Advances in Water Resources 96 (2016) 11–22

Contents lists available at ScienceDirect

Advances in Water Resources

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

Regional flow duration curves: Geostatistical techniques versus multivariate regression

Alessio Pugliese[a,*], William H Farmer[b], Attilio Castellarin[a], Stacey A. Archfield[c], Richard M. Vogel[d]

[a] Department of Civil, Chemical, Environmental and Materials Engineering – DICAM, University of Bologna, Bologna, Italy
[b] U.S. Geological Survey, Denver, CO, USA
[c] U.S. Geological Survey, Reston, VA, USA
[d] Department of Civil and Environmental Engineering, Tufts University, Medford, MA, USA

ARTICLE INFO

Article history:
Received 12 October 2015
Revised 11 June 2016
Accepted 11 June 2016
Available online 18 June 2016

Keywords:
Flow-duration curve
Top-kriging
Linear regression
Prediction in ungauged basins (pub problem)
Regional analysis
Geostatistics
Southeastern United States

ABSTRACT

A period-of-record flow duration curve (FDC) represents the relationship between the magnitude and frequency of daily streamflows. Prediction of FDCs is of great importance for locations characterized by sparse or missing streamflow observations. We present a detailed comparison of two methods which are capable of predicting an FDC at ungauged basins: (1) an adaptation of the geostatistical method, Top-kriging, employing a linear weighted average of dimensionless empirical FDCs, standardised with a reference streamflow value; and (2) regional multiple linear regression of streamflow quantiles, perhaps the most common method for the prediction of FDCs at ungauged sites. In particular, Top-kriging relies on a metric for expressing the similarity between catchments computed as the negative deviation of the FDC from a reference streamflow value, which we termed total negative deviation (TND). Comparisons of these two methods are made in 182 largely unregulated river catchments in the southeastern U.S. using a three-fold cross-validation algorithm. Our results reveal that the two methods perform similarly throughout flow-regimes, with average Nash-Sutcliffe Efficiencies 0.566 and 0.662, (0.883 and 0.829 on log-transformed quantiles) for the geostatistical and the linear regression models, respectively. The differences between the reproduction of FDC's occurred mostly for low flows with exceedance probability (i.e. duration) above 0.98.

1. Introduction

A flow duration curve (FDC) is a graphical depiction of the cumulative distribution of streamflows in a river catchment. Given a record of streamflow, the FDC can be empirically estimated by ranking the streamflows and estimating a corresponding exceedance probability from an appropriate plotting position (e.g., Weibull, Blom, etc., see Stedinger et al., 1993; Vogel and Fennessey, 1994). The resulting curve, which is a relationship between exceedance probabilities (or flow durations) and discharge, indicates the percentage of time a given streamflow value has been equalled or exceeded over an historical period (Vogel and Fennessey, 1994). Although many variations exist, FDCs are often constructed from daily streamflows and consider either each year individually (one FDC for each year of record) or the entire period of record (one FDC for the entire period). The former are useful for assessing the year-to-year variability of streamflow, whereas the latter can be considered a steady-state picture of the entire hydrological regime over the period considered (Hughes and Smakhtin, 1996). This work is concerned with period-of-record, daily FDCs, which are essential to the development and management of water resources and are routinely required in hydropower generation, design and water supply systems, irrigation planning and management, waste-load allocation, sedimentation studies, habitat suitability and many other water resource investigations (see Vogel and Fennessey, 1995; and Castellarin et al. 2013).

While empirical FDCs provide an important characterization of the behaviour of streamflow in a watershed, they require streamflow data in order to be constructed. Therefore, due to a lack of data, FDCs are not readily available along ungauged stream reaches. This is problematic because these are often the very regions where we have the greatest need for an understanding of streamflow behaviour. Recognizing this need, the prediction of FDCs in gauged, partially gauged, and ungauged sites has been, for years, an extremely active area of research (Singh, 1971; Dingman, 1978; Fennessey and Vogel, 1990; Castellarin et al., 2013). Given increasing

* Corresponding author.
E-mail address: alessio.pugliese3@unibo.it (A. Pugliese).

http://dx.doi.org/10.1016/j.advwatres.2016.06.008
0309-1708/

concerns relating to our ability to predict streamflow properties at ungauged locations, the International Association of Hydrological Sciences (IAHS) launched an initiative for Predictions in Ungauged Basins (PUB) (Sivapalan et al., 2003). The prediction of FDCs at ungauged sites, because of their widespread use in water resources engineering, was one of the main goals of the PUB initiative.

Concerning the problem of FDC prediction in ungauged basins, regional models proposed in the literature follow a variety of different approaches and conceptualizations, which are covered e.g. by Castellarin et al. (2013, 2004). We concentrate on two different prediction strategies, namely regional quantile-regression or simply regional regression (see e.g. Castellarin et al., 2013), which is a classical and widely used approach, and an alternative approach based on geostatistical interpolation. We present a detailed and comprehensive comparison of the potential, ease of implementation, and reliability of the two approaches relative to a broad geographical area in the southeastern United States.

Regional regression has long been used to predict daily FDCs in ungauged basins (Fennessey and Vogel, 1990; Klemeš, 2000; Castellarin et al., 2013). An example of such an approach is provided in a recent comparison of techniques for predicting continuous time series of daily streamflow (see Farmer et al., 2014), one of which required the prediction of ungauged FDCs prior to time series prediction. In their study, treating quantiles as nearly independent, Farmer et al. (2014) used a regional multiple-linear regression to produce an estimate of the logarithmically transformed quantiles as a function of at-site basin characteristics. With discontinuous estimates of specific quantiles, log-normal interpolation is used to complete the continuous FDC (Farmer et al., 2014). While this method is objective, reliable, and easy to implement, it does not account for interdependency among quantiles along the FDC. Ignoring quantile dependencies can lead to complications, such as the failure of the resulting FDC to exhibit its expected monotonic property. For example, Archfield et al. (2010) employed a recursive approach to estimation of regional FDCs to ensure the monotonic properties of FDCs are reproduced.

Over the past decade, geostatistical approaches to predicting streamflow indices in ungauged basins have become increasingly popular (see e.g. Chokmani and Ouarda, 2004; Skøien et al., 2006; Castiglioni et al., 2009; Archfield et al., 2013). Such techniques, relying on kriging methods, do not require identification of hydrologically homogeneous regions. Using a kriging-based weighting scheme, Castellarin (2014) introduces the prediction of a continuous FDC within a three-dimensional *xyz* space, where *x* and *y* are functions of the physiographic and climatic catchment descriptors, while *z* represents the streamflow duration in terms of standard-normal variate. Another viable strategy is to predict the FDC as a single, continuous curve along the duration domain, removing the need for interpolation between quantiles. As an example of this approach, Pugliese et al. (2014) show how to predict FDCs through a comprehensive point index of the FDC which characterizes, in some extent, the shape of the curve. This index can be estimated in ungauged basins employing Top-kriging (Skøien et al., 2006), can be used for expressing hydrological similarity between catchments, and can yield weights to directly relate all FDCs in a region.

This paper contrasts the ability to reproduce the FDC at ungauged sites of traditional regional-regression approaches applied by Farmer et al. (2014), and others, with that of the geostatistical technique of Pugliese et al. (2014). We implemented a threefold cross-validation algorithm to verify the accuracy of the resulting predictions for each ungauged site. The main objective of this study is to provide a comprehensive assessment and comparison of the prediction capabilities and reliability for each method. A secondary goal is to provide guidance for future research on estimation of FDCs at ungauged sites, especially in terms of identifying particular aspects of each of the methods that offer opportunities for improvement. Because of the tremendous attention given to regional regression approaches for estimation of streamflow statistics in the past, and the recent innovations relating to the use of geostatistical methods, it is our goal to provide guidance and assessment for new opportunities relating to the use of both of these approaches for estimation of time series of daily streamflow at ungauged sites.

The methods compared in this study are computationally intensive, and are based on several previous studies which have developed regional regression equations (Farmer et al., 2014) and geostatistical methods based on Top-Kriging (Pugliese et al., 2014) for the purpose of estimating FDCs at ungauged sites. Most of the methods have been developed and discussed elsewhere, however, to our knowledge, this if the first effort to compare these computationally intensive methods using a rigorous cross-validation experiment across a broad and hydrologically diverse region consisting of hundreds of gauged basins with rather heterogeneous geomorphological and climatic conditions. Since the methods are complex and have been discussed elsewhere, the following sections only provide a brief overview, concentrating mostly on a large and complex experimental program which was developed for this study to provide a comparative assessment of their performance at ungauged sites.

2. Materials and methods

We provide a comparison between two previously-developed techniques for the prediction of FDCs at ungauged sites: (1) regional multiple-linear regression of independent quantiles and (2) an adaptation of Top-kriging capable of predicting continuous FDCs. In this section we describe the study area, the construction of empirical FDCs, the regional regression techniques, the geostatistical tools, and how the prediction methods were implemented.

2.1. Study area, streamflow data, and empirical flow duration curves

The study area is located in the southeastern United States and covers an area of approximately 355.000 km^2. The climate is generally warm and humid with average temperatures ranging from 19.9°C in the southern part to 10.4°C at the northern reaches; mean annual precipitation spans from 1150 to 2070 mm per year (Gotvald et al., 2009; Farmer et al., 2014). The study area encompasses parts of Alabama, Florida, Georgia, Mississippi, North Carolina, South Carolina, Tennessee, and Virginia and includes 182 gauged river catchments, which are considered to be relatively undeveloped and only minimally impacted by regulation (Falcone, 2011). Fig. 1 shows the spatial distribution of catchments across the study area. Table 1 quantifies the distribution of key hydrologic and climatic characteristics across all catchments.

For all basins, daily streamflows series were obtained from U.S. Geological Survey (USGS) streamgauges (U.S. Geological Survey Water Data for the Nation, http://dx.doi.org/10.5066/F7P55KJN). As described by Farmer et al. (2014), these streamgauges were screened to have at least 6 complete calendar years of daily streamflow recorded between 01/10/1980 and 30/03/2010. The streamflow sequences of a few sites (6 out of 182) contained zero flow values, resulting in 0.3% of the total station-days data, thus, in order to be logarithmically transformed, zeros were censored at 0.001ft^3/s ($\cong$ 0.00003m^3/s) (see details in Farmer et al., 2014, Table 1). The vast majority of the streamgauges were considered reference quality by the GAGES-II database (Falcone, 2011), though some were included on the basis of previous flood-frequency analyses (Gotvald et al., 2009). The tables and appendices of Farmer et al. (2014), provide detailed information on the streamgauges selected and their associated watersheds.

Empirical FDCs were constructed from the daily streamflow series by ranking the streamflows from complete water years

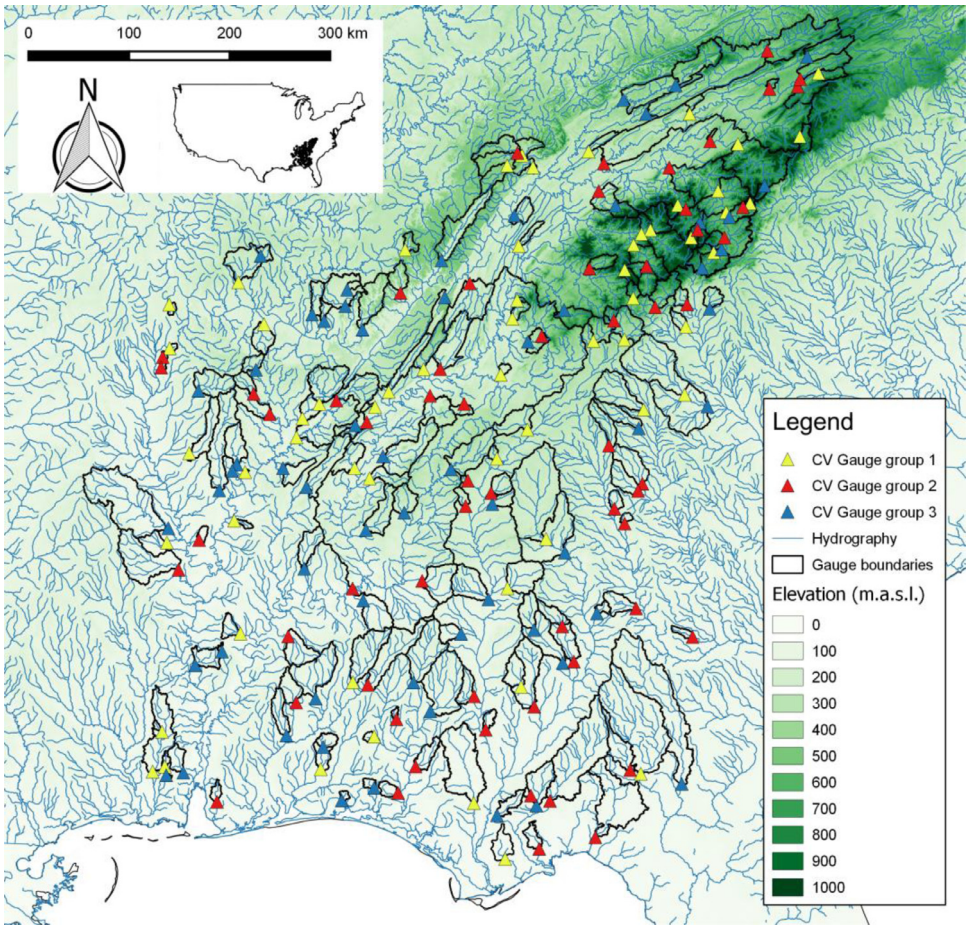

Fig. 1. Study area in the southeastern United States, showing 182 gauged catchments included in the analysis. Triangles indicating gauge locations are coloured according to groups used for the three-fold cross-validation (CV Gauge group k, with $k = 1,2,3$).

Table 1
Statistics of catchment characteristics for 182 gauged catchments included in the study: length of the observed streamflow series (y), drainage area (A), mean annual flow (MAF), 95% exceeded quantiles divided by MAF (q95), mean annual precipitation (MAP), mean annual potential evapotranspiration (PET), mean annual temperature at catchment scale (Tmean), mean basin elevation (Hmean), empirical total negative deviation (TND).

	y (yrs)	A (km^2)	MAF (m^3/s)	q95 (-)	MAP (mm)	PET (mm)	T_{mean} (°C)	H_{mean} (m)	TND (-)
Minimum	6.0	15	0.3	0	1150	577	10	18	1.373
First quartile	18.2	226	3.3	0.05	1360	754	13	116	2.679
Median	30.0	588	6.9	0.12	1460	862	16	252	2.998
Mean	24.3	1427	15.4	0.13	1470	846	15	371	2.995
Third quartile	30.0	1317	15.8	0.21	1550	955	18	521	3.351
Maximum	30.0	56610	598.1	0.59	2070	1042	20	1452	4.054

and assigning an appropriate probability to each rank. Probabilities are assigned using the Blom plotting position, which defines the probability, or duration, of the i^{th} observation as $d_i = (i - 0.375)/(n + 0.25)$, where n is the series length. This plotting position gives unbiased quantiles of the Normal distribution, which is a reasonable approximation for logarithmically transformed streamflows (Stedinger et al., 1993). In order to improve the visualization of streamflow quantiles corresponding to small and large durations, i.e. floods and low flows respectively, we employ standard normal variates for plotting streamflows.

2.2. Regional multiple linear regression of independent quantiles

One of the most common techniques for the prediction of FDCs at ungauged sites is the use of regional regression (Castellarin et al., 2013). For this work, we relied on a previous application of regional regression that was applied to our study area by Farmer et al. (2014). The methodology is summarized here, but further information can be obtained by referring to the original source.

At its simplest, regional regression treats the continuous FDC as discrete points. Regressions of selected FDC points are then built independently of each other. Farmer et al. (2014) chose to discretize the FDC into 27 quantiles, with durations of 0.02, 0.05, 0.1, 0.2, 0.5, 1, 2, 5, 10, 20, 25, 30, 40, 50, 60, 70, 75, 80, 90, 95, 98, 99, 99.5, 99.8, 99.9, 99.95, and 99.98 percent.

A log-linear regression model was developed for each duration using a range of basin characteristics from the GAGES-II database (Falcone, 2011). Some of the explanatory variables used are logarithmically transformed so the "log-linear" regression is linear with respect to the potentially transformed explanatory variables. Weighted-least squares regression was used to weight quantile

Table 2
Summary of the catchment characteristics used for quantiles log-linear regression. A letter "x" distinguishes whether or not the recalled variable is employed in one of the selected durations range, i.e. 0.02–2%, 5–95% and 98–99.98%. Basin characteristics are reported in the first column along with the variable codename used in Farmer et al. (2014).

Catchment characteristic	Flow regime		
	0.02–2%	5–95%	98–99.98%
Drainage Area (DRAIN_SQKM)	x	x	x
Mean watershed slope (SLOPE_PCT)	x	x	
Mean annual precipitation (PPTAVG_BASIN)	x	x	
Average value of total soil thickness (ROCKDEPAVE)	x	x	x
Percentage of the basin classified as planted or cultivated (PLANTNLCD06)	x		
Rainfall and runoff coefficient from the Universal Soil Loss Equation (RFACT)			x
Average silt content of soils (SILTAVE)			x

estimates by the record length associated with each streamgauge. Tobit regression was applied to handle zero streamflow observations which are treated as censored values (Greene, 1997, p. 962–7). To help ensure continuity, streamflow durations were broken into three regimes: durations 0.02–5%, 5- 95%, and 98–99.98%. Each regime was associated with a selected set of basin characteristics, even though some of them are shared among flow regimes. For instance, drainage area and the average value of total soil thickness are used in all the regressions regardless the specific flow regime, while the mean annual precipitation is used for high and median flows only. Within each regime, the variables for prediction were held constant and only the coefficient values were allowed to vary. Table 2 shows a summary of the basin characteristics used for all the three flow regimes in the regional regression method. Further details and the final results of these regressions can be found in Farmer et al. (2014) (see Table 4 of the supplementary material at http://pubs.usgs.gov/sir/2014/5231/table/sir2014-5231_tables%201 – 7.pdf).

For the sake of brevity we will refer to regional regression method as RR in the remainder of the manuscript.

2.3. Geostatistical prediction of continuous flow duration curves

Top-kriging (or topological kriging) is a powerful geostatistical procedure developed by Skøien et al. (2006) for the prediction of hydrological variables. Like all kriging approaches, Top-kriging produces predictions of hydrologic phenomena at ungauged sites with a linear combination of the empirical information collected at neighbouring gauging stations. That is, the unknown value of the streamflow index of interest at prediction location x_0, $Z(x_0)$, can be estimated as a weighted average of the variable measured in the neighbourhood:

$$Z(x_0) = \sum_{i=1}^{n} \lambda_i Z(x_i) \tag{1}$$

where λ_i is the kriging weight for the empirical value $Z(x_i)$ at location x_i, and n is the number of neighbouring stations used for interpolation. Kriging weights λ_i can be found by solving the typical ordinary kriging linear system (2a) with the constraint of unbiased estimation (2b):

$$\sum_{j=1}^{n} \gamma_{i,j}\lambda_j + \theta = \gamma_{i,0} \quad i = 1, \ldots, n \tag{2a}$$

$$\sum_{j=1}^{n} \lambda_j = 1 \tag{2b}$$

where θ is the Lagrange parameter and $\gamma_{i,j}$ is the semi-variance between catchment i and j (Isaaks and Srivastava, 1990). The variogram, which represents the semi-variance of the increment $Z(x_i) - Z(x_j)$ between catchments i and j with respect of catchments distance (see for details Skøien et al., 2014), delivers the spatial variability of the regionalised variable Z across sites. Top-kriging considers the variable defined over a non-zero support S, the catchment drainage area (Cressie, 1993; Skøien et al., 2006). The kriging system of Eqs. (2) remains the same, but the semi-variances between the measurements need to be obtained by regularization, i.e. smoothing the point variogram over the support area. The point variogram can then be back-calculated by fitting aggregated variogram values to the sample variogram (Skøien et al., 2006).

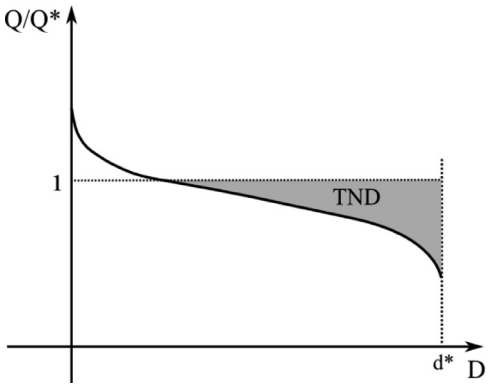

Fig. 2. The shaded area qualitatively illustrates the meaning of Total Negative Deviation (TND).

Pugliese et al. (2014) proposed a method for using Top-kriging to predict continuous FDCs at ungauged locations; indeed, they reduce the dimensionality of the problem by seeking a unique index of site-specific FDCs. Unlike the regional regression approach, which treats quantiles as independent, this Top-kriging-based approach considers the entire curve simultaneously. This is accomplished by first standardising the empirical FDCs at site x, $\Psi(x, d)$, for some reference value, $Q^*(x)$, to yield a dimensionless FDC:

$$\psi(x, d) = \frac{\Psi(x, d)}{Q^*(x)}, \tag{3}$$

where d denotes a specific duration. The reference value can be a given streamflow statistic, such as the long-term average of the daily streamflow series. Pugliese et al. (2014) identified an overall point index that effectively summarizes the entire curve. This index, which the authors termed total negative deviation (TND), is derived by integrating the area between the lower limb of the FDC and the reference streamflow value Q^* (see Fig. 2).

Empirically, TND values are computed as:

$$TND(x) = \sum_{i=1}^{m} |q_i(x) - 1| \Delta_i \tag{4}$$

where $q_i = \frac{Q_i}{Q^*}$ represents the ith empirical dimensionless quantile standardised for the selected reference value Q^*, Δ_i is half of the frequency interval between the $(i+1)^{\text{th}}$ and $(i-1)^{\text{th}}$ quantile and the summation involves only the m standardised quantiles less than 1. The range of the summation, m, in Eq. (4) is a function of the maximum duration d^*. Duration d^* is itself a function of the minimum record length across gauged sites in the study region; therefore, using the Blom plotting position and the information in Table 1, the maximum duration was set to $d^* = 0.9997$.

Having calculated empirical TNDs, Pugliese et al. (2014) propose using the TNDs as a regionalised variable to develop site-specific weighting schemes. The same weights derived through the solution of the linear kriging system (2) for TND are used for a batch prediction of the continuous, dimensionless FDC for the ungauged site x_0:

$$\hat{\psi}(x_0, d) = \sum_{i=1}^{n} \lambda_i \psi(x_i, d) \;\forall\; d \in (0, 1) \tag{5}$$

where λ_i, with $i = 1, \ldots, n$, are the weights resulting from the kriging interpolation of TNDs; $\psi(x_i, d)$ is the dimensionless, empirical FDC at the donor site x_i, and $\hat{\psi}(x_0, d)$ is the predicted dimensionless FDC. It is worth highlighting that the computation of the linear kriging system (2) depends on n, the number of neighbouring sites on which to base the spatial interpolation, a fact that will be explored below.

Once a reliable model (e.g., a regional regression model, or kriging model, etc.) for predicting Q^* at the ungauged site x_0 has been set up for the study region, the prediction of the dimensional FDC, $\hat{\Psi}(x_0, d)$, can be obtained as:

$$\hat{\Psi}(x_0, d) = \hat{Q}^*(x_0)\hat{\psi}(x_0, d) \;\forall\; d \in (0, 1) \tag{6}$$

where $\hat{Q}^*(x_0)$ is the prediction of Q^* at the ungauged site x_0 and $\hat{\psi}(x_0, d)$ has the same meaning as in (5). For the sake of brevity this method of prediction is referred to herein as Total Negative Deviation Top-kriging (TNDTK). TNDTK was applied with the same 27-point resampling of the FDC developed for regional regressions (RR) above.

2.4. Cross validation procedure and comparative assessments

A three-fold cross-validation (3FCV) procedure was used to equitably compare the regional regression model against TNDTK. The dataset was divided into three random subsets (see Fig. 1); each prediction model was calibrated on two-thirds of the data and then applied to produce an 'ungauged' prediction on the remaining third. Iterating for each third, the algorithm leads to predictions for each and every site across the study area. We used the same random subsets as in Farmer et al. (2014).

The performance of each prediction method was assessed using the Nash-Sutcliffe efficiency index either in natural, i.e. real, space (NSE) and in logarithmic space (LNSE) (Nash and Sutcliffe, 1970). These are computed as follows:

$$NSE_j = 1 - \frac{\sum_{k=1}^{n_d} \left(\Psi\left(x_j,\ d_k\right) - \hat{\Psi}\left(x_j, d_k\right)\right)^2}{\sum_{k=1}^{n_d} \left(\Psi\left(x_j, d_k\right) - \mu_j\right)^2} \quad j = 1, \ldots, n_s$$

$$LNSE_j = 1 - \frac{\sum_{k=1}^{n_d} \left(\ln \Psi\left(x_j,\ d_k\right) - \ln \hat{\Psi}\left(x_j, d_k\right)\right)^2}{\sum_{k=1}^{n_d} \left(\ln \Psi\left(x_j,\ d_k\right) - \omega_j\right)^2} \quad j = 1, \ldots, n_s \tag{7}$$

where $\Psi(x_j,\ d_k)(\mathrm{m}^3/\mathrm{s})$ and $\hat{\Psi}(x_j, d_k)$ $(\mathrm{m}^3/\mathrm{s})$ are the empirical and predicted kth streamflow quantiles at site x_j, respectively, μ_j is the mean of the empirical streamflow quantiles at site x_j, ω_j is the mean of the logarithms of the empirical streamflow quantiles at site x_j, while n_s and n_d are respectively the total number of stations (i.e. 182) and the number of selected quantiles for FDC discretization (i.e. 27).

In addition to a site-by-site comparison of performance, the Nash-Sutcliffe efficiencies can be used to estimate the performance across streamflow quantiles. This is accomplished by summing over sites rather than durations in Eqs. (7):

$$NSE_k = 1 - \frac{\sum_{j=1}^{n_s} \left(\Psi\left(x_j,\ d_k\right) - \hat{\Psi}\left(x_j, d_k\right)\right)^2}{\sum_{j=1}^{n_s} \left(\Psi\left(x_j,\ d_k\right) - \mu_k\right)^2} \quad k = 1, \ldots, n_d \tag{8}$$

and likewise with LNSE, with the same meaning of symbols as in (7). This metric allows one to visualize the performance of a selected model as a function of the duration interval and, specifically, to assess how the results vary in different streamflow regimes, e.g. high-flows rather than low-flows. It should be noted, however, that the cross-site range of streamflows for a particular quantile is greater than the within-site range across the FDC, a fact which may affect the usefulness of a quantile-specific NSE.

An additional metric of performance was the overall error of prediction

$$\delta_j = \sum_{k=1}^{n_d} \left|\Psi\left(x_j,\ d_k\right) - \hat{\Psi}\left(x_j, d_k\right)\right| \quad j = 1, \ldots,\ n_s \tag{9}$$

where $\Psi(x_j,\ d_k)$ and $\hat{\Psi}(x_j, d_k)$ $(\mathrm{m}^3/\mathrm{s})$ are the same variables as above and δ_j $(\mathrm{m}^3/\mathrm{s})$ is the overall error computed for a given model. This metric captures the overall distance between empirical and predicted FDCs by computing the absolute error at each duration interval and then summing across the range of durations (see Ganora et al., 2009). Finally, two statistical non-parametric hypothesis tests were employed to verify whether or not the model errors δ_i are significantly greater for TNDTK compared to RR. To accomplish this task we used (1) the Wilcoxon signed-rank test with the null hypothesis that TNDTK errors are larger than RR and (2) the exact binomial test, which performs an exact test about the probability of success in a Bernoulli experiment, both at 5% significance level. For the latter, we considered the random variable X, defined as the number of sites out of 182 for which model errors δ_i are lower for TNDTK relative to RR ones, under the assumption that X follows a Binomial distribution with the number of trials $n = 182$ (i.e. number of catchments) and the hypothesized probability of success $p = 0.5$ (see Hollander and Wolfe, 1999; R Core Team, 2016).

3. Results

3.1. Prediction of mean annual streamflow

For this application we considered the mean annual streamflow (MAF) as the reference streamflow Q^* used to standardise the streamflows, because it is a traditional method to standardise FDCs. However, in our ungauged application, this necessitates the prediction of the mean annual streamflow before FDCs can be back-transformed from TNDTK. A power-law model of MAF as a function of drainage area (A) was employed to investigate how much of the MAF variability is explained by the influence of drainage area.

Using regional regression, we observed a strong ($R^2 \cong 0.927$) log-log relationship between the MAF and A (i.e. scaling exponent resulted equal to 0.93; see Fig. 3a). Fig. 3 demonstrates that drainage area can be considered one of the primary drivers controlling the average discharge across the study area.

This power-law model was therefore used to standardise empirical MAF values to be predicted with Top-kriging, which directly handles drainage area as a key variable of the model. The MAF was scaled by a factor of $A^{0.93}$, where A is the drainage area of

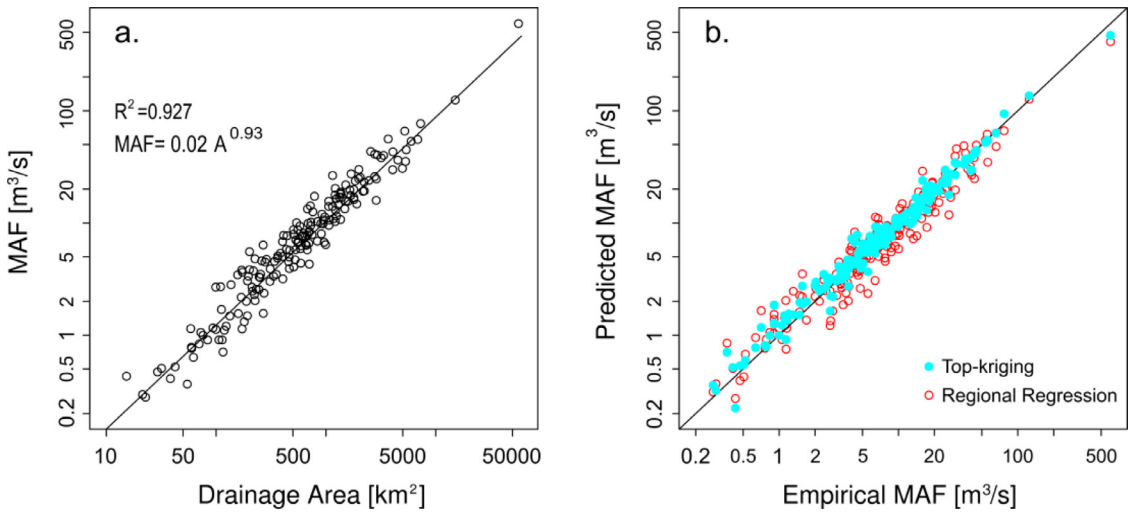

Fig. 3. Left panel: scatter diagram between Mean Annual Flow (MAF) and Drainage Area (A). Right panel: empirical vs. predicted MAF with either Top-kriging (cyan dots) or regional regression (red circles). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.) (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the specific catchment, and Top-kriging was then applied in order to predict MAF using the 3FCV algorithm, for different neighbourhood sizes n, analogous to the sensitivity analysis performed for TNDTK which is described in Section. 3.2. Similarly, we employed the same cross-validation algorithm with the regional regression model, in order to consistently assess which one of the two models performs better for the prediction of MAF in ungauged sites. Fig. 3b shows an example of the performance of Top-kriging using a neighbourhood of $n = 6$ (cyan dots) as well as the performance of regional regression (red circles). Fig. 3b reports on the x-axis the empirical MAF values (m^3/s) against either Top-kriging or regional regression predictions of MAF (m^3/s) on the y-axis, as the result of the 3FCV algorithm. The assessment of the prediction capabilities in terms of Nash-Sutcliffe efficiency, computed either in natural (NSE) or in log-transformed space (LNSE), reveals very good performances for Top-kriging, with NSE and LNSE equal to 0.93 and 0.97, respectively, whereas a slight drop in performance is obtained using regional regression, with NSE and LNSE equal to 0.90 and 0.93, respectively.

It is worth noting that the differences in terms of both NSE and LNSE for the regional linear regression and Top-kriging are rather minor. Therefore, even though the two procedures are interchangeable from a practical viewpoint for the study area, we decided to employ the geostatistical method as the reference method for the prediction of MAF in ungauged basins for the 3FCV cross-validation.

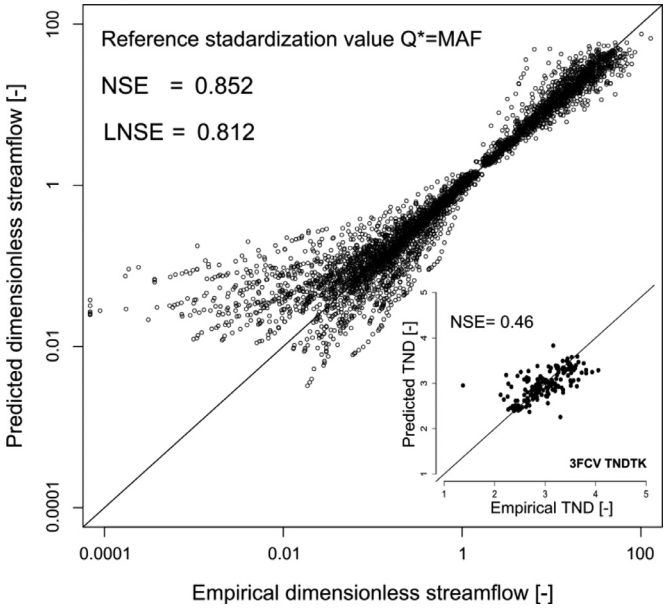

Fig. 4. Prediction of dimensionless FDC in cross-validation for TNDTK: empirical vs. predicted standardised streamflows (big panel); empirical vs. predicted TND values (small panel).

3.2. Prediction of dimensionless FDCs

The TNDTK method produces estimates of dimensionless FDCs. While this is useful for regionalisation studies, it is not directly comparable to the methods employed by Farmer et al. (2014) for estimation of the FDC itself. For this reason, we only briefly consider the predictive performance of TNDTK for dimensionless FDCs. A preliminary sensitivity analysis suggested that a neighbour of $n = 6$ sites provided the best results; these results are summarized in Fig. 4.

The main panel of Fig. 4 shows the relationship between empirical and 3FCV predicted dimensionless quantiles (27 for each site). The sub-panel in the bottom right corner shows the empirical TND values (x-axis) against their prediction in cross-validation (y-axis). The quoted NSEs are for all streamflow quantiles in the plot, across sites and durations. Note that since the NSE is so sensitive to the range of the predictions and observations that this measure may be obscured by outliers, especially on the high end. However, it is an adequate metric for this initial assessment. Surprisingly, the Top-kriging model performs rather poorly for several locations when predicting TND (NSE of 0.46), but very well when predicting dimensionless streamflow quantiles, especially for high-flows (NSE of 0.852). This suggests that the weights from Top-kriging of TND yield significant added value when used for predicting FDCs by averaging empirical dimensionless curves. Also, this finding indicates that simple Top-kriging may not predict TND values in ungauged locations accurately enough for some applications and a more sophisticated prediction technique needs to be identified and tested for a reliable prediction of TND values (e.g. Top-kriging with exogenous variables, see e.g. Chokmani and Ouarda, 2004; Archfield et al., 2013), which is an open avenue for future studies.

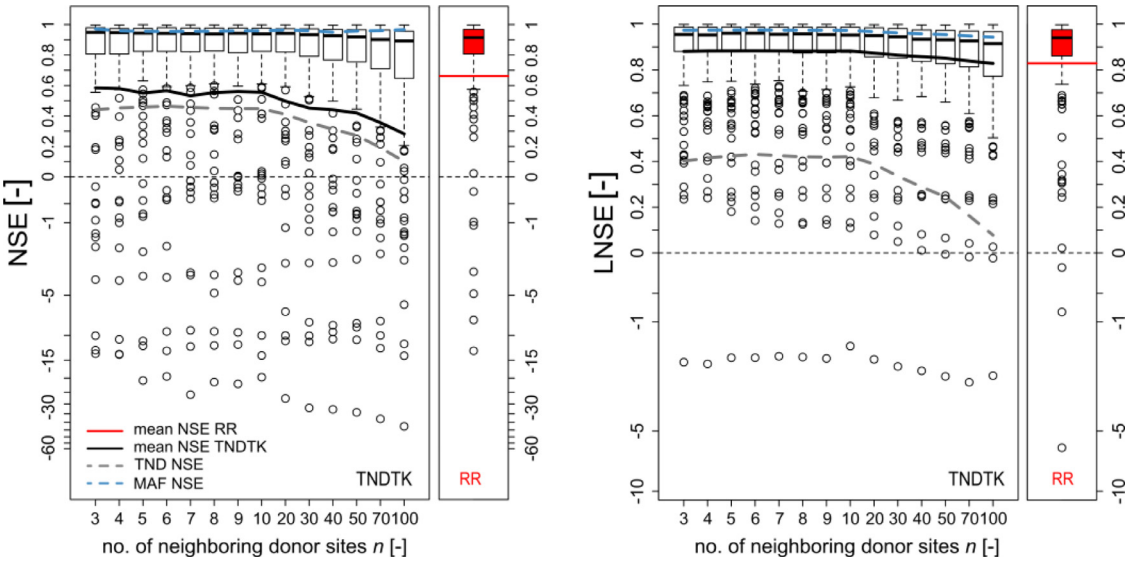

Fig. 5. Distributions of at-site Nash-Sutcliffe efficiencies for natural (NSE, left panel) and log-transformed (LNSE, right panel) streamflows plotted against the number n of neighbouring sites used for the interpolation with TNDTK. Efficiency values are represented as a box-and-whiskers plot summarizing the 1st, 2nd (median) and 3rd quartiles along with whiskers extending to the most extreme data point which is no more than 1.5 times the interquartile-range away from the nearest quartile. Circles indicate extreme members of the distribution. The black solid line illustrates the mean of the distributions; gray and cyan dashed lines, indicate the kriging performance relative to TND and MAF, respectively. The prediction performance for RR is illustrated in a similar fashion using a red box (the red horizontal segment illustrates the mean value). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.3. Comparative assessment of regional regression and TNDTK

In this section a rigorous comparison of estimated dimensional FDCs is provided using both regional regression and top kriging. Concerning TNDTK, we obtained a cross-validated prediction of dimensional FDC at any given site in the region as the product of locally predicted MAF (see Section. 3.1) and dimensionless FDC (see Section. 3.2). Fig. 5 shows the distribution of at-site NSEs and LNSEs for both the regional regression and several iterations of the TNDTK method (i.e. varying the size of the neighbourhood from 3 to 100 members). Before considering the relative performance between regional regression and TNDTK, the most remarkable result is the effect of neighbourhood size on the performance of TNDTK. As can be seen in Fig. 5, for both NSE and LNSE, the mean performance tends to decrease as n increases, especially for $n \geq 10$. Indeed, the mean NSE reaches a minimum of 0.282 at $n = 100$, whereas the best performances are obtained over the interval $3 \leq n \leq 8$, with mean NSE ranging from 0.556 to 0.585 and the median NSE ranging from 0.939 to 0.949. LNSE is generally greater than NSE, but still shows, to a lesser extent, the same generally negative correlation with neighbourhood size. The mean LNSE is always above 0.8, with a maximum of 0.883 at $n = 6$, while the medians are always above 0.91 with a maximum of 0.961 at $n = 6$. Furthermore, the decreasing performance with increasing n is also characterized by an increasing frequency of increasingly more extreme low-end outliers. The source of such decaying performance is likely to be the underlying performance of the kriging system for TND (gray, dashed line). Interestingly, and differently from the prediction of TND and FDC, the prediction of MAF using Top-kriging (cyan, dashed line) does not seem to be affected by the number of neighbouring donor sites and results in efficiencies equal to 0.959 and 0.966 for NSE and LNSE, respectively.

NSE values are similar for RR and TNDTK, though TNDTK exhibits substantially lower efficiency values for a larger number of sites relative to RR, as clearly showed by mean NSE values depicted in Fig. 5. The RR approach yields a mean NSE of 0.662, while the best iteration of TNDTK yields a mean NSE of only 0.585. The median goodness of fit are more competitive: regional regression demonstrates a median NSE of 0.914, while the best of the TNDTK iterations presents a median NSE of 0.949. However, NSE is notoriously sensitive to extreme values, a problem which is particularly significant when considering values across the entire range of streamflows at a site. Instead, LNSE values are much less sensitive to outliers and provide a better overall reflection of the goodness of fit corresponding to the two methods. Performance in terms of NSE values is also valuable, though, as it enables one to compare the results of this study with several other studies that were previously published on the same topic.

The LNSEs show a much more equitable performance between regional regression and TNDTK. This metric, being less sensitive to extreme realizations and variability, is a more honest indicator of performance in predicting FDCs for the same reason that logarithmic estimates of cross correlation provide better estimates of cross correlation than untransformed estimators (see Stedinger, 1981), particularly for streamflow regimes which exhibit high variability. While RR appears to outperform TNDTK in terms of NSE, TNDTK yields a greater LNSE, on average. Mean LNSE reaches a maximum value of 0.883 for TNDTK and neighbourhood size of $n = 6$; whereas mean LNSE is 0.829 for RR. Medians of at-site LNSE values are 0.961 and 0.940, for TNDK and RR respectively.

Overall, the distributions of NSEs and LNSEs are relatively similar and moderately skewed. Nevertheless, the distributions of at-site NSE and LNSE both support the conclusions that the best neighbourhood size for TNDTK is $n = 6$. For the remainder of this manuscript we will focus more closely on the performance of the 6-neighbour iteration of TNDTK and delve into its direct comparison with RR.

In addition to at-site performance, it is useful to consider the performance of both RR and TNDTK for specific quantiles. Fig. 6 shows the cross-site NSE and LNSE of both regional regression and TNDTK. It should be noted that the cross-site variability, in mixing potentially dissimilar sites, can have a significant impact on the interpretations of NSE and LNSE. TNDTK outperforms the regional regression in the low-flow regimes in terms of cross-site NSE, for durations ranging from 0.7 to 1, and in median-high flow regimes, from 0.001 to 0.02. However, RR outperforms the geosta-

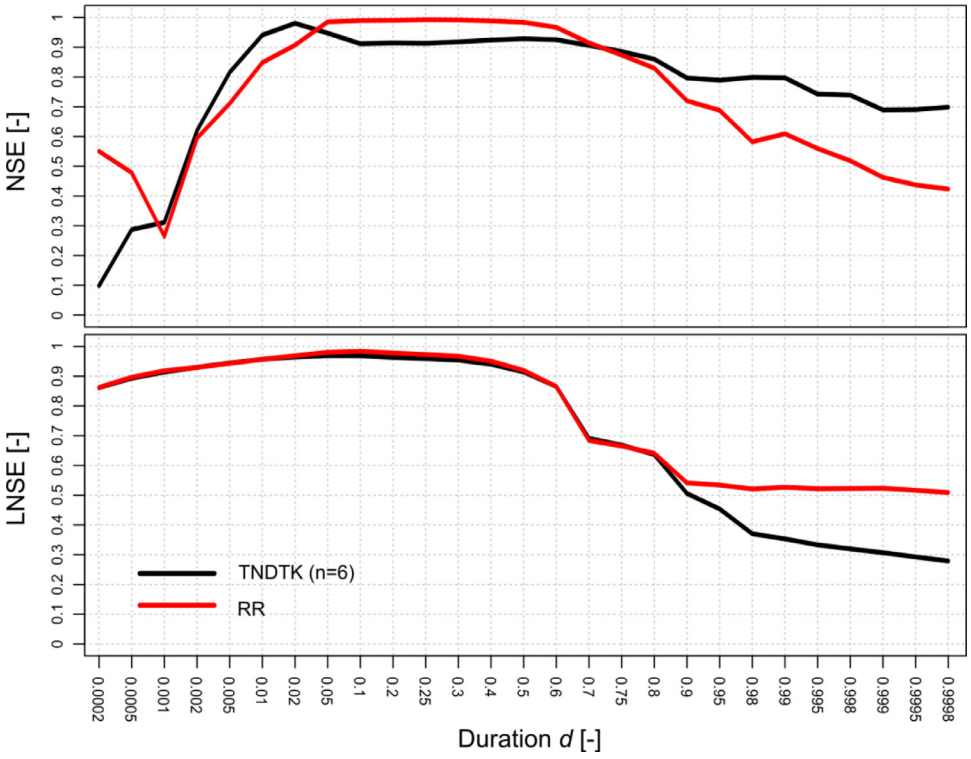

Fig. 6. Nash-Sutcliffe efficiencies for natural (NSE, upper panel) and log-transformed (LNSE, bottom panel) streamflows computed at each considered duration for TNDTK (black line) and RR (red line). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3
TNDTK vs RR: model predictive accuracy in terms of Nash-Sutcliffe Efficiencies in natural (NSE) and in log space (LNSE); areas associated with the 50, 80 and 90% error-bands depicted in Fig. 7 (central panel) are also reported.

	TNDTK	RR
NSE	0.566	0.662
LNSE	0.883	0.829
Area Under 50%	0.528	0.547
Area Under 80%	1.422	1.289
Area under 90%	3.756	1.929

tistical model in the median flows, from 0.05 to 0.7, and the very high flows, i.e. duration lower than 0.001. Despite this variability, the discrepancy appears only significant for the low-flow regimes.

The LNSEs show a notably different story. The performance of TNDTK and regional regression are essentially the same from the high to the mid-low flows regime, i.e. d lower than 0.9, while regional regression model outperforms TNDTK in the low-flow regimes. Furthermore, while NSE indicated that high flow regimes were more poorly estimated than low-flow regimes, the LNSEs suggest the reverse. This discrepancy is likely due to the marked sensitivity of NSE to poor predictions of extremely high streamflow values, and the enhanced diagnosing capability of LNSE when it comes to low-flow predictions.

The upper panel of Fig. 7 further elucidates the comparative assessment between the 6-neighbor TNDTK and RR by plotting the observed quantiles against the modelled quantiles for each method (Fig. 7, upper panel). The results of each method overlap substantially, indicating only slight differences in their corresponding prediction capability. This behaviour is confirmed through the numerical assessment driven by performance indices computed for both TNDTK and RR, reported above and summarized more succinctly in Table. 3. The average at-site three-fold cross-validation NSE is equal to 0.566 and 0.662 for TNDTK and RR, respectively, while the average at-site validation LNSE is equal to 0.883 and 0.829, in the same order. However, the variability of performance is much greater for lower streamflow values. TNDTK tends to overestimate the low-flows regime much more so than regional regression. This suggests that the relative performances seen in the LNSEs of Fig. 6 are more indicative of a general behaviour.

The middle panel of Fig. 7 shows the relative residuals, i.e. the difference for a given site between the predicted and empirical values divided by the empirical one, computed for each of the 27 quantiles. It is worth noting that such a diagram is not symmetric, because the y-axis has a lower bound equal to -1. This plot confirms the overestimation of low-lows for TNDTK and highlights an underestimation of streamflow quantiles for RR in the same duration range; indeed the median behaviour of both models is relatively unbiased for duration below 0.5, then deflections from 0 are clear for durations ranging from 0.6 to 1, i.e. low-flows regime, though in dissimilar direction. Table. 3 reports the magnitude of the areas within error-duration bands illustrated in Fig. 7 (middle panel), which can be interpreted as an average relative error across duration for a fixed accuracy level. There is, between the two methods, a substantial equality in terms of 50 and 80% accuracy levels. Conversely, TNDTK results in a doubled value for the 90% band, an area that is driven by the variability of low-flow regimes in Fig. 6; this further supports the tendency towards overestimation previously highlighted.

Finally, the bottom panel of Fig. 7 shows the at-site comparison in terms of the overall, i.e. throughout duration, distance between empirical and predicted FDCs computed, for both models, per each site with Eq. (9). In this representation equivalence between the models is represented by the solid bisecting line; therefore, if one point falls above the 1:1 line, TNDTK provides a better overall prediction of the empirical FDC than RR for that site, and vice-versa if the point falls below the 1:1 line. This plot shows that TNDTK outperforms regional regression for 109 sites out of 182 (≅60%); the binomial test reveals that the hypothesis for which model

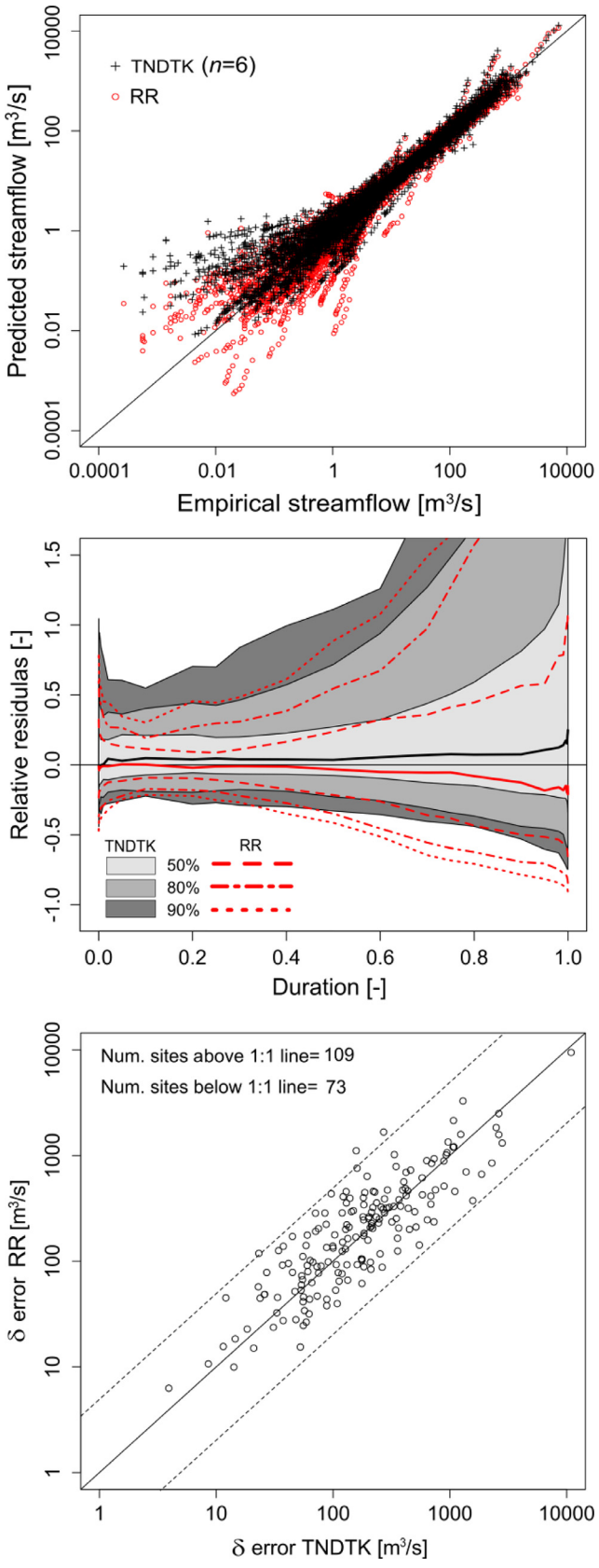

Fig. 7. Comparison between dimensional FDC predictions in cross-validation via TNDTK and RR: empirical vs. predicted dimensional streamflows (upper panel); error-duration bands reporting the median of relative errors against duration (solid lines) as well as error bands containing 50, 80 and 90% of prediction relative errors (middle panel); overall at-site prediction errors computed with Eq. (9) (bottom panel).

errors δ_i are greater for TNDTK relative to RR should be rejected (p-value = 0.0046), which suggests that TDNK is better than RR at 5% significance level. This result is also confirmed by the Wilcoxon signed rank test that reveals the null hypothesis that TNDTK errors are larger than RR should be rejected at a 5% significance level (p-value = 0.0242). However, it is worth noting that, similar to NSE, this error measure defined in Eq. (9) is mostly driven by high-flows, which are generally related to larger absolute errors, thus it likely fails to capture the performance in the low-flow regime.

4. Discussion

4.1. Comprehensive assessment of RR and TNDTK prediction performance

The results of this study reveal that the two approaches compared perform similarly regardless of the specific choice of the model settings. To some extent, they could be considered interchangeable, showing the same results for most of the streamflow regimes, i.e. from very high flows to low-flows. The exception is for very high durations, i.e. $d \geq 0.95$, for which TNDTK is characterized by a positive bias, while RR, even if less emphasized, shows a negative bias (see middle panel in Fig. 7). Indeed, for the very high durations, which are commonly dominated by subsurface flows, the two methods might be seen as complementary to each other: as shown in Fig. 6 by the discrepancies in the behaviour of NSE and LNSE for both high and low durations.

Nevertheless, TNDTK is a promising tool for predicting FDCs in ungauged basins, given the small amount of input data required by such a model, which mainly relies on streamflow series and catchments' size as well as their mutual position. Still, TNDTK exhibited intrinsic weaknesses when contrasted against RR. In terms of at-site performance, the overall metric of LNSE suggested that TNDTK possessed a marked advantage over RR, but, as in Fig. 7, it is clear that TNDTK overestimates low-flows. This result was to be expected. Because a linear weighting scheme that uses only positive weights is adopted, TNDTK is expected to overestimate low-flows for sites exhibiting extremely low dimensionless low-flows and to underestimate high-flows for sites characterized by extremely high dimensionless high-flows (see Fig. 4). The same smoothing does not necessarily apply to multiple regression methods, such as RR. Despite this concern, TNDTK provides methodological advances over regional regression in that it is capable of producing continuous FDCs rather than being constrained to point quantiles. Furthermore, RR, in treating quantiles as independent, may introduce non-monotonic behaviour into the FDCs; although not explored here, the smooth prediction method of TNDTK may enforce monotonicity. In this application we did not find any issue related to non-monotonicity for both methods, yet it may not be the case in a different study area. Further research will address this specific point and will deal with the monotonic property of the curve among different approaches.

One of the most interesting aspects of the models' comparative performance corresponds to the few sites that exhibited extreme negative results. For instance, TNDTK produces in turn 10 and 1, out of 182, negatives NSE or LNSE respectively, while RR results in 7 and 3 negatives NSE or LNSE. The maps in Fig. 8 (right panels) show the locations where negative NSE values are produced respectively by both models (upper-right, blue dots), by RR only (middle-right, light green dot) and by TNDTK only (bottom-right, black dots). Six sites are associated with the worst performances for both approaches (blue dots); these sites are located mainly at the periphery of the study area and are grouped in two different climatic regions: a humid group in the south and a more continental group in the north. The poor results obtained for these sites

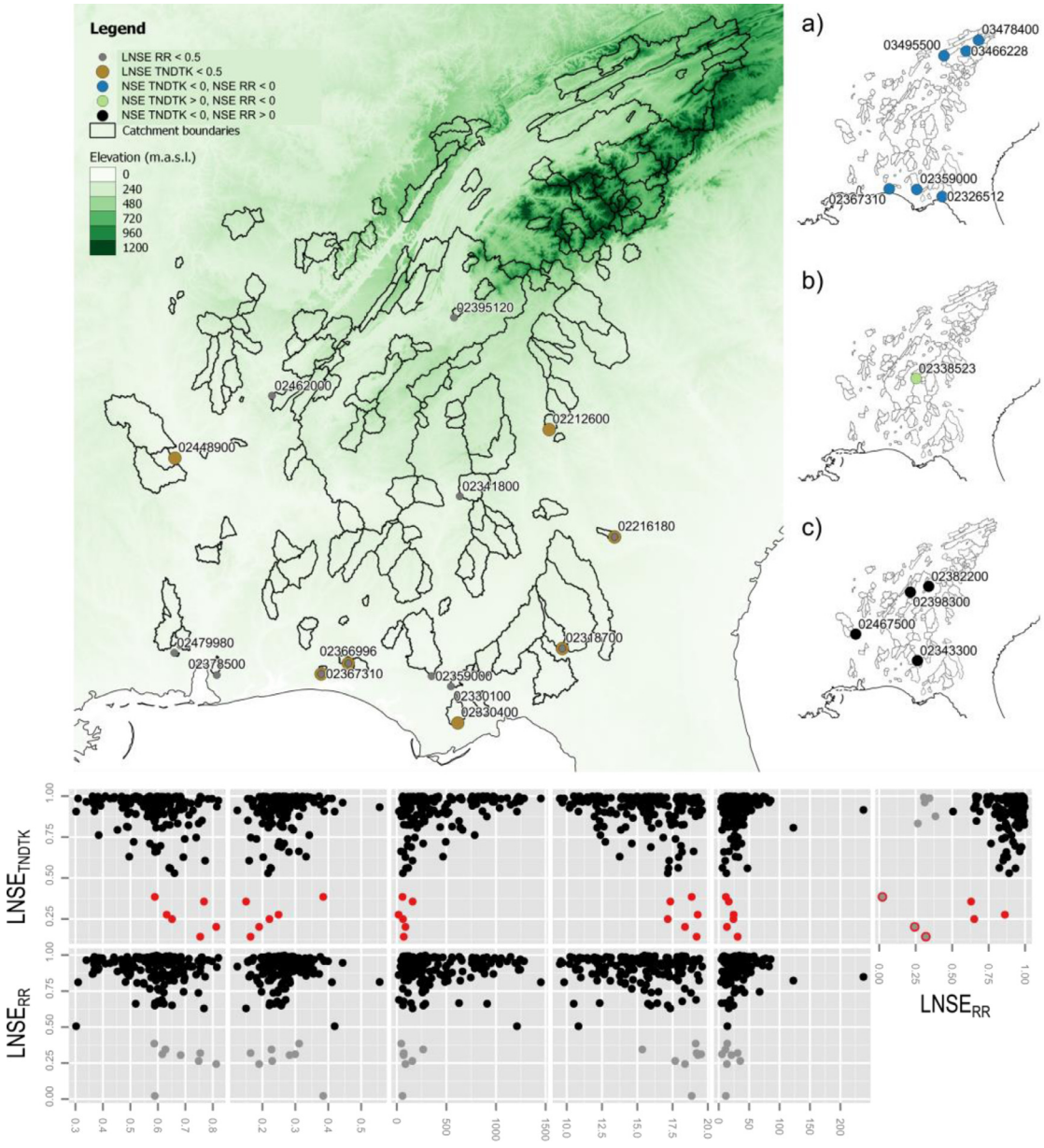

Fig. 8. Top panels: the large central map reports those sites which resulted with LNSE < 0.5 for either TNDTK (red dots) or RR (grey dots); the three small maps in the right side report, respectively, those catchments with a) negative NSEs for both models (blue dots), b) negative NSE for RR only (light green dot), and c) negative NSE for TNDTK only (black dots). Bottom panels: log space Nash-Sutcliffe efficiencies (LNSE) for TNDTK and RR ranging within the [0,1]interval, vs. potential evapotranspiration (PETR), run-off ratio (ROR), mean basin elevation (H_{mean}), mean basin temperature (T_{mean}), square root of drainage area ($\sqrt{A}$), in this order; the rightmost scatterplot reports TNDTK LNSE values vs. RR values. Sites with LNSE values ranging from 0 to 0.5 are highlighted with red dots for TNDTK and grey dots for RR. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

are not surprising, as they correspond to very small or very large basins. Our study highlights that drainage area, mutual position of catchments, nested structure, and differences in climate play a significant role in the prediction of FDCs with Top-kriging and regional regression.

Furthermore, both methods suffer deficiencies in predictive accuracy when the spatial density of gauging stations is low (see e.g. Castiglioni et al., 2011; Parajka et al., 2015). Table 4 presents a summary of drainage areas and nested structure along with the NSE negative efficiencies. For instance, sites 0,349,5500 and 0,347,8400 are nested and should, therefore, benefit from along-stream or nested stream neighbours. However, there is a large difference of more than three orders of magnitude in terms of drainage areas and both sites belong to the same cross-validation group (i.e. group 2) and therefore prediction in one cannot benefit from data collected for the other, and vice-versa. These considerations could be extended to all nested sites that belong to the same cross-validation group. Also, when sites 0,349,5500 and 0,347,8400 are removed from the dataset (i.e. prediction of cross-validated empirical FDCs for all sites belonging to group 2) the cross-validated variogram does not capture the overall variability in terms of drainage area any longer, leading to poorer performance.

Conversely, negative efficiencies obtained at the southern catchments could be reasonably attributed to changes in climatic and geomorphological conditions. A similar reasoning could be adopted for the 4 catchments associated with extremely low NSE values for TNDTK only (back dots); the combination of small catchment areas and wide climatic variability might affect models' ability to predict variation in streamflow regimes. With this in mind, we investigated how such drivers influence the FDC predictions.

Table 4
Focus on those sites where both TNDTK and RR report negative NSE (see also upper-right map in Fig. 8). From left to right, columns report respectively the gauge identification number (ID), the catchment drainage area (A), a logical value that describes whether or not the catchment is nested (NEST), and the cross-validation group (CVG).

	ID	A (km^2)	NEST	CVG (-)	NSE_{TNDTK} (-)	NSE_{RR} (-)
Southern catchments	02367310	99	No	2	−12.89	−7.71
	02359000	2946	No	1	−3.57	−0.24
	02326512	2868	No	3	−9.38	−4.86
Northern catchments	03466228	55	No	3	−0.41	−0.97
	0349,5500	15041	Yes	2	−19.47	−12.89
	03478400	106	Yes	2	−1.86	−3.20

Fig. 8 suggests a climatological pattern to the poor performance of either method. LNSEs of each method, with values lower than 0.5 in grey (RR) or red (TNDTK), are graphed against several explanatory variables derived at a catchment scale (see Falcone, 2011). These include the square-root of drainage area, $\sqrt{A}$ (km), the mean annual basin temperature, T_{mean} (°C), the mean basin elevation, H_{mean} (m), the run-off ratio, $ROR = \frac{MAF}{MAP}$ (-), and potential evapotranspiration ratio, $PETR = \frac{PET}{MAP}$ (-) (Budyko, 1974), while the spatial pattern of the poorly-performing sites is illustrated in the large map in Fig. 8 (note that the bottom panels highlight LNSEs within the [0.1] interval only, which is why the number of dots in the large map does not match with those in the scatterplots). The peculiar sites seems to be small or very small catchments, characterized by high mean annual temperature, low mean elevation, low run-off ratio, and high potential evapotranspiration ratio.

As another interpretation, since ROR and PETR mainly summarize the average annual hydrological water balance of a given catchment, it follows that the hydrological regime of the southern catchments is dominated by subsurface flows along with an increased capacity of storing and retaining water (low ROR), while the sub-humid climate (high PETR) gains the seasonal variability of streamflows (Ponce et al., 2000; Berghuijs et al., 2014). The same hypotheses can be developed for the regional regressions, though in both cases there are sites with similar characteristics that perform quite well. Likely, subsurface flows together with climatic changes along the NE-SW direction could deeply influence the final prediction of low flows for either method.

4.2. Guidance for future research

Consistent with previous work by Pugliese et al (2014), a neighbourhood of six donor sites produced the greatest predictive capacity for TNDTK. Different from previous work, application of Top-kriging resulted in rather poor predictions of empirical TND values. NSE associated with cross-validated TND values dropped from 0.81 (see Pugliese et al., 2014, p. 3808) to 0.46. Although it is worth pointing out that we used a different cross-validation procedure in this study, and thus a direct comparison between these two case studies might be flawed. However, analogous to what is presented in Pugliese et al. (2014), poor predictions of TND did not automatically result in poor FDC predictions by using kriging weights resulting from Top-kriging application to the prediction of TND values. This outcome seems to suggest that TND is a rather complex signature of streamflow regime, which is difficult to capture and predict, yet it is highly descriptive in terms of hydrological similarity and future analyses should focus on how to improve TND predictions.

Understanding how to couple and blend the two methods, combining their complementarities is definitely an interesting open question for future research. For instance, future analyses should look at (*i*) the possibility to incorporate a bias correction module, within the TNDTK method e.g. by directly accounting for external drifts associated with geomorphological and climatic characteristics, and (*ii*) different and "duration-oriented" weighting schemes, which might be applied over constrained duration intervals, e.g. an exclusive set of weights for the low-flow regime, one for the high-flow, etc. Moreover, a rather complex issue is finding a comprehensive descriptor capable of expressing the similarity between catchments in terms of FDCs. Future research should address this issue and move towards the delineation of better metrics for quantifying the similarity/dissimilarity between the curves, instead of resorting to signatures such as TND, which evidently provide only a partial description of an FDC.

Finally, this study showed how climatic and geomorphologic patterns could play a significant role in the prediction of FDCs in ungauged basins, thus the practice of dispensing with the delineation of homogenous regions, commonly adopted in geostatistical applications, might be unsuitable for large and very large study areas, as in this study. Although this practice could introduce further elements of subjectivity in the procedure, this feature could be taken into account in future analyses and, furthermore, dealing with prediction of FDCs in a changing environment, might be an interesting research avenue exploring how much the performances rely on the assumption of different across-space climatic and geomorphologic conditions.

5. Conclusions

This study focuses on a comparative assessment of two different methods for the prediction of flow-duration curve (FDCs) at hundreds of ungauged basins in the southeast United States. The first method proposed is an adaptation of Top-kriging, capable of predicting the FDC in a given ungauged catchment by employing a linear weighted average of empirical standardised FDCs, belonging to n donor sites. The prediction is carried out via an empirical dissimilarity index defined as the negative deviation of the FDC from a selected reference streamflow value. The reference streamflow value chosen in this study for standardising each curve is the Mean Annual Flow (MAF), which is a traditional method to standardise FDCs. The second method explored in this study is the regional multiple linear regression method, which has been widely used in several climatic and geomorphological contexts around the world. This approach treats the FDC as a collection of independent streamflow quantiles and predicts every quantiles with a regression based on a different combination of climatic and geomorphologic catchments descriptors.

This study demonstrated that the two procedures perform quite similarly across a broad range of streamflow regimes ranging from yearly floods to droughts. Overall, the regional regression method, termed RR, appeared more robust than TNDTK for very high durations (i.e. severe droughts), showing better performance indices and lower bias, while TNDTK tends to overestimate low-flows.

Still, TNDTK proved to be a reliable procedure for the prediction of FDCs at ungauged sites while simultaneously ensuring its unique

characteristics including that: (*i*) it is able to predict the entire FDC as a single object regardless of the number of points used in a resampling scheme; (*ii*) it preserves the monotonic non-increasing property of the FDC, as a fundamental requirement of cumulative frequency distributions; (*iii*) it works with a limited amount of input data, so that, it only requires a reasonable number of streamflow series and their related catchments' boundaries.

Our comparative assessments have revealed several useful conclusions which hopefully will inspire future research. In particular, the TNDTK approach which only uses streamflow information was shown to be competitive with the regression approach requiring much more information concerning differences among drainage basins. Numerous opportunities exist for improvement of both methods. For instance, our results highlight that drainage area, mutual position of catchments, nested structure and differences in climate all play a significant role in the prediction of FDCs with Top-kriging and regional regression. Our results also highlight the importance of the identification of a more reliable metric than the TND approach employed here, capable of describing similarity among FDCs, which is still a challenging science question to be addressed in the future. Indeed, a more complex conceptualization of the differences, or "distance", between curves, might lead to better and more unbiased performance. Perhaps our most important findings relate to the situations in which these methods performed poorly, which occurred at sites which were either small or very small catchments, characterized by high mean annual temperature, low mean elevation, low run-off ratio and high potential evapotranspiration ratio. Future research might benefit from defining hydrologic homogeneity in those terms to better contrast our ability to estimate FDCs at ungauged sites.

Acknowledgments

The contribution from European Commission FP7 funded research project SWITCH-ON "Sharing Water-related Information to Tackle Changes in the Hydrosphere – for Operational Needs" (grant agreement number 603587) is thankfully acknowledged. The present work was partially developed within the framework of the Panta Rhei Research Initiative of the International Association of Hydrological Sciences (IAHS). Also, we would like to acknowledge A. Liguori and A. Bononi for their contributions to this research work with preliminary analyses and J. E. Kiang for her help providing data and information useful for the realization of the manuscript.

References

Archfield, S.A., Pugliese, A., Castellarin, A., Skøien, J.O., Kiang, J.E., 2013. Topological and canonical kriging for design flood prediction in ungauged catchments: an improvement over a traditional regional regression approach? Hydrol. Earth. Syst. Sci. 17, 1575–1588. http://dx.doi.org/10.5194/hess-17-1575-2013.

Archfield, S.A., Vogel, R.M., Steeves, P.A., Brandt, S.L., Weiskel, P.K., Garabedian, S.P., 2010. The Massachusetts Sustainable-Yield Estimator: A Decision-Support Tool to Assess Water Availability at Ungaged Stream Locations in Massachusetts, U.S Geological Survey Scientific Investigations Report. US Department of the Interior, US Geological Survey.

Berghuijs, W.R., Sivapalan, M., Woods, R.A., Savenije, H.H.G., 2014. Patterns of similarity of seasonal water balances: A window into streamflow variability over a range of time scales. Water Resour. Res. 50, 5638–5661. http://dx.doi.org/10.1002/2014WR015692.

Budyko, M.I., 1974. Climate and Life, International Geophysics Series. Academic Press.

Castellarin, A., 2014. Regional prediction of flow-duration curves using a three-dimensional kriging. J. Hydrol. 513, 179–191. http://dx.doi.org/10.1016/j.jhydrol.2014.03.050.

Castellarin, A., Botter, G., Hughes, D.A., Liu, S., Ouarda, T.B.M.J., Parajka, J., Post, M., Sivapalan, M., Spence, C., Viglione, A., Vogel, R., 2013. Prediction of flow duration curves in ungauged basins. In: Blöschl, G., Sivapalan, M., Wagener, T., Viglione, A., Savenije, H. (Eds.), Runoff Prediction In Ungauged Basins: Synthesis Across Processes, Places and Scales. Cambridge University Press.

Castellarin, A., Galeati, G., Brandimarte, L., Montanari, A., Brath, A., 2004. Regional flow-duration curves: reliability for ungauged basins. Adv. Water Resour. 27, 953–965. http://dx.doi.org/10.1016/j.advwatres.2004.08.005.

Castiglioni, S., Castellarin, A., Montanari, A., 2009. Prediction of low-flow indices in ungauged basins through physiographical space-based interpolation. J. Hydrol. 378, 272–280. http://dx.doi.org/10.1016/j.jhydrol.2009.09.032.

Castiglioni, S., Castellarin, A., Montanari, A., Skøien, J.O., Laaha, G., Blöschl, G., 2011. Smooth regional estimation of low-flow indices: physiographical space based interpolation and top-kriging. Hydrol. Earth Syst. *Sc* 15, 715–727. http://dx.doi.org/10.5194/hess-15-715-2011.

Chokmani, K., Ouarda, T.B.M.J., 2004. Physiographical space-based kriging for regional flood frequency estimation at ungauged sites. Water Resour. Res. 40, W12514.

Cressie, N.A.C., 1993. Statistics for Spatial Data, Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. J. Wiley.

Dingman, S.L., 1978. Synthesis of flow-duration curves for unregulated streams in New Hampshire. J. Am. Water Resour. Assoc. 14, 1481–1502. http://dx.doi.org/10.1111/j.1752-1688.1978.tb02298.x.

Falcone, J., 2011. GAGES-II: geospatial attributes of gages for evaluating streamflow, digital dataset. http://water.usgs.gov/GIS/metadata/usgswrd/XML/gagesII_Sept2011.xml.

Farmer, W.H., Archfield, S.A., Over, T.M., Hay, L.E., LaFontaine, J.H., Kiang, J.E., 2014. A comparison of methods to predict historical daily streamflow time series in the southeastern United States (No. 2014–5231), U.S. Geological Survey Scientific Investigations Report.

Fennessey, N., Vogel, R., 1990. Regional flow-duration curves for ungauged sites in Massachusetts. J. Water Resour. Plan. Manag.-Asce 116, 530–549. http://dx.doi.org/10.1061/(ASCE)0733-9496(1990)116:4(530).

Ganora, D., Claps, P., Laio, F., Viglione, A., 2009. An approach to estimate nonparametric flow duration curves in ungauged basins. Water Resour. Res. 45. http://dx.doi.org/10.1029/2008WR007472.

Gotvald, A.J., Feaster, T.D., Weaver, J.C., 2009. Magnitude and Frequency of Rural Floods in the Southeastern United States, 2006: Volume 1, Georgia (No. 2009–5043), U.S. Geological Survey Scientific Investigations Report. U.S. Geological Survey, Reston, Virginia, USA.

Greene, W.H., 1997. Econometric Analysis. Prentice Hall, New York.

Hollander, M., Wolfe, D.A., 1999. Nonparametric Statistical Methods. Wiley.

Hughes, D.A., Smakhtin, V., 1996. Daily flow time series patching or extension: a spatial interpolation approach based on flow duration curves. Hydrol. Sci. J. 41, 851–871. http://dx.doi.org/10.1080/02626669609491555.

Isaaks, E.H., Srivastava, R.M., 1990. Applied Geostatistics. OUP USA.

Klemeš, V., 2000. Tall tales about tails of hydrological distributions. II. J. Hydrol. Eng. 5, 232–239. http://dx.doi.org/10.1061/(ASCE)1084-0699(2000)5:3(232).

Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I - A discussion of principles. J. Hydrol. 10, 282–290. http://dx.doi.org/10.1016/0022-1694(70)90255-6.

Parajka, J., Merz, R., Skøien, J.O., Viglione, A., 2015. The role of station density for predicting daily runoff by top-kriging interpolation in Austria. J. Hydrol. Hydromech. 63. http://dx.doi.org/10.1515/johh-2015-0024.

Ponce, V.M., Rejendra, P.P., Sezar, E., 2000. Characterization of drought across climatic spectrum. J. Hydrol. Eng. 5, 222–224. http://dx.doi.org/10.1061/(ASCE)1084-0699(2000)5:2(222).

Pugliese, A., Castellarin, A., Brath, A., 2014. Geostatistical prediction of flow–duration curves in an index-flow framework. Hydrol. Earth Syst. Sci. 18, 3801–3816. http://dx.doi.org/10.5194/hess-18-3801-2014.

R Core Team, 2016. R: A Language and Environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Singh, K.P., 1971. Model flow duration and streamflow variability. Water Resour. Res. 7, 1031–1036. http://dx.doi.org/10.1029/WR007i004p01031.

Sivapalan, M., Takeuchi, K., Franks, S.W., Gupta, V.K., Karambiri, H., Lakshmi, V., Liang, X., McDonnel, J.J., Mendiondo, E.M., O'Connell, P.E., Oki, T., Pomeroy, J.W., Schertzer, D., Uhlenbrook, S., Zehe, E., 2003. IAHS decade on predictions in ungauged Basins (PUB), 2003-2012: shaping an exciting future for the hydrological sciences. Hydrol. Sci. J. 48, 857–880. http://dx.doi.org/10.1623/hysj.48.6.857.51421.

Skøien, J.O., Blöschl, G., Laaha, G., Pebesma, E., Parajka, J., Viglione, A., 2014. Rtop: an R package for interpolation of data with a variable spatial support, with an example from river networks. Comput. Geosci. 67, 180–190. http://dx.doi.org/10.1016/j.cageo.2014.02.009.

Skøien, J.O., Merz, R., Blöschl, G., 2006. Top-kriging - geostatistics on stream networks. Hydrol. Earth Syst. Sci. 10, 277–287. http://dx.doi.org/10.5194/hess-10-277-2006.

Stedinger, J.R., 1981. Estimating correlations in multivariate streamflow models. Water Resour. Res. 17, 200–208. http://dx.doi.org/10.1029/WR017i001p00200.

Stedinger, J.R., Vogel, RichardM., Foufula-Georgiou, E., 1993. Frequency analysis of extreme events. In: Maidment, R. (Ed.), Handbook of Hydrology. McGraw-Hill, New York, pp. 18.11–18.66.

Vogel, R., Fennessey, N., 1994. Flow-duration curves. I: new interpretation and confidence intervals. J. Water Resour. Plan. Manag. 120, 485–504. http://dx.doi.org/10.1061/(ASCE)0733-9496(1994)120:4(485).

Vogel, R.M., Fennessey, N.M., 1995. Flow duration curves II: a review of applications in water resources planning. J. Am. Water Resour. Assoc. 31, 1029–1039. http://dx.doi.org/10.1111/j.1752-1688.1995.tb03419.x.