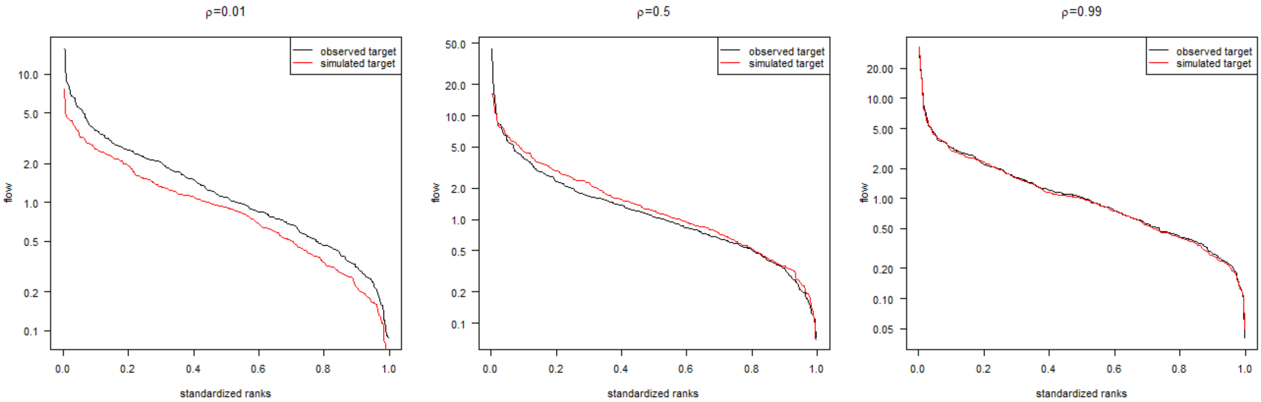
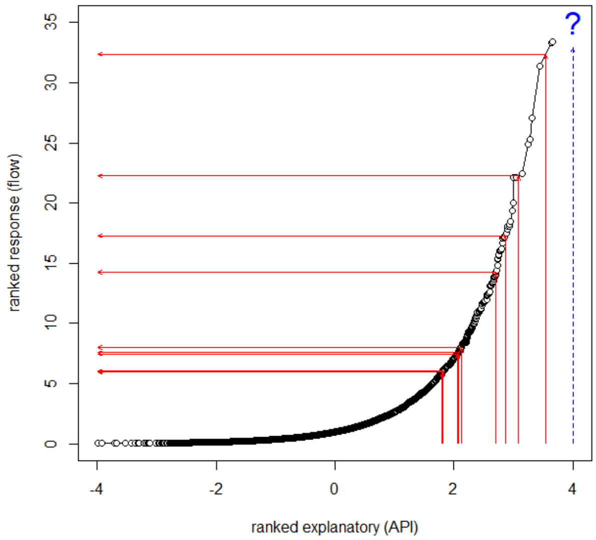
Some further remarks

I broadly agree with the very good review of Dr. Farmer whose remarks pushed me to do some basic experiment with the code I provided in my previous comment. Since the method is a probability matching method widely used in climatology, the correlation between the target variable (flow) and the explanatory variable (API or whatever else) play a role. As shown in the figure below better results are obtained for higher correlation values. The diagrams show random simulations, meaning that better results can be obtained also for low correlation values. The diagram is just an example, but it is representative.



As Dr. Farmer said, in principle we have perfect matching for perfect correlation (comonotonicity). In this respect, we can use API if it is correlated with flow, but also whatever variable that is strongly correlated with the target flow, and has observations at both reference and target period. For example, streamflow at nearby upstream or downstream locations can do the job. In any case, as mentioned by Dr. Farmer and myself, it worth exploring in depth the correlation between flow and API and check how it affects the performance. By the way, some experiments with other variables showing stronger correlation can help. As mentioned, basically, this is a probability matching method: the closer the co-movement of the variables the better the matching is. Moreover, since the performance depends on correlation, is the method really better than some well-devised regression? What about a comparison?

Below, there is also a figure showing an attempt to explain the procedure (it is a q-q plot).



However, I think that the figure 2 of Smakthin and Masse (2000) suggested by Dr. Farmer describes precisely the methodology: new observations of an explanatory variable X are passed through the relationships of ranked X and Y, mathematically y=FY-1(FX(x)) (note that this is the equation used to build q-q plots), where FY-1 is the quantile function (i.e. the inverse of FY), to obtain the unobserved y. As mentioned earlier, this seems to me the most primitive version of the probability matching method used to correct GCM output for "bias". Note that this method is fully general; we can use whatever co-moving X and Y recorded at the same site, at different sites, etc.

Concerning the differences of annual FDCs (or FDCs for sub periods), this is well known and led e.g. Vogel, Castellarin and colleagues to introduce annual FCDs and index-flow FDCs, which attempt to collapse annual FDCs dividing by the annual mean flow values of each year in the spirit of Darlymple's index-flow approach. In general, substantial (multi)annual fluctuations are expected. However, while Dr. Farmer suggests exploring stationarity/nonstationarity, I would suggest starting from temporal dependence, which can provide a good description of such a behavior and is less demanding in terms of assumptions (see e.g. Serinaldi and Kilsby 2015, 2016; Serinaldi et al. 2018; Montanari 2012, Koutsoyiannis' works, among others). In this respect, the correction assumed for the KS test can be insufficient because the effective sample size can be much smaller than one can believe. In fact, streamflows generally show short and long-term persistence going beyond three days (there is extensive literature about this). Even though I am a bit critical on the application of statistical tests, my suggestion is to perform KS tests computing the empirical distribution of the KS statistic under the null by resampling flow values via methods such as IAAFT, which yield stationary processes that preserve marginals and ACF. This way, you can assess if the observed inter-annual/decadal fluctuations of FDCs are consistent with persistence, without introducing arbitrary 'effective' sample sizes. The effect of persistence can be surprisingly striking (e.g. Douglas et al. 2000, Serinaldi and Kilsby 2018; Serinaldi et al. 2018), and you can discover that apparently large differences are large only under too restrictive assumptions.

NSE: to be honest, I never understood very well why hydrologists love this index so much, and why they use it as an index of performance for almost everything. NSE is the similarity index corresponding to the mean squared error (MSE) with benchmark models equal to the average of the time series. Whatever decent model that is able to reproduce seasonality usually gives high NSE values (whose admissible range is also strongly skewed). In my opinion, performance indices should be chosen according to the aim, bearing in mind that they measure different types of discrepancies and some of them are redundant, thus not adding any insight when used together (see e.g. Hyndman and Koehler, 2006; Jachner et al., 2007; Dawson et al., 2007).

Dr. Farmer states 'When talking about general duration curves, more commonly known as cumulative distribution functions, it is better to say "exceedance frequency" rather than "exceedance time"'. Honestly, I do not fully agree. As mentioned in my previous posts, FDCs cannot be considered proper distribution functions. To represent a distribution, the ECDF should satisfy three criterion that are (too) often overlooked (Wilk and Gnanadesikan 1968): '(i) that the order of the observations is immaterial; (ii) that there is no classification of the observations, based on extraneous considerations, which one wishes to employ; and (iii) if the sample is non-random, then appropriate weights are specified.'. Streamflow data does not fulfill these conditions; in fact, the same value in winter and summer has different probabilities of occurrence. If we merge all values, we assign similar probabilities to close values that occur in different seasons with very different probability. To obtain a meaningful distribution, we need to average the distributions of data recorded over periods in which the process can be considered homogeneous (see e.g. Allamano et al. 2011 for an illustration). Therefore, we can use 'exceedance time' or 'exceedance frequency', but please bear in mind that we are not dealing with proper CDFs and probabilities/frequencies having a true probabilistic sense.

References

Allamano, P., Laio, F., and Claps, P.: Effects of disregarding seasonality on the distribution of hydrological extremes, Hydrol. Earth Syst. Sci., 15, 3207–3215.

Dawson C.W., R.J. Abrahart, L.M. See (2007) Hydrotest: a web-based toolbox of evaluation metrics for the standardised assessment of hydrological forecasts Environ. Modell. Soft., 22,1034-1052

Douglas, E., Vogel, R., & Kroll, C. (2000). Trends in floods and low flows in the United States: Impact of spatial correlation. Journal of Hydrology, 240(12), 90–105.

Hyndman R.J., A.B. Koehler (2006) Another look at measures of forecast accuracy Int. J. Forecast., 22 679-688

Jachner S., K.G. van den Boogaart, T. Petzoldt (2007) Statistical methods for the qualitative assessment of dynamic models with time delay J. Stat. Softw., 22 (8) 1-30

Montanari, A. (2012) Hydrology of the Po River: looking for changing patterns in river discharge, Hydrol. Earth Syst. Sci., 16, 3739-3747

Serinaldi F, Kilsby CG. Stationarity is undead: Uncertainty dominates the distribution of extremes. Advances in Water Resources 2015, 77, 17-36.

Serinaldi F, Kilsby CG. (2016) Understanding persistence to avoid underestimation of collective flood risk. Water, 8(4), 152.

Serinaldi F, Kilsby CG. (2018) Unsurprising surprises: The frequency of record-breaking and over-threshold hydrological extremes under spatial and temporal dependence. Water Resources Research.

Serinaldi F, Kilsby CG, Lombardo F. (2018) Untenable nonstationarity: An assessment of the fitness for purpose of trend tests in hydrology. Advances in Water Resources, 111, 132-155.

Wilk, M., Gnanadesikan, R. (1968). Probability Plotting Methods for the Analysis of Data. Biometrika, 55(1), 1-17