Hydrol. Earth Syst. Sci. Discuss.,
doi:10.5194/hess-2016-460-RC2, 2016
© Author(s) 2016. CC-BY 3.0 License.

Hydrology and Earth System Sciences
Discussions

Interactive comment on "The probability distribution of daily streamflow in the conterminous United States" *by* Annalise G. Blum et al.

F. Serinaldi (Referee)

francesco.serinaldi@ncl.ac.uk

Received and published: 18 October 2016

General comments

In this paper, the Authors perform a large scale analysis in order to identify a parametric distribution function providing reasonable approximation for flow duration curves (FDCs) across the conterminous United States. The paper relies on classical "weapons" in the "statistical" arsenal commonly applied in hydrology (L-moments, Nash-Sutcliffe performance index, linear regression in logarithmic space for regionalization, etc.). So, taking for granted that such tools are sound and correctly applied, the interest in this paper is not surely methodological, but concerns the empirical re-

sults. Considering that downloading data and analyzing them with R packages such as `lmom`, `lmomco` and some build-in regression functions is a matter of few hours (most of which are needed to slightly customize the default diagrams yielded by R), in my opinion, it is a bit hard to classify this kind of works as research papers. My very personal opinion, is that they can be at most technical reports or case studies (likely resulting from some master thesis). Anyway, I leave the paper classification to the Editors; from my side, I can only say that I cannot see significant insights, while there are some inconsistencies resulting in misleading conclusions. Just to make an example, there are works providing L-moments for gridded rainfall worldwide with quite limited insight about the nature of rainfall (e.g., Maeda et al., 2013), while others (e.g. Papalexiou and Koutsoyiannis, 2016) make similar analysis on gauged data but considering distribution families derived by entropy maximization, introducing a new test for seasonal variation, and providing a number of new insights. What I mean is that we can analyze a large data set "passively", by running e.g. R codes quite blindly, or we can decide to use data in order to understand the underlying processes in more depth. Said that, if we accept the first approach, the paper is ready to publish, once removed some nonsense discussed below (concerning the comparison of MA-FDC and POR-FDC); in the second case, we are far from a good quality work. In any case, I would like to use this opportunity to share my point of view on flow duration curves (FDC), stressing that the philosophy behind them probably needs some rethinking.

Specific comments

The Authors stress twice in the text that FDC (actually POR-FDC) "ignore the important serial stochastic structure of daily flows, including such issues as autocorrelation and seasonality" and they also recognize that simple 3- or 4-parameter distributions can only approximate FDCs. These statements are given in passing, but they are actually the core of the problem. In principle we can get whatever time series of numerical values, arranging it in ascending (descending) order and then plotting the sorted values against their rescaled ranks. Irrespective of the nature of the (numerical) data, the

result is always a monotonic pattern describing the function $g : \mathcal{R} \rightarrow [0, 1]$ (of course, the domain can be a subset of $\mathcal{R}$, and the function is strictly monotonic if there are not statistical ties (i.e. identical values)). If the aim is to fit a simple analytical function to such curves, theoretical cumulative distribution functions (CDFs) seem to be natural candidates. However, CDFs are not simple curves useful for fitting data, but represent the nonexceedance probability of a random variable and work if the data are independent and identically distributed (*iid*). All these concepts are trivial and the Authors know them better than me. However, since daily stream flow records surely do not fulfill any of these conditions, why should a single distribution fit FDCs? In other words, in spite of the efforts made along the years to find suitable CDFs for modeling FDCs, the problem is ill-posed by definition: even the most parametrized CDFs cannot mimic FDCs unless the flow series is characterized by strong mixing (e.g. weak seasonal pattern compared to non-seasonal (essentially "random") fluctuations). So, if the FDCs analysis reduces to a simple exercise of curve fitting, the overall analysis performed in this type of studies can make sense; otherwise, if the aim is to fit a CDF, and then concluding that such model describe probability of (non)exceedance or something like that, this statement can be much more problematic, unless the model is a mixture of CDFs describing data approximately 'identically distributed' (id) such as seasonal or monthly subsets. In fact, the analysis reported by e.g. Basso et al. (2015) is performed on a seasonal basis.

More generally, as the Authors know, stream flows are characterized by two properties that play a fundamental role in this context: seasonality and persistence (often long range persistence; see e.g. Montanari et al. (1997,2000) or more recently Serinaldi and Kilsby (2015)). Seasonality is often the main source departure from *id* condition. This is well known for instance in rainfall modeling where simple 2-parameter Weibull distributions are surely insufficient to describe daily rainfall over the entire year, but their performance is very good if we introduce parameters varying with the seasonality. Indeed the fact that stream flow values can cover two or three orders of magnitude simply depends (obviously) from the alternation of high-flow and low-flow seasons, in which

the *id* hypothesis is far from being realistic. On the other hand, long-range dependence results in inter-annual variability, which is what the index-flood method attempts to take into account in quite a naïve way. However, the index-flood still overlooks the problem of non-*id* conditions within calendar or water year. When the seasonal signal is strong, this can be the main reason for the lack of fitting of simple parametric distributions, and index-flood cannot improve the fitting very much. Moreover, while seasonality impacts on the overall shape of flow distribution (imagine to mix e.g. 12 different distributions, each reproducing approximately *id* monthly flows), long range dependence induces inter-annual fluctuations that impact especially on the tails. Therefore, the index-flood method adjusts more easily tail behavior than the overall shape of the parametric FDC.

The above remarks, can help to understand how to improve FDC if we want to avoid physical approaches *à la* Botter (...but overlooking physical arguments is never a good choice) and keep the model purely statistical, but a little bit more coherent with the nature of the data. The easiest approach is surely splitting data at e.g. seasonal scale. On the other hand, we can build on the fact that the regionalization procedure commonly applied in hydrology (and summarized in this study) is only a rough and naïve version of generalized linear/additive models (GLM/GAM an their extensions) $f(y; \theta(\mathbf{X}))$, where f is the distribution of flows Y, θ is a vector of parameters (e.g., the three parameters of the Generalized Pareto) and $\mathbf{X}$ is a design matrix of covariates (e.g., the variables in Eqs. 7-9). In this framework, seasonality can easily be introduced by simple sine and cosine functions describing the seasonal cycles; since a couple of waves are generally sufficient to describe the seasonal flow regime, GLMs imply only a couple of additional parameters. Alternatively, a factor index can be used in the fitting procedure to distinguish e.g. between the four seasons or the 12 months. In all cases, the resulting model not only account for the spatial variability but also for the non-id conditions by a few additional parameters that have a clear physical interpretation (they represent the seasonal regimes across the area of study). Of course, the usual graphical representation (as in Fig. 1) is possible only if we compare observations and simulations because such a diagrams merge quantiles coming from a set of distributions (devised

for id data), roughly speaking one for each season (or month). However, this is not surprising because the observed FDCs themselves incorporate values coming from different (seasonal) distributions, thus explaining the lack of fiit of simple models. This approach also helps overcoming the problem of MA-FDC simulation mentioned in the paper. Notice that the effect of seasonal variation as well as long range dependence can be recognized in Figs. 7(a-b) and 8(b-c) in the form of multimodality, while the stepwise pattern in some regions of the FDCs in Fig. 7(a) and 8(a) denotes the presence of statistical ties, which generally results from limits in the resolution of measurement devices or round-off procedures. The first aspect denotes the intrinsic inadequacy of whatever classical unimodal distribution, while the latter often affects estimation procedures (so, I'm not so surprised about the poor fitting). In this respect I have to say that the scale of the x-axis does not help fitting assessment. I'm a bit surprised because after Vogel and Fennessey (1994), we know that stretched axes enhancing the linearity of FDCs and CDFs allow much better assessment, in agreement with recommendations available in the literature on visual perception and data visualization (see e.g., works by Tufte, Cleveland, etc.).

Another concern is about the comparison of POR-FDCs and MA-FDC. The Authors conclude that fitting MA-FDCs is easier and more reliable than POR-FDCs as "prediction of POR-FDCs was less consistent" (consistent?). The comparison between MA-FDCs and POR-FDCs is ill-posed by itself and in the interpretation of NSE. Firstly, for MA-FDC, we always fit a CDF on 365 data points, where each one is the median (or mean) of a set of M values, where M is the number of years (here 40-60); for POR-FDCs we are trying to fit a CDF on $365 \cdot M$ values (i.e. a sample 40-60 times larger), where each values (order statistics) should be the point estimates of the corresponding quantiles. In the first case, we seek the fitting in the range of probabilities $\left(\frac{1}{365+1} \approx 3 \cdot 10^{-3}, \frac{365}{365+1} \approx 0.997\right)$, whereas in the second we pretend to fit quantiles corresponding to probabilities between $\frac{1}{365M+1} \approx 5 \cdot 10^{-5}$ and $\frac{365M}{365M+1} \approx 0.99995$. So, is it so surprising that fitting a curve on 365 "smoothed" values (medians) is easier than

on 18250 values (being already aware that such values cannot come, by definition, from a unique distribution)?

Secondly, the above remark allows some reflection on the (mis)use of performance metrics and their interpretation. As for every performance index (absolute metrics, relative errors, deviance or similarity measures, information criteria, etc.), NSE (which is simply the similarity index corresponding to the mean squared error) is devised to compare the performance of a set of models for the ***same*** data set; in our case, not only the sample size of the data sets and error terms is completely different (365 against about 18250), but also the nature of the data is completely incomparable (raw data against medians resulting from a very specific selection procedure). Thus, stating that NSE for MA-FDC is generally smaller than that of POR-FDCs is nonsense, as we are comparing apples with pears. Moreover, even though I know that hydrologists have fallen in love with NSE for some esoteric reason, I would like to stress that a performance index should be chosen according to the particular type of discrepancy one wants to highlight, and not because it is popular. To be more specific, NSE is a similarity index comparing the errors from the selected model (numerator) with those from a benchmark or reference model (denominator), where the reference model is, in this case, the sample average (aka 'reference climatology' in climatological literature or "naïve" reference in forecasting literature...it seems that people in each discipline like renaming the same concepts many times, just to increment a little bit the already widespread confusion...). The choice of this "naïve" reference has two consequences: (1) the range of possible NSE values is strongly asymmetric, and (2) every model more complex than the simple average easily yields relatively high NSE values; this is usually interpreted as a good performance, but actually it is not, because the way NSE values populate the range $(-\infty, 1)$ is strongly nonlinear. Since the average is not a sufficient statistics even for data coming from a Gaussian distribution, it is easy to recognize that whatever model provides great improvement and (relatively high NSE) compared to such "naïve" reference. Therefore, sentences such as "Despite this comparable fit, the NSE coefficients are quite different: 0.89 for POR-FDC GPA3 versus the much

higher 0.96 for MA-FDC GPA3. This discrepancy reflects a challenge in the use of the metric and indicates why visual inspection of FDC plots is particularly important for understanding overall GOF", make little sense because (1) the two values refer to different data sets (comparisons can be done only between at-site and regional models for the same data set, MA and POR, respectively), and (2) even if they referred to different models for the same data set, NSE is not equipped with criteria allowing to say if the difference between two values is significant or not (unlike methods based on maximum likelihood and/or information criteria). Concerning the rationale, choice and interpretation of performance measures please see Dawson et al. (2007), Hyndman and Koehler (2006), Jachner et al. (2007), Burnham and Anderson (2004), Reusser et al. (2009), among others.

Technical remarks

Please use homogeneous notation: "2-,3-,4-parameter distributions" or "two-,three-,four-parameter distributions" throughout the text.

P3L16: it can be worth citing Doulatyari et al (2005), Basso et al. (2015), and Schaefli et al. (2013)

P6L10-15: the Authors refer to other quantile estimators; however, Weibull plotting position is not a quantile estimator. In this respect , it can also be worth having a look at Makkonen (2006), and Hutson (2000)

P7L8: "Hosking and Wallis 1997"

P7L16: "natural logarithm"

P7L16: "linear combination of order statistics" can better reflect their actual rationale (linear combination with weighted moments is a consequence)

P8L16: "see e.g. Rianna et al. (2011) and references therein"

P9L10-15: I may have missed something, but I cannot see where the effect of sample

size on L-moment scattering is shown. Moreover, the similarity between L-moments of POR-FDC and MA-FDC (Fig. 3) are likely due to the fact that L-moments are less sensitive to tail behavior by definition. Since the body of POR-FDC and MA-FDC are similar, L-moment ratios are similar.

P9L25: I understand the attitude of simulating everything, but sometimes it is not strictly necessary; in this case, we already known that daily stream flows cannot be distributed as GPA (or whatever else common unimodal distribution) because they are non-iid.

P11L5-25: concerning the simulation of MA-FDC, I think you can do an attempt by fitting the standardized annual sequences (dividing each year of data by annual median), then simulating from this distribution, and multiplying each simulated block of 365 elements by resampled (bootstrap) values of annual medians. This way, you do not need to fit any distribution for annual medians, but can explore the effect of inter-annual variability (see discussion above). Concerning the index-flow method, please note that my paper highlights a problem that is probably more serious than the uncertainty of the distribution of annual medians. Actually, the analysis of confidence intervals highlighted that the distribution of the product $Q \cdot X'$ is not coherent with the idea of constant median over blocks of 365 elements, because the distribution of the product of two random variables implies the product of independent realizations, whereas the annual (constant median) introduces redundancy. This is one of the reasons why I think that common FDC frameworks, even if simple, actually provide very rough approximations, and should be replaced by more coherent methods, even if this means to loose this (perhaps excessive) simplicity.

P14L5-10: see comment above about a more careful choice of graphical properties, axes scales, etc.

P16L1: which is the physical interpretation of GPA scale and shape?

P17L4: "Table C2"

Appendix A: I understand the rationale of this discussion, but the comparison between index-flood models and classical FDCs should be done in terms final output. As mentioned above, index-flood is an attempt to account for one of the key aspects of stream flows (inter-annual variability or long range fluctuations and persistence if you prefer). It suffers some statistical inconsistencies (that may be overcome by moving from the distribution of a product to e.g. compound distributions, but this needs to be explored) and does not account for the second and perhaps more important aspect, i.e. non-*id* intra-annual conditions. This is just to say that the problem goes slightly beyond distribution fitting, but requires a more careful consideration of the nature of data and underlying process, otherwise it reduces to what Vit Klemes called "dilettantism in hydrology" (i.e. replacing physics with (often misused) statistics).

Sincerely,

Francesco Serinaldi

References

Basso S, M Schirmer, G Botter (2015) On the emergence of heavy-tailed streamflow distributions Advances in Water Resources 82, 98-105

Burnham K. P. and Anderson D. R. (2004) Multimodel Inference: Understanding AIC and BIC in Model Selection, Sociological Methods Research 2004; 33; 261

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

Doulatyari B, A Betterle, S Basso, B Biswal, M Schirmer, G Botter (2015) Predicting streamflow distributions and flow duration curves from landscape and climate Advances in Water Resources 83, 285-298,4

Hutson, A.D. (2000) A composite quantile function estimator with applications in boot-

strapping, Journal of Applied Statistics, 27 (2000), pp. 567–577

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

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

Maeda, E. E., Arevalo Torres, J. and Carmona-Moreno, C. (2013), Characterisation of global precipitation frequency through the L-moments approach. Area, 45: 98–108. doi:10.1111/j.1475-4762.2012.01127.x

Makkonen L (2006), Plotting positions in extreme value analysis, Journal of Applied Meteorology and Climatology 45 (2), 334-340

Montanari, A.; Rosso, R.; Taqqu, M.S. (1997) Fractionally differenced ARIMA models applied to hydrologic time series: Identification, estimation, and simulation. Water Resour. Res., 33, 1035–1044.

Montanari, A.; Rosso, R.; Taqqu, M.S. (2000) A seasonal fractional ARIMA model applied to the Nile River monthly flows at Aswan. Water Resour. Res., 36, 1249–1259.

Papalexiou S.M., and D. Koutsoyiannis (2016), A global survey on the seasonal variation of the marginal distribution of daily precipitation, Advances in Water Resources, 94, 131–145, doi:10.1016/j.advwatres.2016.05.005.

Reusser, D.E., Blume, T., Schaefli, B., Zehe, E., (2009). Analysing the temporal dynamics of model performance for hydrological models. Hydrol. Earth Syst. Sci. 13, 999–1018.

Rianna, M., Russo, F., and Napolitano, F. (2011) Stochastic index model for intermittent regimes: from preliminary analysis to regionalisation, Nat. Hazards Earth Syst. Sci., 11, 1189-1203, doi:10.5194/nhess-11-1189-2011

Schaefli B, A Rinaldo, G Botter (2013) Analytic probability distributions for snow-

dominated streamflow Water Resources Research 49 (5), 2701-2713

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

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-460, 2016.