Reconstruction of missing daily streamflow data using dynamic regression models

Patricia Tencaliec[1], Anne-Catherine Favre[2], Clémentine Prieur[3] & Thibault Mathevet[4]

[1] *Univ. Grenoble Alpes, CNRS, LJK, Inria project/team AIRSEA, F-38000 Grenoble, France, patricia.tencaliec@imag.fr*

[2] *Univ. Grenoble Alpes, CNRS, IRD, LTHE, F-38000 Grenoble, France, Anne-Catherine.Favre-Pugin@ense3.grenoble-inp.fr*

[3] *Univ. Grenoble Alpes, CNRS, LJK, Inria project/team AIRSEA, F-38000 Grenoble, France, clementine.prieur@imag.fr*

[4] *Electricité de France (EDF), Division Technique Générale(DTG), Grenoble, France, thibault.mathevet@edf.fr*

Résumé. Pour de nombreuses applications de recherche ou de recherche opérationnelle, telles que par exemple la gestion des ressources en eau, la prévision d'événements extrêmes tels que les inondations ou les sécheresses, la prévision des débits et l'analyse de la variabilité du climat, il est nécessaire de travailler avec des séries chronologiques fiables. Par ailleurs, pour l'étude des phénomènes extrêmes, disposer de longues séries complètes est un plus indéniable. Nous présentons dans ce travail une technique efficace pour la reconstruction des données manquantes de débits journaliers, dans un contexte où les données de débits journaliers sont les seules à disposition. La méthode proposée repose essentiellement sur la combinaison de modèles de régression linéaire multiple à des modèles à moyenne mobile autorégressifs intégrés (ARIMA). Plus spécifiquement, elle fait appel aux modèles de type régression dynamique. Plus précisément, l'approche proposée exploite les corrélations linéaires entre stations voisines, puis ajuste les résidus par un processus ARIMA. Cette approche est flexible est appliquée aux données de débits journaliers du bassin versant de la Durance (France). Par ailleurs, une étude à base de simulations est présentée, afin d'illustrer l'efficacité de l'approche proposée pour la reconstructione données manquantes.

Mots-clés. imputation, débit, ARIMA, modèle de régression dynamique, bassin versant de la Durance

Abstract. Numerous research and operational applications, such as water resources management, extreme flood or drought predetermination, streamflow forecast and climate variability analysis, require reliable time series. Since extreme events are seldom by definition, long and continuous time series are necessary. In this work we introduce an effective technique for reconstructing missing daily discharge data when one has access to only daily streamflow data. The proposed procedure uses a combination of regression and autoregressive integrated moving average models (ARIMA) called dynamic regression model. It uses the linear relationship between neighbor and correlated stations and then adjusts the residual term by fitting an ARIMA structure. This technique has a very

general formulation, making possible the imputation for a wide range of situations. Application of the model to daily streamflow data for the Durance river watershed (France) showed that the model yields reliable estimates for the missing data in the time series. These results were further confirmed by simulation studies.

Keywords. imputation, streamflow, ARIMA, dynamic regression models, Durance watershed

1 Introduction

The reconstruction of missing streamflow data is a problem studied from decades ago and, even nowadays, it continues to be a challenge. There are several methods reported in the literature, among these, Wallis et al. (1991), that discuss infilling approaches for daily data using data from the nearby station(s), Woodhouse et al. (2006) that recommend the use of regression analysis for reconstructing the missing data, or more recent studies present approaches that involve artificial neural networks as detailed in Coulibaly and Baldwin (2005).

In this study we use the dynamic regression models (DRMs) to estimate the missing streamflow data. The DRM estimates an output variable based on one or multiple input variables and also adjusts the correlation from the remainder part (residuals) by fitting an autoregressive integrated moving average (ARIMA) structure.

This approach was used before, among others, by Greenhouse et al.(1987) to fit biological rhythm data, Miaou (1990) to estimate the water demand in some states from USA, or, more recently, by Bercu and Proia (2013) to forecast energy consumption.

2 Statistical modeling: theory and methodology

A dynamic regression model states how a response variable (Y_t) is related to present and past values of one or more explanatory variables ($X_{t,1}, ..., X_{t,l}$). Besides this, it allows for the residual term of the regression (i.e., the difference between observations and the estimates of the regression part of the model) to be modeled with a seasonal autoregressive integrated moving average (SARIMA) model.

A SARIMA model is an extension of the well-known ARIMA model that addresses seasonality. Therefore, apart from the relationships between observations of successive periods, SARIMA incorporates the relationships between observations at certain period distance, for example a week, a quarter, etc. (seasonal part). A short notation for this model is SARIMA$(p, d, q)(P, D, Q)_s$.

The general dynamic regression model formulation, in terms of the backshift operator B (defined as $B^i Y_t = Y_{t-i}$), with l explanatory variables and a SARIMA$(p, d, q)(P, D, Q)_s$ model for the residuals, is

$$Y_t = \beta_0 + \alpha_1(B)X_{t,1} + ... + \alpha_l(B)X_{t,l} + Z_t \tag{1}$$

where the residual term Z_t is expressed as

$$\phi(B)\phi_s(B^s)\nabla^d\nabla_s^D Z_t = \theta(B)\theta_s(B^s)e_t \tag{2}$$

The polynomials $\alpha_i(B)$ in (1) represent how Y_t reacts over a time period to a change in $X_{t,i}$. They are called so far *transfer functions* and are defined as $\alpha_i(B) = \frac{\omega_i(B)}{\delta_i(B)}B^{b_i}$. Thus, the transfer functions have three orders, m_i and r_i (orders of the polynomial $\omega_i(B)$ and $\delta_i(B)$), and b_i, that must be set.

In the formulation (2) we have the polynomials of the SARIMA model for the non-seasonal part ($\phi(B)$, $\theta(B)$ with orders p and q) and seasonal part ($\phi_s(B^s)$,$\theta_s(B^s)$ with orders P and Q), representing the autoregressive (polynomials ϕ) and moving average components (polynomials θ). The operators ∇^d and ∇_s^D are used in case of non-stationary series, and they represent the differencing of order d for the non-seasonal part, respectively the differencing of order D for the seasonal part with s time units per season. Thus, SARIMA model has seven orders that must be set (p, d, q, P, D, Q, s).

Readers can find an extended theory presentation of this model in Pankratz (1991) or Box and Jenkins (1976).

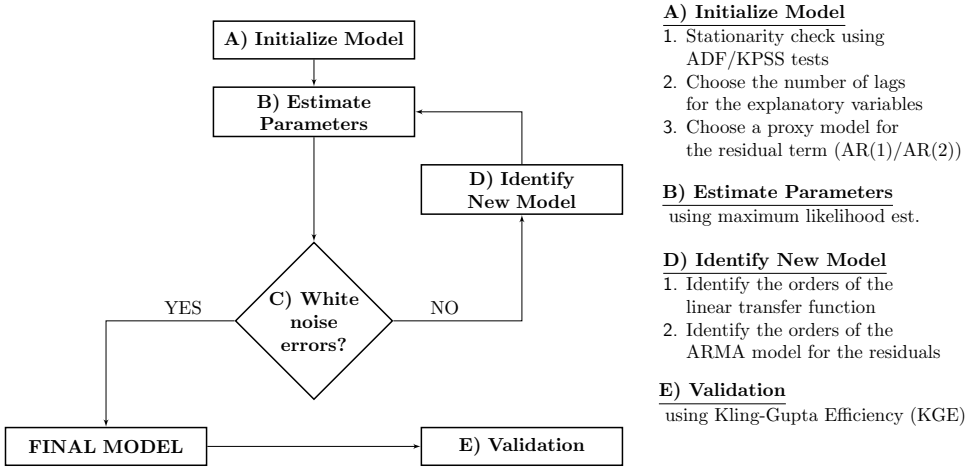

Figure 1: Schematic representation of the estimation methodology

We illustrate the procedure for the model estimation and validation schematically, see Figure 1. The first step (step A) is to choose a proxy model for both parts: a multiple linear regression with a large enough number of lags for each explanatory variable, and a low-order AR(1)/AR(2) model for the residuals. The estimates of the parameters and the errors of the initial-proxy model are then analyzed and, if necessary, a new model is identified and estimated again, until the errors of the selected model are a white noise process.

The procedure and rules for identifying the new model (step D) requires to find first the orders of both the linear transfer functions and (S)ARIMA, procedure that can be find in Pankratz (1991) and Box and Jenkins (1976). Shortly, for the (S)ARIMA models,

the order identification is done by analyzing the sample autocorrelation and partial autocorrelation coefficients, while for the transfer function one must examine the pattern of the coefficients for each explanatory variable.

3 Application on the Durance watershed

The method discussed previously is now applied to eight stations of the Durance watershed (France) for the daily flow measurements of 107 years (1904-2010). For the estimation of the parameters, we used the longest part of the dataset that has no missing values, namely a sequence of 22 years (1980-2001). The validation of the models was handled on three different test sets each containing four years of daily flow data, that is: 1918-1921, 1931-1934 and 2002-2005.

The performance of the models was study by comparing our results with a simpler, but common method of reconstructing missing meteorological data, the nearest-neighbors technique (NN), and also with a more complex one, a meteorological data reconstruction called ANATEM, see Kuentz et al. (2015).

	S1	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**
all data & warm season	S3	S1,S4	S1,S4	S1,S5	S4,S7	S7,S8	S5,S6,S8	S6,S7
cold season	S3	S1,S4	S1,S4	S3	S7	S7,S8	S5,S6,S8	S6,S7

Table 1: The neighbors of each station for all-year data (Jan-Dec), cold season data (Sep-Feb) and warm season data (Mar-Aug).

Model Identification. After an extended exploratory analysis and preprocessing of the eight stations data (i.e.,monthly means, correlation and cluster analysis, or multicolinearity and stationarity examination), we set the neighbors for each station, so the explanatory variable in the regression part of our model (see Table 1). As the relationships might slightly change when different subsets were analyzed (all-year data, only the cold season data, or only the warm season data), we studied all the situations.

Applying the model identification procedure described in the previous section, we found for each station a range of four models that we further validated. These models are different in the number of lags considered for the explanatory variables X_t (i.e., 0- or 1-lag) and in the number of seasons considered, i.e., one season (all the data) or two seasons (cold and warm). Regarding the SARIMA model-part (models for the residuals), the results showed that they are the same regardless how many seasons or lags are considered, excepting two stations that, in fact, presented also a weak weekly seasonality.

Model validation and performance evaluation. The results reveal that, except some isolated cases (3 out of 24 for NN and 3 out of 24 for ANATEM-RR), our best-model (best out of the four we validated) performs better in each case. An important aspect that must be highlighted is that with DRM the efficiency of the models (KGE) is never lower than 0.72 (1 meaning perfect accuracy), while NN and ANATEM, due to lack of robustness, reduce up to a level of 0.41 and 0.22, respectively.

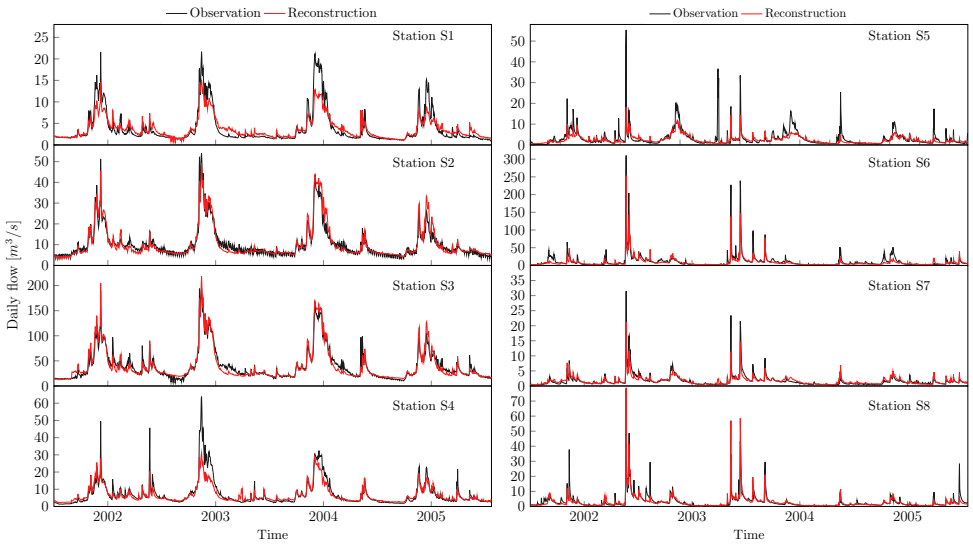

Figure 2: Daily flow estimations vs. observations for period 2002-2005 of the Durance

An illustration of the estimations vs. observations for period 2002-2005 is shown in Figure 2. We have similar representations for the other two periods.

To further test our models, we study also the case when all or some of the explanatory variables for one station are missing too (situation often met in our dataset). Therefore, in order to be able to apply the estimated models, we use for the missing covariates the weighted values from the correlated-neighbor stations. When all the covariates are missing, we use the daily mean (mean of the non-missing values for a certain day for that stations).

The results in this case show that we slightly decrease in performance, but, overall, the KGE is still above 0.5.

These results were also checked by simulation study. The complete-covariates model on simulated data shows that we have a very good performance (average KGE above 0.96), and for missing-covariates model we decrease in performance, but the average KGE remains, mainly, above 0.5.

Finally, the reconstructed series for the eight stations can be seen in Figure 3. The reconstructions show once more that in case of an infilling using the complete-covariates model (all covariates are present) the estimations are extremely good (see stations S1,S2,S3,S4). They slightly decrease in performance when we deal with missing explanatory variables in the model, see the case of the stations S5,S6,S7,S8.

4 Conclusion

In this study we present a way of reconstructing daily streamflow data by using dynamic regression. It is an accessible approach and it can handle even large amount of observations in a short run-time period. Apart from this, our study was performed on a large watershed

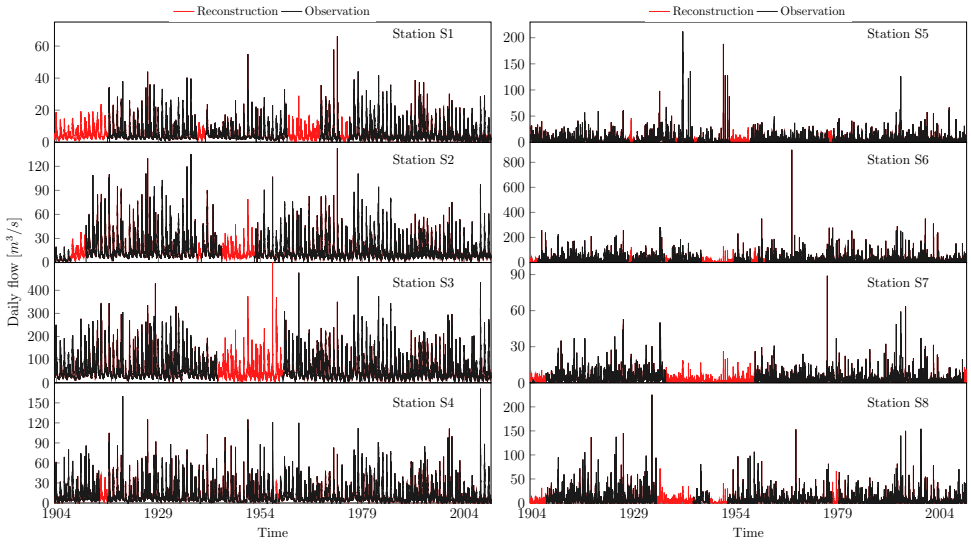

Figure 3: Daily flow Reconstructed series of the eight stations of the Durance watershed

characterized by several hydrological regimes and various data quality issues, so it brings a solid and complex analysis.

Bibliography

[1] Bercu,S., and Proïa,F. (2013) A SARIMAX coupled modelling applied to individual load curves intraday forecasting, *Journal of Applied Statistics*, 40(6), 1333–1348
[2] Box,G.E.P., and Jenkins,G.M. (1976) *Time Series Analysis: Forecasting and Control*, Holden-Day
[3] Coulibaly,P., and Baldwin,C.K. (2005) Nonstationary hydrological time series forecasting using nonlinear dynamic methods, *Journal of Hydrology*, 307(1-4), 164–174
[4] Greenhouse,J.B., Kass,R.E., and Tsay,R.S. (1987) Fitting nonlinear models with ARMA errors to biological rhythm data, *Statistics in Medicine*, 6(2), 167–183
[5] Kuentz,A., Mathevet,T., Gailhard,J., and Hingray,B. (2015) Building long-term and high spatio-temporal resolution precipitation and air temperature reanalyses by mixing local observations and global atmospheric reanalyses: the ANATEM method, *Hydrology and Earth System Sciences Discussions*, 12(1), 311–361
[6] Miaou, S.P. (1990) A stepwise time series regression procedure for water demand model identification, *Water Resources Research*, 26(9), 1887–1897
[7] Pankratz,A. (1991) *Forecasting with Dynamic Regression Models*, Wiley
[8] Wallis,J.R., Lettenmaier,D.P., and Wood,E.F. (1991) A daily hydroclimatological data set for the continental United States, *Water Resources Research*, 27(7), 1657–1663
[9] Woodhouse,C.A., Gray,S.T., and Meko,D.M. (2006) Updated streamflow reconstructions for the Upper Colorado River Basin, *Water Resources Research*, 42(5), 1–16