Summary and Overall assessment

The study aims to propose a new statistical technique to improve the predictive power of multiple linear regression models used for streamflow predictions in ungauged basins. The contribution of the technique, in essence, is based on defining new (dis)similarity indices—called point-to-point, lateral, and vertical distance measures—applied to monthly flow regimes. They have used monthly flow regimes, instead of using flow duration curves (FDCs), to compare the ungauged basin with its neighboring basins arguing that FDCs loss the time aspect of hydrologic data information content. Further, they seek to improve the flow estimation in the ungauged by "swapping" the original model (used for flow estimation) by other models *if* certain conditions are met; namely (1) increasing the covered space around the ungauged basin, and (2) reducing the overall model error. Authors developed their regression models based on climatic and geomorphologic descriptors of neighboring basins, and applied their framework across 124 basins in northwestern Italy.

In my evaluation the manuscript is **rejected**, with the kind encouragement of submitting a substantially revised version. The fundamental requirement of reporting a scientific research is not met, i.e. its clarity and testability/reproducibility. The wording and phrasing on many instances within the manuscript are short of a scientific text; needless to mention that the manuscript is not even proof-read. The manuscript is filled with unsubstantiated, poorly defined, and vague statements—instances are discussed within the section below. Parts of the discussions lack logical coherence and internal consistency (comments 2.f and 2.g). The work is not properly contextualized within the extant literature. The descriptors used in developing original models are not explained properly, and the results discussion lack a discussion on the uncertainties associated with the data used, developed method, as well as the discussion of results.

I encourage the authors to ask one or few experts in regionalization as well as a native English speaker to review their manuscript thoroughly before the re-submission.

Major comments

1. The manuscript is poorly structured.
   a. Parts of the Introduction (lines 64-82) are in fact methodology. Within the Introduction is important to discuss "why" your work matters and "why" the method is a contribution. That said, "how" the method works, should be discussed in the Methods section.
   b. The objective of the work is not clearly stated in the Introduction, instead mentioned at the end of the Study Area (lines 103-105). Further, you mentioned your primary focus is "accurate prediction". You need to exactly define what you mean by accuracy. If by accuracy you mean reducing bias, then you need to present the values of the models' biases to demonstrate the improvement (e.g. on table 3).

1

c. The manuscript does not have a Methods section. To better elaborate and emphasize your proposed method, I think it is essential to put sections 3-5 under an overarching section of Methods, consolidate them (some parts are hard to follow) and remove repetitions, and perhaps before section 2 Study Area. Further, the procedure explained within lines 21-221 (including the figure) is hard to follow. Please rephrase and explain more clearly.

d. Lines 93-95: you listed "some of the descriptors (out of 74 descriptors)" on table 1. Why did you choose these descriptors? Are these the ones that you used in developing your regression models? You did not elaborate which descriptors are used for model development (see also lines 145-150). You can possibly list all the descriptors as supplementary material.

2. The literature is not cited representatively nor discussed properly (lack of proper contextualization). Many sentences are vague and open to interpretation, with poorly defined terms particularly remote from the common hydrologic modelling terminology. Further, some arguments/discussions lack logical coherence. Here are a few examples—but not limited to this list.

a. Take the first 2 sentences of the Introduction. In the first sentence you discussed "the prediction of flow regimes <u>in general</u>", but in the next sentences of the paragraph you are specifically and only talking about prediction in ungauged basins. In the second sentence you mentioned "widely studied" yet you only cited 3 papers. Also, stated "prediction of hydrological data"; we do not *predict* 'data', we observe/estimate/measure them. Moreover, the methods you are referring to are only for predicting streamflow (and not other hydrological variables) in ungauged basins (and not in general).

b. Further, this very first paragraph of the Introduction, particularly lines 49-63, is filled with unsubstantiated fact-like claims. Many sentences such as lines 49-50 and 58-60 need referencing. For lines 49-50, please refer to *Ali et al.* [2012], and consider rephrasing to "hydro-climate, geomorphologic, and physical catchment characteristics (e.g., land use, soil types and geology)".

c. Line 54: what does the "process of model constitution" mean? developing a model structure, parametrization/calibration of a model?

d. Lines 55-56: term "global performance" could be interpreted in two ways: (1) performance of models on a global scale (around the globe), (2) the overall performance of model when conditioned (e.g. using global calibration methods) given a particular performance metric. Also, the term "performance parameters" is unclear to me; do you mean performance metrics that a model is evaluated against, or model (free) parameters that are calibrated for reaching a particular performance level? Moreover, it seems to me that you are referring to the problem of parameter equifinality in models here; if so, please acknowledge the terminology and its relevant literature.

e. 56-57 what "restrictive assumptions"? If the differences are very small, how they are selected then?

f. Lines 57-58: there is a rich literature on regionalization emphasizing on the importance of physical understanding of the catchment dynamics, particularly dominant runoff

generation; which you ignored here. You can refer to the extant literature discussed in *Hrachowitz et al.* [2013]. Moreover, you argued the selection criteria of the model is merely based on tradeoffs between "statistical parameters". First, what do you mean by "statistical parameters"; error metrics? Second, this is exactly what you based your model swapping on; particularly the criterion of reducing model overall error (as discussed later on table 3). Further, in many of the cases discussed in your results (presented on table 3), the improvement of swapped model (SM_i) over the original model (OM)—measured by root mean squared error, Nash-Sutcliffe Efficiency, and mean absolute error—is trivial, such as 28, 32 40, 44, 57, 68, 78, 82, 108, and 114. In cases that the improvement in error metric is not trivial (such as 90, 95, 101, 103, and 116), it is important to present the modelled hydrographs as well; certain aspects of the hydrograph are visually discernable, while ignored or cancelled out when numerical metrics of model evaluation are used.

g. Lines 64-65: this seems to me like a straw man fallacy. There are many studies which used an ensemble of behavioral models and multi-objective framework to identify feasible parameterizations for ungauged basins (again, you can refer to *Hrachowitz et al.* [2013] for examples). In fact, using a single model framework is now considered as incomplete within mainstream hydrologic research/publications.

h. Lines 101 and 145 simply do not make sense to me!

i. Line 120 and equation 2: you stated "L_{sp} which describes the time difference"; yet given the equation 2, the unit of L_{sp} is not time but flow magnitude.

j. Line 147: what do you exactly mean by "complex descriptors"? please define.

k. For equations 5 and 6 cite the original literature [*Ganora et al.*, 2009].

l. Lines 179: elaborate what you mean by "extraneous (junk) variable" and "standard errors".

m. Lines 226-227: what do you mean by "better satisfies the condition of [a] meaningful transformation"?

n. Lines 237-239, need rephrasing and substantiation, e.g. in "river flow varies unpredictably over a short distance", what do you mean by the underlined term exactly?

3. It is not clear to me what exactly is "comprehensive" about the proposed method (mentioned multiple times e.g. lines 74, 108, 144, …). Using a FDC the "time" aspect of your hydrologic data information content is missed, yet all the observed data points are still preserved and used (i.e. their magnitude). But in the proposed method, the hydrologic data is summarized into monthly means and reduced from years of data points to a monthly so-called "representative flow regime". Arguably, you are shrinking the information content of the data in time, both in their sequencing and magnitude (by averaging).

4. Perhaps one of the major uncertainties (or undiscussed questions) about your work is that although you tested the method on 124 basins, they are all located within a small region. That is, it is not clear how effective your proposed method would be in other regions of the world with a different hydro-climatology. Even if all the results are accurate and model predictability is improved, it is still difficult to claim that it is a superior method. So, the result discussion should be realistic, rather than over-promising about the proposed methods.

5. Equation 6: I'd like to point out to fundamental limitations of R^2 in assessing the predictive power (or goodness-of-fit) of regression models. First, R^2 is independent of model bias [*Legates and McCabe*, 1999]. Secondly, R^2 monotonically increases with the number of variables included in the regression (i.e. the number of descriptors p). In other words, R^2 will never decrease when you add on new descriptors to the model the number of p. It is useful to discuss this problem of over-fitting, and to what extent it is (ir)relevant to the methodology you proposed. Additionally, Akaike-based criteria are also useful for ranking models (from best to worst); you can see *Saft et al.* [2016] as an example.

    a. Line 163, n is not the total number of basins, it is the number of basins considered. Likewise, for equation 7 properly define n.

Minor comments and suggestions

1. Poor use of English language
   a. Line 54: use ; instead of :.
   b. Lines 69-71, rewrite please.
   c. Line 84: "is" tested in Italy.
   d. Stick to American or British English across the manuscript. For instance, "viz." is a British adverb in a manuscript filled with the American verb-forming suffix "-ize".
   e. Poor punctuations, e.g. using underscore '__' instead of em dash '—', e.g. lines 39, 74, 108.
2. Figure 3, location of Italy on the globe is not necessary.
3. Line 146: basin "mean" elevation
4. The term "normalized" is used throughout the literature (particularly line 140). In statistics normalization mainly refers to transforming data distribution into a normal distribution. Given the statistical nature of your work, either use a different term (e.g. transformation, re-expression, etc.), or clearly emphasize that you are not normalizing data in classical sense of the term within statistics literature. Further, d within the in-text equation should be defined.

References

Ali, G., D. Tetzlaff, C. Soulsby, J. J. McDonnell, and R. Capell (2012), A comparison of similarity indices for catchment classification using a cross-regional dataset, *Advances in Water Resources*, *40*, 11-22.

Ganora, D., P. Claps, F. Laio, and A. Viglione (2009), An approach to estimate nonparametric flow duration curves in ungauged basins, *Water Resources Research*, *45*(10).

Hrachowitz, M., et al. (2013), A decade of Predictions in Ungauged Basins (PUB)—a review, *Hydrological Sciences Journal*, *58*(6), 1198-1255.

Legates, D. R., and G. J. McCabe (1999), Evaluating the use of "goodness-of-fit" Measures in hydrologic and hydroclimatic model validation, *Water Resources Research*, *35*(1), 233-241.

Saft, M., M. C. Peel, A. W. Western, and L. Zhang (2016), Predicting shifts in rainfall-runoff partitioning during multiyear drought: Roles of dry period and catchment characteristics, *Water Resources Research*, *52*(12), 9290-9305.