Journal of Hydrology 556 (2018) 262–278

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Research papers

Conditioned empirical orthogonal functions for interpolation of runoff time series along rivers: Application to reconstruction of missing monthly records

Lingqi Li [a], Lars Gottschalk [b], Irina Krasovskaia [c], Lihua Xiong [a,*]

[a] State Key Laboratory of Water Resources and Hydropower Engineering Science, Wuhan University, Wuhan, PR China
[b] Department of Geosciences, University of Oslo, P.O. Box 1022 Blindern, N-0315 Oslo, Norway
[c] IGC, Lergravsvägen 33, SE-26352 Lerberget, Sweden

ARTICLE INFO

Article history:
Received 11 March 2017
Received in revised form 25 September 2017
Accepted 8 November 2017
Available online 11 November 2017
This manuscript was handled by A. Bardossy, Editor-in-Chief, with the assistance of Purna Chandra Nayak, Associate Editor

Keywords:
Empirical orthogonal functions
Amplitude functions
Reconstruction
Gap filling
Runoff time series

ABSTRACT

Reconstruction of missing runoff data is of important significance to solve contradictions between the common situation of gaps and the fundamental necessity of complete time series for reliable hydrological research. The conventional empirical orthogonal functions (EOF) approach has been documented to be useful for interpolating hydrological series based upon spatiotemporal decomposition of runoff variation patterns, without additional measurements (e.g., precipitation, land cover). This study develops a new EOF-based approach (abbreviated as CEOF) that conditions EOF expansion on the oscillations at outlet (or any other reference station) of a target basin and creates a set of residual series by removing the dependence on this reference series, in order to redefine the amplitude functions (components). This development allows a transparent hydrological interpretation of the dimensionless components and thereby strengthens their capacities to explain various runoff regimes in a basin. The two approaches are demonstrated on an application of discharge observations from the Ganjiang basin, China. Two alternatives for determining amplitude functions based on centred and standardised series, respectively, are tested. The convergence in the reconstruction of observations at different sites as a function of the number of components and its relation to the characteristics of the site are analysed. Results indicate that the CEOF approach offers an efficient way to restore runoff records with only one to four components; it shows more superiority in nested large basins than at headwater sites and often performs better than the EOF approach when using standardised series, especially in improving infilling accuracy for low flows. Comparisons against other interpolation methods (i.e., nearest neighbour, linear regression, inverse distance weighting) further confirm the advantage of the EOF-based approaches in avoiding spatial and temporal inconsistencies in estimated series.

1. Introduction

Availability of complete runoff time series is essential for hydrological studies on such as hydraulic infrastructures design, water resources management, and flood or drought forecasting. Unfortunately, there usually exists such a predicament that data are poorly recorded due to accidental or human-induced causes. The loss of information resulting from the adverse impact of scarce records threatens the reliability of hydrological studies, which consequently poses great demands on reconstruction of missing data (Gottschalk et al., 2015).

Problem of interpolation for gaps in incomplete time series has long been an interesting topic in hydrologic, environmental, and other geoscientific studies. Several methods are proposed and applied successfully within their own applicable scopes, such as the auto-regression model and its various derivatives (Tencaliec et al., 2015; Wang et al., 2015), multiple linear regression linking with simultaneous covariates (precipitation, soil moisture, land cover, etc.) (Daly et al., 1994), weighting average method based on spatial proximity between observation and prediction locations (Arnell, 1995; Kurtzman et al., 2009), spectral analysis (Kondrashov and Ghil, 2006; Mariethoz et al., 2012), Bayesian inference (Tingley and Huybers, 2010), copula modelling (Bárdossy and Pegram, 2014), and the classical methods of objective analysis like Gandin interpolation (Gandin, 1965) and empirical orthogonal functions (EOF) approach (Gottschalk et al., 2015).

* Corresponding author.
E-mail address: xionglh@whu.edu.cn (L. Xiong).

https://doi.org/10.1016/j.jhydrol.2017.11.014

In addition to the aforementioned methods, interpolation of missing data has also been achieved by using process-based approaches, e.g., running a rainfall-runoff model (Hisdal and Tveito, 1993; Wagner et al., 2012), or by incorporating artificial intelligence like artificial neural network (Coulibaly and Evora, 2007; Kim and Pachepsky, 2010) or fuzzy logic theory (Abebe et al., 2000). It has been widely acknowledged that improvements for the accuracy of reconstructed time series for a large basin are still challenging tasks in consideration of the inherent deficiencies in some existing approaches, such as computational complexity, a certain level of subjectivity in calibration and validation of models, or absence of prerequisite for other covariate measurements (Pappas et al., 2014; Tencaliec et al., 2015). Selection of a suitable infilling method often depends on the type of interpolated variables in different disciplines, spatial and temporal distribution features of gaps, and application conditions of interpolation formulation (Gottschalk et al., 2011).

In this paper, reconstruction of missing runoff series is investigated with an emphasis on the EOF approach. In different publications, this approach is also perceived as principal component analysis, factor analysis, and Karhunen-Loève expansion in stochastic approach (Gottschalk et al., 2015). It allows an analysis of spatiotemporal variation by characterising the variance-covariance/correlation structure of runoff series (seen as a random process that evolves in space and time) over a common period and decomposing observed series into a linear combination of orthogonal patterns and uncorrelated components (amplitude functions) in spatial and temporal domain. The determined amplitude functions, invariant across sites in a target region, can be then applied for a large-scale interpolation of incomplete runoff series, at any site in this region, for not only intermittent time slices but also successively long-term time periods.

The EOF approach, being originally popular in the scientific progress of meteorology and climatology, has increasingly received attention in hydrology as a useful tool for, e.g., dimensionality reduction, pattern recognition of hydrological characteristics, classification of runoff pattern, description of runoff characteristics (Gottschalk, 1985; Krasovskaia and Gottschalk, 1995; Krasovskaia et al., 1999, 2003; Johnston and Shmagin, 2008; Ionita et al., 2014; Li et al., 2017), and more importantly, as a powerful stochastic interpolation scheme for "mapping" of river runoff series across space and time (Creutin and Obled, 1982; Obled and Creutin, 1987; Hisdal and Tveito, 1992, 1993; Gottschalk, 1993; Krasovskaia and Gottschalk, 1995; Sauquet et al., 2000, 2008; Beckers and Rixen, 2003; Henn et al., 2013; Gottschalk et al., 2015; Obled and Braud, 1989). For example, Hisdal and Tveito (1993) compared the EOF approach with linear regression and hydrological model for reconstructing daily runoff series from nine basins in southern Norway and concluded that the EOF approach yielded results as good as the other two. Beckers and Rixen (2003) interpolated the incomplete oceanographic datasets based on the ordinary EOF approach. Henn et al. (2013) studied the EOF approach for filling in gaps in hourly near-surface air temperature data and found that the approach was sensitive to the vertical separation of stations and spatial correlation between them. Gottschalk et al. (2015) applied the ordinary EOF approach to interpolate the monthly runoff in the Upper Magdalena River, Colombia for gauging stations with missing records and for ungauged sites. They pointed out that the moderately good results in some mountain headwater stations could be ascribed to the relatively weaker representation of the amplitude functions at these sites where runoff series might tend to present irregular variability. Since the amplitude functions are rarely explained with a clear hydrological meaning, investigation on how to enhance their physical significance becomes necessary to improve the capacity for accounting for various runoff regimes in a basin.

With the above background, the main aim of this paper is to develop the new algorithm based on the EOF approach, namely, the conditioned EOF (CEOF) approach, to improve the accuracy of reconstruction of missing data. The mathematical treatment of this algorithm is presented in Section 3, following a brief presentation of the theoretical background of the conventional EOF approach in Section 2. Section 4 describes how to reconstruct complete time series using the EOF and CEOF approaches and further employs other five interpolation schemes for comparisons. All the approaches are evaluated through an application to monthly discharge observations from the 28 gauging stations in the Gan River (Ganjiang), China in Section 5. We summarise the study in Section 6 with the discussion and conclusions.

2. Background for empirical orthogonal functions

The theoretical background of the conventional EOF approach is briefly recalled, supposing an application to river discharge observations. Readers interested in more details can refer to Gottschalk et al. (2015). Define the drainage area down to a gauging outlet station as the total domain Ω. The discharge series at the outlet is denoted by $Q(t,\Omega)$. The denested (non-overlapping) areas $\omega_i;\ i = 1,\ldots,M$ related to M gauging sites upstream of this outlet station are extracted using $\omega_i = A_i - \sum_s \omega_s$ so that $\sum_{i=1}^{M}\omega_i = \Omega$, where s denotes all sites upstream of site i and A_i is the drainage area of site i. When i is a headwater site (without any upstream site), ω_i is exactly equal to A_i. The discharge $Q(t,\omega_i)$ for a denested drainage area ω_i is defined as the net discharge after subtracting the inflows from upstream areas (except for headwater sites), i.e., $Q(t,\omega_i) = Q(t,A_i) - \sum_s Q(t,A_s)$, which can be written as the integral of flows over space $\mathbf{u}$ across this area (smooth variations are assumed so that the integration can be separated).

$$\begin{aligned} Q(t,\omega_i) &= \int\int_{\mathbf{u}\in\omega_i} Q(t,\mathbf{u})d\mathbf{u} = m_Q(\omega_i) + \int\int_{\mathbf{u}\in\omega_i} X(t,\mathbf{u})d\mathbf{u} \\ &= m_Q(\omega_i) + X(t,\omega_i) \end{aligned} \quad (1)$$

where $Q(t,\omega_i)$ is expressed as the sum of long-term mean value of the discharge $m_Q(\omega_i)$ for a denested area ω_i and the fluctuations around this mean, namely the centred discharge series $X(t,\omega_i)$. In terms of the EOF expansion, this centred discharge can be linearly decomposed into double orthogonal series of spatial and temporal functions:

$$X(t,\omega_i) = \sum_{k=1}^{M}\psi_k(t)\beta_k(\omega_i) \quad (2)$$

where the temporal functions $\psi_k(t);\ k = 1,\ldots,M$ are known as principal components representing time series that are not directly linked to any specific points of the domain Ω. The spatial functions $\beta_k(\omega_i)$, i.e., empirical orthogonal functions, are the weight coefficients of the kth amplitude function $\psi_k(t)$ for a denested basin ω_i. In Holmström (1970), $\beta_k(\omega_i)$ and $\psi_k(t)$ are also termed as weights and amplitude functions, respectively. They both are dimensionless with values in the range $(-\infty,+\infty)$. The centred series $X(t,\omega_i)$ for an area ω_i are thus obtained as a linear combination of different amplitude functions projected on the weight vectors for this area. The basic equation to determine $\beta_k(\omega_i)$ for the case with discharge Q expressed in m^3/s is written down as:

$$\sum_{j=1}^{M}\mathrm{cov}_Q(\omega_i,\omega_j)\beta_k(\omega_j) = \lambda_k\beta_k(\omega_i);\ i = 1,\ldots,M \quad (3)$$

The covariance matrix $\mathrm{cov}_Q(\omega_i,\omega_j)$ in Eq. (3) between denested time series at gauging stations i and j can be directly estimated from the sample spatial variance-covariance matrix $[\hat{c}(\omega_i,\omega_j)]$.