Hydrol. Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/hess-2018-427-RC2, 2018

Hydrology and Earth System Sciences
Discussions

Interactive comment on "Identifying rainfall-runoff events in discharge time series: A data-driven method based on Information Theory" *by* Stephanie Thiesen et al.

Anonymous Referee #2

Received and published: 27 October 2018

Review of hess-2018-427

Identifying rainfall-runoff events in discharge time series: A data-driven method based on Information Theory

Stephanie Thiesen, Paul Darscheid, Uwe Ehret

Summary

———-

This paper discusses data-driven modeling approaches to identifying rainfall-runoff events. The authors construct probabilistic models using combinations of precipita-

tion and discharge to classify whether or not a given discharge should be considered an "event". Multiple models, with varying number of predictors are constructed. Each of these models are ranked by the reduction in uncertainty of the user-classified data. Overall, the paper has novel approaches and contains interesting results. However, in its current form the paper lacks several clarifying details and analyses and I cannot recommend publication in its current form. I suggest major revisions before the manuscript may be acceptable for publication,

Major comments

————————

1. The paper goes into detail with regards to hypothesis selection, model construction, and model evaluation, but lacks necessary details on model analysis. Model evaluation uses reduction in uncertainty of the user-defined classification, making this a supervised learning approach. However, there is very little analysis of how well the resulting models perform on new data. Figure 7 shows the application of the 4-predictor model, but it is unclear whether the application is on data that the model was trained on. A clear application of the resulting model on data which was not used during the model selection process is necessary to build credibility of the technique.

2. Motivations in this paper claim that existing automated event detection techniques are not adequate. In order to understand the benefits of this new technique there should be a comparison with one or more alternatively generated classifications. As currently written, it is unclear what baseline is used to decide that the method is good and whether the method is better than existing methods.

3. Figures 2, 5, and 6 don't provide very much insight. All simply show the convergence of the sub-sampled data to the overall data as sample sizes increase. Further, in figure 6 if the bar groupings are normalized by $\frac{D_{kl}}{H(X|Y)}[N=50]$ all of the curves would fall on top of each other. This is just an illustration of the ratio of the number of data points to the number of bins the estimators use. The caption text is also unclear,

as "curse of dimensionality" is not a formal quantity.

4. Page 7 line 23 claims that the prediction is not biased, but it is known that histogram-based entropy estimators systematically overpredict entropy [1]. This issue is particularly large in more than 3 dimensions [2]. If there is a systematic bias in the entropy estimation, it seems likely that the underlying probability distributions are then biased systematically as well. This is particularly true of the 4-predictor model.

Minor comments

_____

* Neither information theory nor curse of dimensionality need to be capitalized throughout.

* p.1 l.23: In the abstract it is unclear what "relative magnitude of discharge in a 65-hour time window" means. "Relative" to what?

* p.2 l.3: The quote from Chow 1988 appears to be missing a word, "... physiographic and climatic [word missing] that govern..."

* p.2 l.5-7: Aren't i) and iv) basically the same?

* p.2 l.20-29: The discussion of the history of event detection doesn't provide a particularly historic view. Event-detection (and baseflow separation) have a history that goes back a lot further than 2006.

* p.3 l.11-15: Claims about the bias and confidence from data driven methods need citation.

* p.4 l.1: It is unclear what is meant by "1:1" mapping between target and predictor data means on page 4 line 1.

* p.4 l.21: What happens when the system is not stationary?

* Equation 2 has an errant dot before $\log_2$.

* Page 6 line 26: good -> well.

* Section 2.3.1: The phrase "over the same underlying set of events" is unclear.

* The last sentence of page 6 requires a citation as well as clarification of what is meant by "how hard the Curse of Dimensionality hits"

* There are multiple (consistent) definitions of cross entropy.

* p.7 l.24-25: How do you determine whether "appropriate binning choice were made"?

* p.8 l.2: the area of the catchment should be in $kmˆ2$.

* p.8 l.7: The arguments of the KL divergence need to be explained.

* Throughout section 3 discharge units should be $mˆ3 sˆ{-1}$

* Section 3.1: How is snow taken into account in the discharge time series (since this changes the timing between precipitation and streamflow)?

* Section 3.2.5: One sentence sections are rather odd.

* In the application of the method to a new time series, what happens when you encounter conditions that did not previously occur and which are outside the range of your empirical PMF?

* The heading for section 4.2 is vague (non-descriptive)

* p.12 l.19: "Computationally expensive" - what does that mean in this context? 2 minutes on a laptop or a week on a 30 thousand core computing cluster?

* Conclusions: The first part of the conclusions is just a summary of the paper. I think this can be shortened.

References

_____-

[1] Steuer, R., Kurths, J., Daub, C. O., Weiseand, J., & Selbig, J. (2002). The mutual information: Detecting and evaluating depencies between variables. Bioinformatics, 18(2), S231–S240.

[2] Hlaváčková-Schindler, K., Paluš, M., Vejmelka, M., & Bhattacharya, J. (2007). Causality detection based on information-theoretic approaches in time series analysis. Physics Reports, 441(1), 1–46. https://doi.org/10.1016/j.physrep.2006.12.004

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-427, 2018.