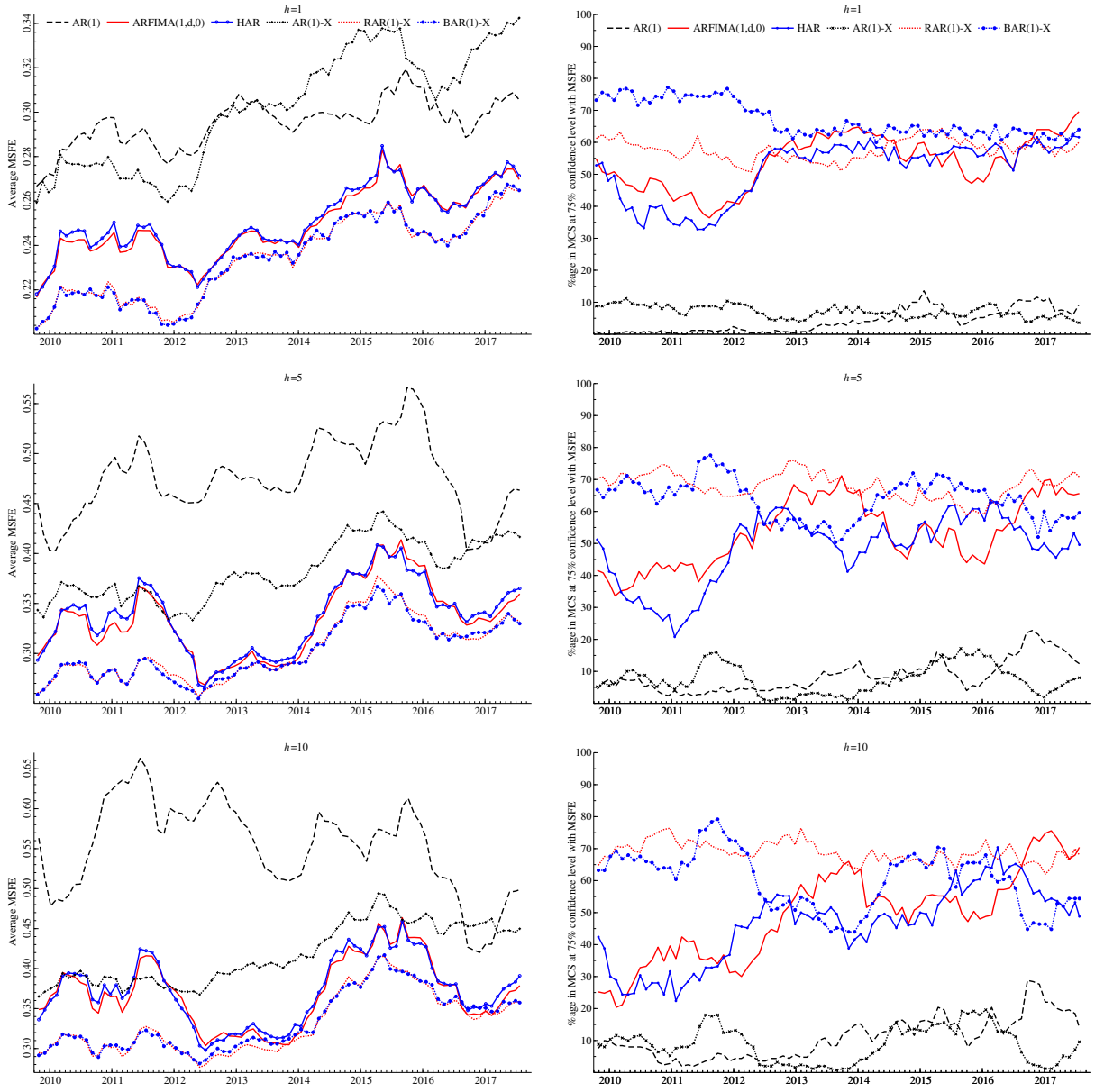

Figure 6: Application to realized volatilities. Left panels: Average (over the 250 series) MSFE loss computed on rolling windows of 250 observations. Right panels: Frequencies (over the 250 series), at each date, at which each model belongs to the MCS (at the 75% confidence level) for the MSFE loss function. The top block is for the forecast horizon $h = 1$, the middel one for $h = 5$, and the bottom one for $h = 10$.

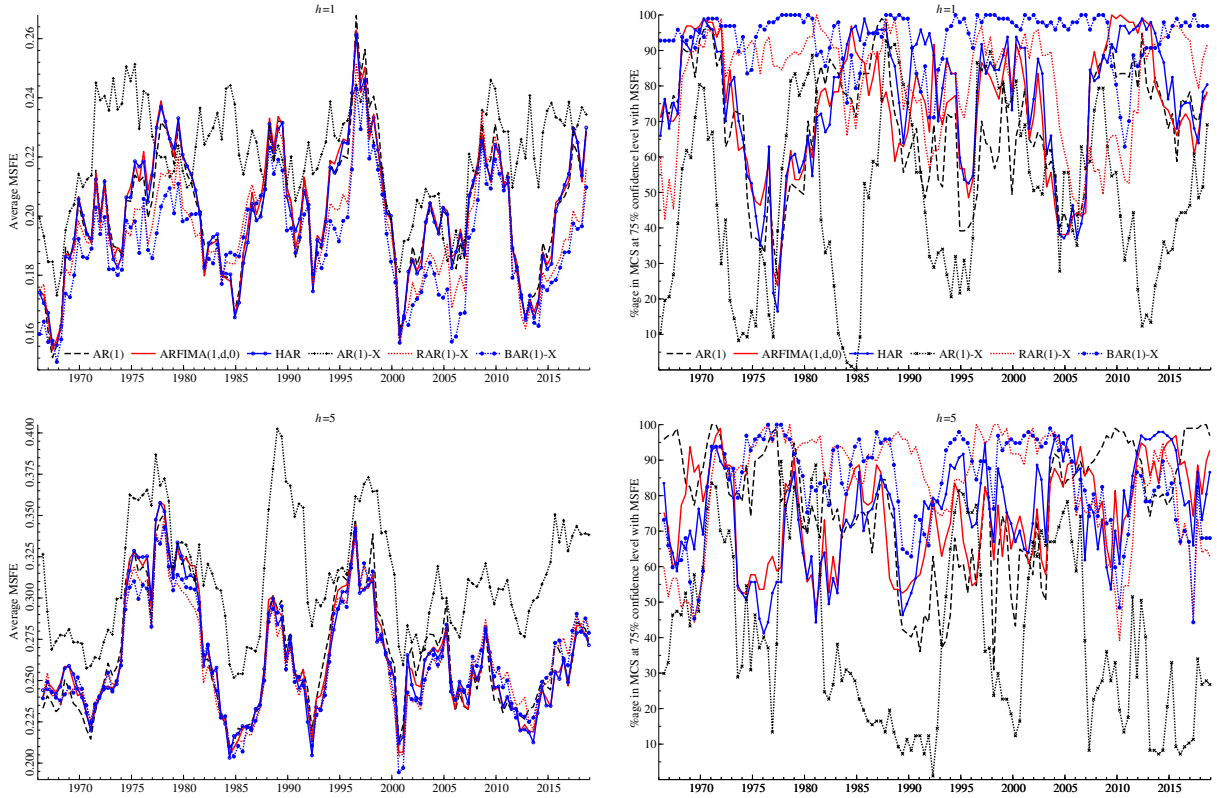

Figure 7: Application to modified river streamflows. Left panels: Average (over the 97 series) MSFE loss computed on rolling windows of 50 observations. Right panels: Frequencies (over the 97 series), at each date, at which each model belongs to the MCS (at the 75% confidence level) for the MSFE loss function. The top block is for the forecast horizon $h = 1$, the bottom one for $h = 5$.

References

Abadir, K. M. and G. Talmain (2002). Aggregation, persistence and volatility in a macro model. *Review of Economic Studies 69*(4), 749–79.

Andersen, T., D. Dobrev, and E. Schaumburg (2012). Jump robust volatility estimation using nearest neighbor truncation. *Journal of Econometrics 169*(1), 75–93.

Andersen, T. G., T. Bollerslev, F. X. Diebold, and H. Ebens (2001). The distribution of realized stock return volatility. *Journal of financial economics 61*(1), 43–76.

Anderson, H. M. and F. Vahid (2007). Forecasting the volatility of Australian stock returns: Do common factors help? *Journal of Business & Economic Statistics 25*(1), 76–90.

Baillie, R., T. Bollerslev, and H. Mikkelsen (1996). Fractionally integrated generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics 74*, 3–30.

Baillie, R. T. (1996). Long memory processes and fractional integration in econometrics. *Journal of Econometrics 73*, 5–59.

Bauwens, L., M. Lubrano, and J.-F. Richard (1999). *Bayesian Inference in Dynamic Econometric Models*. Oxford University Press.

Beran, J. (1992). Statistical methods for data with long-range dependence. *Statistical Science 7*(4), 404–416.

Breidt, F. J., N. Crato, and P. de Lima (1998). The detection and estimation of long memory in stochastic volatility. *Journal of Econometrics 83*(1-2), 325–348.

Chen, X., L. P. Hansen, and M. Carrasco (2010). Nonlinearity and temporal dependence. *Journal of Econometrics 155*(2), 155–169.

Chevillon, G., A. Hecq, and S. Laurent (2018). Generating univariate fractional integration within a large VAR(1). *Journal of Econometrics 204*(1), 54–65.

Chevillon, G. and D. F. Hendry (2005). Non-parametric direct multi-step estimation for forecasting economic processes. *International Journal of Forecasting 21*(2), 201–218.

Chevillon, G. and S. Mavroeidis (2017). Learning can generate long memory. *Journal of Econometrics 198*, 1–9.

Chevillon, G. and S. Mavroeidis (2018). Perpetual learning and apparent long memory. *Journal of Economic Dynamics and Control 90*, 343–365.

Comte, F. and E. Renault (1998). Long memory in continuous-time stochastic volatility models. *Mathematical Finance 8*, 291–323.

Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics 7*(2), 174–196.

Cox, D. R. and M. W. H. Townsend (1947). The use of the correlogram in measuring yarn irregularities. *Proceedings of the Royal Society of Edinburgh, Section A 63*, 290–311.

Cubadda, G., A. Hecq, and A. Riccardo (2019). Forecasting realized volatility measures with multivariate and univariate models: The case of the us banking sector. In J. Chevallier, S. Goutte, G. D., S. Saglio, and B. Sanhaji (Eds.), *Financial Mathematics, Volatility and Covariance Modelling*, Volume 2, Chapter 11, pp. 791–897. Routledge.

Dakhlalla, A., S. Hughes, A. M. E. Pytlak, T. R. Roth, and R. van der Zweep (2020). 2020 level modified streamflow: 1928-2018. Historical streamflow reports, Bonneville Power Administration, Department of Energy.

Davidson, J. and P. Sibbertsen (2005). Generating schemes for long memory processes: regimes, aggregation and linearity. *Journal of Econometrics 128*(2), 253–82.

Diebold, F. X. and A. Inoue (2001). Long memory and regime switching. *Journal of Econometrics 105*(1), 131–159.

Diebold, F. X. and K. Yılmaz (2009). Measuring financial asset return and volatility spillovers, with application to global equity markets. *The Economic Journal 119*, 158–171.

Diebold, F. X. and K. Yılmaz (2014). On the network topology of variance decompositions: Measuring the connectedness of financial firms. *Journal of Econometrics 182*, 119–134.

Doan, T., R. Litterman, and C. Sims (1984). Forecasting and conditional projection under realistic prior distributions. *Econometric Reviews 3*, 1–100.

Giannone, D., M. Lenza, and G. Primiceri (2019). Priors for the long run. *Journal of the American Statistical Association 114*, 565–580.

Giraitis, L., P. M. Robinson, and D. Surgailis (2000). A model for long memory conditional heteroscedasticity. *Annals of Applied Probability*, 1002–1024.

Gourieroux, C. and J. Jasiak (2001). Memory and infrequent breaks. *Economics Letters 70*, 29–41.

Granger, C. W. J. (1966). The typical spectral shape of an economic variable. *Econometrica 34*, 150–161.

Granger, C. W. J. (1980). Long memory relationships and the aggregation of dynamic models. *Journal of Econometrics 14*(2), 227–238.

Granger, C. W. J. and R. Joyeux (1980). An introduction to long-memory time series models and fractional differencing. *Journal of time series analysis 1*(1), 15–29.

Haldrup, N. and J. E. Vera Valdés (2017). Long memory, fractional integration, and cross-sectional aggregation. *Journal of Econometrics 199*(1), 1–11.

Hansen, P., A. Lunde, and J. Nason (2011). The model confidence set. *Econometrica 79*, 453–497.

Hualde, J. and M. Ø. Nielsen (2022). Fractional integration and cointegration. In *Oxford Research Encyclopedia of Economics and Finance*, pp. forthcoming. Basingstoke.

Hurst, H. E. (1951). Long-term storage capacity of reservoirs. *Transactions of the American society of civil engineers 116*(1), 770–799.

Hurvich, C. M., E. Moulines, and P. Soulier (2005). Estimating long memory in volatility. *Econometrica 73*(4), 1283–1328.

Johansen, S. and M. Ø. Nielsen (2012). Likelihood inference for a fractionally cointegrated vector autoregressive model. *Econometrica 80*(6), 2667–2732.

Karlsson, S. (2013). Forecasting with Bayesian vector autoregressions. In G. Elliot and A. Timmermann (Eds.), *Handbook of Economic Forecasting*, Volume 2, Chapter 15, pp. 791–897. Elsevier.

Leipus, R. and D. Surgailis (2003). Random coefficient autoregression, regime switching and long memory. *Advances in Applied Probability 35*(3), 737–754.

Lieberman, O. and P. C. Phillips (2008). Refined inference on long memory in realized volatility. *Econometric reviews 27*(1-3), 254–267.

Marmol, F. and C. Velasco (2004). Consistent testing of cointegrating relationships. *Econometrica 72*(6), 1809–1844.

Miller, J. I. and J. Y. Park (2010). Nonlinearity, nonstationarity, and thick tails: How they interact to generate persistence in memory. *Journal of Econometrics 155*(1), 83 – 89.

Montanari, A., R. Rosso, and M. S. Taqqu (1997). Fractionally differenced arima models applied to hydrologic time series: Identification, estimation, and simulation. *Water resources research 33*(5), 1035–1044.

Nelson, C. R. and C. R. Plosser (1982). Trends and random walks in macroeconomic time series: some evidence and implications. *Journal of Monetary Economics 10*(2), 139–162.

Ooms, M. and P. H. Franses (2001). A seasonal periodic long memory model for monthly river flows. *Environmental Modelling & Software 16*(6), 559–569. Economics and Environmental Modelling.

Papacharalampous, G., H. Tyralis, and D. Koutsoyiannis (2019). Comparison of stochastic and machine learning methods for multi-step ahead forecasting of hydrological processes. *Stochastic Environmental Research and Risk Assessment 33*, 481—514.

Perron, P. and Z. Qu (2010). Long-memory and level shifts in the volatility of stock market return indices. *Journal of Business & Economic Statistics 28*(2), 275–290.

Rao, A. R. and D. Bhattacharya (1999). Hypothesis testing for long-term memory in hydrologic series. *Journal of Hydrology 216*(3-4), 183–196.

Robinson, P. M. (2003). Long memory time series. In P. M. Robinson (Ed.), *Time Series With Long Memory*, pp. 1–48. Oxford: Oxford University Press.

Robinson, P. M. and J. Hualde (2003). Cointegration in fractional systems with unknown integration orders. *Econometrica 71*(6), 1727–1766.

Robinson, P. M. and P. Zaffaroni (1998). Nonlinear time series with long memory: a model for stochastic volatility. *Journal of Statistical Planning and Inference 68*(2), 359–371.

Schennach, S. M. (2018). Long memory via networking. *Econometrica 86*(6), 2221–2248.

Schorfheide, F. (2005). VAR forecasting under misspecification. *Journal of Econometrics 128*(1), 99–136.

Shi, S. and J. Yu (2021). Volatility puzzle, forthcoming in *Management Science*.

Smith, H. F. (1938). An empirical law describing heterogeneity in the yields of agricultural crops. *The Journal of Agricultural Science 28*(01), 1–23.

Stein, M. (2005). Space-time covariance functions. *Journal of the American Statistical Association 100*, 310–321.

Wang, W., P. van Gelder, and J. Vrijling (2005). Long-memory in streamflow processes of the Yellow river. In *International conference on water economics, statistics and finance, Rethymno, Crete*, pp. 481–490. University of Crete.