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 results. 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).

The reviewer’s comments are very far from the truth concerning the effort involved in an empirical study of this type. The ‘empirical analysis’ which was performed in this paper began as empirical analyses in Figures 2 and 3 in the paper by Vogel and Fennessey (1993) which only examined the pdf of daily streamflow for 23 sites in Massachusetts. To date, the paper by Vogel and Fennessey (1993) has 346 google scholar citations, and arose entirely from the empirical challenge described in the manuscript under consideration by HESS. That empirical work was later extended by Archfield (2009) to the New England region, but was never published. Furthermore, the process of fitting a probability distribution to daily streamflow observations led to numerous challenges associated with zero occurrences, and obtaining plausible estimates of the lower bound of daily streamflow. It is only through such empirical studies that researchers can become aware of many of the very practical (and perhaps mundane to the reviewer) challenges and concerns in hydrology.

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.

We welcome the detailed and insightful comments on our work, however, we note a tremendous gap between the personal thinking on this problem of this reviewer and the vast literature on FDC’s.

In spite of the fact that daily streamflows are certainly not independent, and probably not identically distributed, FDC’s have found widespread use in practice, and numerous studies have shown that the assumption of a fixed pdf for daily streamflow can lead to improvements in our ability to estimate streamflows at ungaged sites and numerous other applications. Thus, as a practical matter, in spite of the fact that daily streamflows may not really arise from an iid process, there are many advantages of making the empirical assumption of an identically distributed process.

We encourage theoretical work by the reviewer and others to explore the possibility of modeling the probability distribution of daily streamflow as a mixture of several pdfs, however we remain quite confident that our empirical explorations of the distribution of daily streamflow may be quite useful in practice, particularly for the case of prediction in ungauged basins. It is this motivation - prediction at ungaged basins - that motivated this analysis, as the search for parsimonious solutions to fundamental problems in hydrology will always remain of use in our field. Furthermore, we note that this particular paper is our first attempt to summarize more recent investigations mentioned above which are the result of several decades of research on this topic. Thus we take great exception to numerous

comments of Serinaldi when he appears to denigrate the empirical nature of work on FDC's.

We appreciate the recommendation of the author concerning Papalexiou and Koutsoyiannis (2016) which serves to further motivate the importance and relevance of our work. We will add a discussion of Papalexiou and Koutsoyiannis (2016) to our revised paper.

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 : R ! [0, 1] (of course, the domain can be a subset of 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).

We fully agree with the reviewer that daily flow series are neither independent nor identically distributed, and thus a search for a single CDF for modeling FDC's may be ill-posed, in a theoretical sense. However, there has been a continuing discussion of exactly the same issues concerning fitting of a single frequency distribution to flood series which often arise from many different (non-identical) physical mechanisms such as cyclonic, frontal and other meteorological processes. Yet it is still common practice to make the iid assumption for flood series. Similar arguments could be made for stochastic models of precipitation, drought and other hydrologic variables. Nevertheless, we agree with the reviewers concern, especially for daily streamflow which are far more complex than flood series. Thus our revised manuscript will include a section which both discusses the ill-posed nature of this important practical problem, and points to numerous approaches for deriving a composite distribution of daily streamflow based on a mixture of various processes, and we will review the literature to ensure that our work is put in the proper context. That is, this is the first comprehensive analysis of the ability of a single CDF to represent FDC's, thus it makes sense for us to summarize our results and to provide a broader context for the problem to enable others to follow up on our findings. The manuscript already provides a very brief discussion of a few studies which have sought to derive a distribution of daily streamflow from a physically based watershed model. We will use those studies to frame the 'ill-posed' nature of the problem and to provide direction for future research.

In the reviewer's opinion, the ill-posed nature of our problem is both a settled and damning hypothesis; however, this is not the case within a broader scientific context. For the problem under consideration, a parsimonious, yet empirical approach to the estimation of the FDC can have profound practical implications for estimation of FDC's and even for estimation of daily streamflow series at ungaged locations. This has been shown in numerous previous studies and the textbook "Runoff Prediction in Ungauged Basins" devotes an entire chapter to predicting of FDCs at unagaged locations.

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.

This comment reflects a possible misunderstanding of our work as being only a "curve-fitting exercise." We feel the reviewer's confusion may arise from the fact that we have not explicitly made clear the motivations for our work which involve numerous applications of FDC's which stem from the findings of this paper. Although we note this in the introduction (lines 10-11), the revised manuscript will address this issue more fully. For example, there are numerous applications of FDCs that require a complete analytical model of the FDC to implement. A common approach for estimation of daily streamflow series at ungaged sites is based on transfer of streamflow information via the CDF of streamflow from a gaged site to a nearby ungaged site. This method, has found to be a significant improvement over numerous other methods for estimation of time series of daily streamflow at ungaged sites, yet it depends on the assumption of a single pdf of streamflow. For an example of one of the first applications of this approach, using a lognormal distribution of daily streamflows, see Fennessey and Vogel (1990). We will also clearly acknowledge in the revised manuscript that our contribution is largely of a practical nature and may not satisfy one's theoretical scientific curiosity relating to the true underlying probability distribution of daily streamflow.

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 reviewer raises numerous constructive points. Our revised manuscript will include an analysis which evaluates the degree to which breaking up the year into seasons can be used to improve our ability to select and model the probability distribution of daily streamflow. In addition we will add a general discussion of these issues and will point to studies which have considered these issues for analogous problems such for flood and drought problems. These comments and the seasonal analysis described, will be made in an effort to steer the reader in a direction for improvements in future studies.

However, we still feel our analysis, which assumes iid conditions, (or possibly breaks up the year into seasons), is a reasonable assumption for this initial paper, given that (1) a single pdf is now used widely in regional FDC studies and (2) this is exactly the assumption which is made in practice for a very wide range of other hydrologic problems, which the reviewer mentions, including rainfall over a wide range of temporal scales, and of course floods and low flows. In other words, a solution which may not be scientifically correct is widely employed in the interest of obtaining parsimonious models which, to a first approximation, can be used to solve a very wide range of practical problems.

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; _(X)), where f is the distribution of flows Y , _ is a vector of parameters (e.g., the three parameters of the Generalized Pareto) and 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).

We agree that this is a useful comment and will revise our manuscript as described above. We will examine the impact of breaking the year into seasons to examine regional differences in our ability to fit a distribution to the observations. We suspect that when we break the year into four seasons, that a two-parameter Generalized Pareto distribution will fit nicely in all four seasons, resulting in four GPA distributions for a total of eight parameters. Such an analysis would be a very nice addition to our paper and would extend our conclusions and provide others with many new opportunities for research in the future involving mixtures of distributions for daily streamflow.

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 fit of simple models.

Figure 1 is simply illustrating annual flow duration curves along with their median and mean annual counterparts. Such figures are now nearly ubiquitous in hydrology, and their interpretation is quite useful in practice regardless, or in spite of, the comments of the reviewer.

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.).

Figures 7 and 8 both employ a logarithmic scale for streamflow on the y axis, but the reviewer is apparently suggesting the use of a 'stretched axis' for the exceedance probabilities using perhaps the inverse of a normal quantile, so that the resulting plots become equivalent to quantile-quantile lognormal plots. We elected to use the more common approach which plots the logarithms of streamflow versus exceedance probability using an ordinary arithimetic scale, because this is by far the most common approach to the graphical illustration of FDC's in practice.

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
$\frac{1}{365+1}$ _ $3 \cdot 10^{-3}$, $\frac{365}{365+1}$ _ 0.997
_ , whereas in the second we pretend to fit quantiles corresponding to probabilities between $\frac{1}{365M+1}$ _ $5 \cdot 10^{-5}$ and $\frac{365M}{365M+1}$ _ 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)?

We agree with the reviewer that comparisons between the goodness of fit of a pdf to POR-FDC and MA-FDC's is problematic due to the reasons outlined above. However, MA-FDC's are used widely for problems in which ones interest focuses on streamflow conditions in a typical or atypical year, thus it is very important for us to consider this case in our paper. The revised manuscript will include a detailed discussion of the issues raised by the reviewer and will drop all comparisons of goodness of fit between MA-FDC's and POR-FDC's and treat them as separate problems.

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 (−1, 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.

We agree with the comments of the reviewer and, as a result, there will be no comparisons of the goodness-of-fit between the MA-FDC's and the POR-FDC's due to the reasons outlined and the revised manuscript will make this point very clearly. The use of NSE, a standardized form of mean square error, is perhaps the most commonly used goodness-of-fit metric in hydrology. We will only report log space values of NSE to deal with the fact that this goodness of fit statistic has very poor sampling properties when used with highly skewed samples as is the case for daily streamflow when NSE is computed in real space. We will continue to report these values for each case to enable comparisons of the goodness-of-fit of either MA-FDC's or POR-FDC's.

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

Thank you for these technical comments, we will address them in the revised paper.

References Cited:

Archfield, S.A., 2009, Chapter 2 – The Probability Distribution of Daily Streamflow, in: Estimation of continuous daily streamflow at ungaged locations in southern New England, PhD Dissertation. Tufts University.

Fennessey, N. and R.M. Vogel, Regional Flow Duration Curves for Ungaged Sites in Massachusetts, ASCE, Journal of Water Resources Planning and Management, Vol. 116, No. 4, pp. 530-549, 1990.

Vogel, R.M. and N.M. Fennessey, L-Moment Diagrams Should Replace Product-Moment Diagrams, Water Resources Research, Vol. 29, No. 6, pp 1745-1752, 1993.