Anonymous Referee 1

The paper fits theoretical distributions to a large dataset of empirical streamflow observations covering the conterminous US. The study finds that median annual flow duration curves (FDC), which portray flow distribution in a typical year, can be reasonably fitted to three-parameter distributions. In contrast, period of record FDC that incorporate extreme streamflow variations over numerous years cannot be appropriately fitted, even to more complex theoretical distributions. The authors explore the implications of that finding on predictions in ungauged catchments using linear regressions in case studies. Predicting streamflow signatures (particularly FDC) in ungauged basins is both extremely useful and challenging and the findings of this study are interesting and important, particularly the insight that MAFDC might be both easier to predict and more practically relevant than PoRFDC. However, there are two points that I would like to see further discussed before publication, as well as the few minor comments listed below.

First, the study is an impressive effort to fit FDCs to a very large dataset of unregulated catchments – this is definitely a key contribution of the paper. However, by covering the whole conterminous US, the dataset covers a wide variety of climates, catchment characteristics and flow regimes, and it would have been interesting to explore how the fit to specific distributions varies regionally. The shape of FDCs depicts the local flow regime, which are themselves related to climate and catchment characteristics (see e.g., Botter 2013). It would be nice to see whether there is a link between flow regimes, climate/catchment characteristics and the best fitted theoretical distribution. It would also be nice to discuss how the best-fit distributions relate to the distributions that might be expected from process-based models (Botter 2007, Botter 2009, Muller 2014, Muneepeerakul 2010, etc), given the dominant flow processes in particular catchments.

We appreciate these excellent suggestions. In the revised manuscript, we will add an analysis that assesses the fit of the FDCs within each of 19 major hydrologic regions of the United States to supplement the nationwide results. We also intend to evaluate how the goodness of fit of FDC models relate to various climate and catchment characteristics to enable a better understanding of those situations which are most challenging to characterize FDCs.

Second, while I appreciate the effort to extend an already complex and large scale analysis to prediction in ungauged basins, I would like to see more details on how the regression models were obtained (i.e. how the regression covariates were selected for Eqn 7-9), and a discussion on whether these regression models have a physical interpretation. Specifically, I am concerned about using linear regressions to estimate distribution parameters, which arguably have a more ambiguous physical interpretation as moments. Mean flow (first moment) for instance can be argued to be a linear combination of observable characteristics like mean rainfall, as per the water balance equation. The issue in regressing GPA3 parameters is that they are not linear combinations of the moments of the distribution, so using linear regressions to estimate the parameters does not allow moments to be linearly related. In other words, in this spe-

cific case, linearly regressed parameters are not compatible with a linear water balance relation on mean flow. To address this issue, please either apply linear regressions on the moments of the distributions instead of the parameters (and discuss the physical interpretation of the linear models when appropriate), or make the case that Eqn 7-9 are not incompatible with water balance principles.
[To illustrate my point on linear regressions, let's assume the simplest linear model possible, where predictions are simply taken as the mean of the observed sample (this can happen in the specific case of the water balance model above if all catchments have an identical mean rainfall). Let's say that we have a sample of three catchments with the following GPA3 parameters and mean flow (computed from the parameters):
Basin | location param | scale param| shape param | Mean
1 | 0 | 100 | 0.3 | 143
2 | 0 | 1 |0.05 | 1
3 | 0 | 20 | 0.1 | 22
The predicted mean flow in a fourth catchment obtained from the observed mean flows (i.e. by taking the mean of the mean) is 47, whereas the mean flow computed from predicted GPA3 parameters (i.e. computed from the mean value of each parameter) is 55.]

We agree that assuming that the GPA model parameters are independent of each other is a poor assumption. Based on this comment, as well as several comments below along with concerns raised by other reviewers, we have decided to remove this case study from the revised manuscript and add to the manuscript a deeper exploration of the regional and seasonal behavior of the goodness of fit of FDC models, in addition to our national analysis.

Minor comments:
p5 l 17: please define GOF.

We will add this definition.

p12 l14-19: I have seen this issue most often addressed by taking the logarithm of flow quantiles before computing the NSE. Is there are reason why you preferred the selected approach?
We agree that taking the logarithm of the flow quantiles before computing the NSE is a preferred method. We had originally employed NSE in real space due to the occurrence of zero flow values and, therefore, we could not take logarithms of the flows. However, we subsequently removed those streamgages with zero flows, but kept the original reporting of the goodness of fit. Thus, your point is valuable and in the revised manuscript, we will report the NSE values of the logarithms of the streamflow.

p14 l13-15: I agree that modelling errors on FDCs are best assessed graphically and appreciate the effort of showing fits for particular basins with low, median and high NSE. Error duration curves (e.g., Muller 2016) are great way of visualizing performance fits over large samples (as opposed to individual basins), and it would be informative in my opinion to display the relevant EFDCs for the whole dataset.

Thand you for this useful comments. We will produce this graphic in our revisioins and explore the value of adding it to the manuscript.

p.16 l3-4: I realize that concerns of overfitting regression models are to some extent addressed in the LOO cross validation analysis, but please display covariates statistics (e.g., quartiles and range or boxplot) to show that there is enough variability in the samples to credibly argue that the LOO performance is externally valid.

This is a good point, however this Leave One Out Analysis will be removed along with the case study.

p.17 l2-4: Have you tested for serial correlation? Serial correlation affects the estimation of OLS standard errors and are particularly likely to occur between flow-connected gauges (i.e. gauges located on the same river).

We thank the author for pointing out this question; we had not examined if there are nested basins in our dataset that may affect the OLS standard errors. Nevertheless, we are removing the case study now from the analysis.

p.18 l.11-20: The paper makes the great case that MAFDCs are easier to fit and have more practical relevance than PoRFDC. However, I would appreciate a more complete discussion of the tradeoff involved: mAFDC loses information on inter-annual variability, hence their better fit to "simpler" distributions. This is an important caveat that has strong implications for practical applications and should be made clearer in the discussion/conclusion in my opinion.

We will emphasize this important distinction between MAFDC's and PoRFDC's in the revised manuscript.

Table C1: I have trouble understanding how the BFI_AVE is a regression covariate for prediction in ungauged basins. My understanding is that flow observations are necessary to compute the BFI in the first place.

The BFI_AVE is available as a grid generated by inverse-distance-weighting of base-flow index values computed at USGS streamgages for the entire United States (Falcone et al., 2010). Nevertheless, we are removing the case study from the analysis.

Falcone, J. A., D. M. Carlisle, D. M. Wolock, and M. R. Meador (2010), GAGES: A stream gage database for evaluating natural and altered flow conditions in the conterminous United States, *Ecology*, *91*(2), 621–621, doi:10.1890/09-0889.1.