Hydrol. Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/hess-2018-427-AC2, 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.

Stephanie Thiesen et al.

stephanie.thiesen@kit.edu

Received and published: 19 November 2018

We thank the second referee for cautiously reviewing our article and providing his/her feedback. Following, we have addressed all of his/her comments and discussed them. The observations were complementary to referee #1 and very helpful to identify some unclear issues.

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

Comment 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.

Response 1: The application presented in Figure 7 is indeed on a subset of the training data, as mentioned in the manuscript (page 14 line 6, and legend of Figure 7). From this and the following major comments, as well as comments made by referee #1 we conclude that there is a need to both better explain and justify the approach we took in the paper for training and evaluating (t+e in the following) the models, and to also include for comparison approaches to t+e the reader is more familiar with, and to compare in this standard setting our model to benchmark models. We start by explaining our approach, and will then propose additions to the manuscript. Our approach to t+e is summarized in Fig. 2, and corresponding results are shown in Figure 5. What the figures show is a summary of a large number of t+e operations, where from the available data randomly chosen subsets of various sizes are used for calibration/supervised learning, and the resulting model is then applied to and evaluated on all data. So the referee is correct in stating that the model is partly evaluated on data it has seen during learning, but at the same time it is also evaluated against data it has not seen yet. So

this is different from a standard split sample approach, where the data sets for training and evaluation are mutually exclusive. However, we would like to argue that the standard split sample approach has the problem of using different data sets of different sizes for validation (depending on the choice and length of the calibration period), which makes comparison of results difficult. Also, in our approach for a given length of training period we conduct many tests and average the results, which yields more robust results than the standard approach of dividing the available data only once into a calibration and validation period, and providing results only from that single test (however, we are aware that for split sampling also more elaborate approaches exist which include more than a single split). So we think that our approach to t+e has some advantages to offer, among them convergence to the best achievable result for increasing sample sizes, which we would like to keep, and we would like to stress that our approach also includes an evaluation of the model on data it has not seen during training, even though this portion varies with the chosen sample size. Nevertheless, we think it will be beneficial to include a standard approach to t+e into the paper: Firstly because it will be easier to explain the specialties of our approach by contrasting it to an established approach, secondly because it will be easier for the reader to judge the model quality by comparing it to own experiences and to the benchmark model in a familiar setting. So we suggest adding a new section to a revised version of the manuscript, in which we will train our model and, as a benchmark, the method proposed by Mei and Anagnostou (2015), in a single training period (the first 9 years of the available data) and then apply, evaluate and compare the models in the remaining last year.

Comment 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.

Response 2: We agree and suggest applying as a benchmark the method proposed

by Mei and Anagnostou (2015) and discuss the relative benefits of each method via comparison. Please also see our reply to comment 1.

Comment 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.

Response 3: Please see our response to comment 1: We would like to maintain that Fig. 2 and 5 provide essential summaries both of our approach to evaluate models (Fig. 2) and actual results for the given data (Fig. 5). Both figures contain a joint visualization of model analysis and model evaluation, and at the same time provide the opportunity of comparing models with different numbers of predictors. That is, they provide an opportunity to decide, for a given amount of data, which number of predictors is optimal in the sense of avoiding both ignoring available information (by choosing too few predictors) and overfitting (by choosing too many predictors). Fig. 6 was meant as an illustration of our concept to formalize (and make objective) the decision what amount of learning data provides a sufficiently good model for the entire data set. The percentage Dkl/H(X|Y) measures what share of total predictive uncertainty is due to a less-than-optimal model (because it was trained only on a subset of the data). If a user decides on such a limit (e.g. 5%), the minimum amount of training data to assure this can be read from the figure for models with different numbers of predictors. However, we agree that in its current form Fig. 6 may be overly detailed. We suggest removing the figure, and instead add to table 7, column 'Sample size where Dkl/H(X|Y)$\leq$5% the relative sample size in addition to the absolute sample size.

Comment 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.

Response 4: The referee raises several interesting and interconnected aspects of calculating information measures from limited samples and/or binned representations of distributions. The effect of binning: In general, the choice of binning affects the absolute values of information measures of a single variate (entropy) as well as information measures of dependency between variates (conditional entropy, mutual information). If this effect is not of interest, it can be avoided by sticking to the same binning throughout an analysis, such that only the relative, not the absolute magnitude of the information measures matter. This is what we did in our study, as we were not interested in the 'true' information measures calculated from 'true' continuous data (or at least data mapped to very high-resolution bins), but on the relative magnitudes of the information measures for various choices and numbers of predictors. The effect of limited samples: Computing information measures from limited samples can indeed introduce biases in the estimation of information measures: entropy is systematically underestimated albeit the underestimation is bounded (see Paninski, 2003), mutual information is systematically overestimated (see Steuer et al., 2002 as mentioned by the referee). In the extreme, if the sample consists of a single observation or single pair of observations, estimated entropy will be zero, and there will be full mutual information (knowing one of the paired values unambiguously identifies the other). So especially in the case of small training data sets, if we would assume that the information measures derived in the training data set would also hold for the case of applying the model to new data, we would systematically over- or underestimate the true values. However, in our study we evaluate the performance of each model always against the same reference, which is the full data set. The measure we use (cross entropy) incorporates both the agreement of true relation among the data and the relation as expressed by the model via Kullback-Leibler divergence and the predictability of the target given the predictors for the full data set (conditional entropy). As a consequence, the negative effects of learning from limited samples is considered in the relative ranking of the various tested mod-

els, which is what we want. Altogether, we suggest that including the above discussion in the manuscript would be beyond its scope. So in order to avoid misinterpretations, we suggest rephrasing the text in section 2.4 in a revised version of the manuscript: "[. . .]. The model returns a probabilistic representation of the target value. If the model was trained on all available data, and is applied within the domain of these data, the predictions will be unbiased and neither over- nor underconfident. If instead a model using deterministic functions is trained and applied in the same manner, the resulting single-valued predictions may also be unbiased, but due to their single-value nature will surely be overconfident.".

Minor Comments:

Comment 5: Neither information theory nor curse of dimensionality need to be capitalized throughout.

Response 5: Thanks. We will adapt the writing style in a revised version of the manuscript.

Comment 6: 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?

Response 6: We mean "relative to all values in the time window". We suggest rephrasing the sentence "[. . .]: discharge from two distinct time steps, the relative magnitude of discharge compared to all discharge values in a surrounding 65-hour time window, and event predictions from the previous time step.

Comment 7: p.2 l.3: The quote from Chow 1988 appears to be missing a word, "... physiographic and climatic [word missing] that govern..."

Response 7: Thank you. The correct quote is indeed "[...] physiographic and climatic characteristics that govern [...]". We will include the correct quote in a revised version of the manuscript.

Comment 8: p.2 l.5-7: Aren't i) and iv) basically the same?

Response 8: We rethought the arguments made here and suggest rephrasing the sentence: "Discharge time series are a fundamental component of hydrological learning and prediction since they i) are relatively easy-to-obtain, in high quality and from widespread and long-existing observation networks. ii) carry robust and integral information about the catchment state, and iii) are an important target quantity for hydrological prediction and decision-making."

Comment 9: 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.

Response 9: Correct. In our discussion we focused on relatively recent techniques suitable for automated event detection. In order to put the paper in the proper historical context, we suggest adding in a revised version of the manuscript a brief overview on older methods of event detection and baseflow separation.

Comment 10: p.3 l.11-15: Claims about the bias and confidence from data driven methods need citation.

Response 10: We suggest rephrasing to clarify what we wanted to say: "Predictions based on probabilistic models that learn relations among data directly from the data, with few or no prior assumptions about the nature of these relations, are less bias-prone (because there are no prior assumptions potentially obstructing convergence towards observed mean behavior), and less likely to be overconfident compared to established models (because applying deterministic models is still standard hydrological practice, and they are overconfident in all but the very few cases of perfect models). This applies at least if there is sufficient data to learn from, appropriate binning choices were made and the application remains within the domain of the data that were used for learning." We suggest that these claims are self-evident and do not require a citation.

Comment 11: 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.

Response 11: It means that each target has exactly one corresponding predictor, i.e., one particular value of target is related to one particular value of predictor (in contrast to 1:n or n:m relationships). We suggest including a more elaborate explanation in a revised version of the manuscript.

Comment 12: p.4 l.21: What happens when the system is not stationary?

Response 12: If the system is non-stationary, i.e. system properties change with time, the inconsistency between the learning and the prediction situation will result in additional predictive uncertainty. The problems associated with predictions of nonstationary systems apply to all modeling approaches. If a stable trend can be identified, a possible countermeasure is to do learning and prediction on detrended data and then re-impose the trend in a post-processing step. We suggest adding this sentence to a revised version of the manuscript.

Comment 13: Equation 2 has an errant dot before $\log_2$.

Response 13: Thank you. We will correct this in a revised version of the manuscript.

Comment 14: Page 6 line 26: good -> well.

Response 14: Thank you. We will correct this in a revised version of the manuscript.

Comment 15: Section 2.3.1: The phrase "over the same underlying set of events" is unclear.

Response 15: We agree that the meaning is not clear. We suggest rephrasing "It is also possible to compare two probability distributions p and q."

Comment 16: 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"

Response 16: We have explained the Curse of Dimensionality in detail and with citations in the same section (p. 6 l.12-23), so we think there is no need to further explain it at the end of page 6. However, to make clearer what we mean we suggest rephrasing

the sentence to "[. . .] i.e., it is a measure of the impact of the Curse of Dimensionality.".

Comment 17: There are multiple (consistent) definitions of cross entropy.

Response 17: Sorry, we do not understand this comment. In the text we provide several definitions and interpretations of cross entropy to provide the reader with a comprehensive perspective of the subject.

Comment 18: p.7 l.24-25: How do you determine whether "appropriate binning choice were made"?

Response 18: A befitting binning choice is indeed a subject of ongoing research, see e.g. Gong et al. (2014) or Pechlivanidis et al. (2016). In general terms, as mentioned in the manuscript (page 4, lines 8-16) an appropriate binning choice yields bins which are neither too narrow (which leads to a overfitted model) nor too wide (which leads to overly smoothed histograms, which will introduce a significant amount of bias and also discards information about the high resolution details of the distribution). We suggest rephrasing the sentence in a revised version of the manuscript: "[. . .], and appropriate binning choices were made (see the related discussion in section 2.2).".

Comment 19: p.8 l.2: the area of the catchment should be in km^2.

Response 19: Sorry, we do not understand this comment. The catchment area is already expressed in km^2.

Comment 20: p.8 l.7: The arguments of the KL divergence need to be explained.

Response 20: Sorry, we do not understand this comment. We could not find arguments regarding KL divergence in the mentioned page.

Comment 21: Throughout section 3 discharge units should be $m^3 s^{-1}$

Response 21: Sorry, we do not understand this comment. Discharge values are already expressed in m^3 s^-1.

Comment 22: Section 3.1: How is snow taken into account in the discharge time series (since this changes the timing between precipitation and streamflow)?

Response 22: Please note that referee #1 had a similar question, so we mainly provide the same reply here as to referee #1, comment 5. Effects of snow accumulation and melting on discharge events were not the central point of the study. Thus, they were not explicitly considered or classified by the model. One of the strengths of the data-based approach we describe is that it potentially accepts any kind of additional predictors such as air temperature, nitrate concentrations, or snow. We agree with the referee that snow-related observations could be a potentially very useful predictor to identify snowmelt events. However, we suggest that doing so would add another facet of application to the manuscript, but would not add to the method description as such, which is the main goal of our paper. We therefore suggest adding a brief discussion of this topic in a revised version of the manuscript, but not an application.

Comment 23: Section 3.2.5: One sentence sections are rather odd.

Response 23: We agree that this looks odd. To avoid it we considered merging subsections of section 3.2. However, due to the nature of the different predictors (many discharge-based predictors and few others), in order to merge precipitation we'd have to merge them all, which would greatly reduce readability. We think that the one-sentence section is the lesser evil and suggest keeping the subsection structure in 3.2 as it is.

Comment 24: 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?

Response 24: Good point. If the conditions are outside of the range of the empirical PMF, the model will fail, i.e. it will provide no prediction. The same problem will also occur if the conditions are within the range of the empirical PMF, but have never been observed, at least not as seen through the filter of the chosen binning. In that case, the

predictive distribution of the target (event Yes/No) will also be empty. If 'no answer' by the model is not acceptable, several methods exist to guarantee an answer, however at the cost of reduced precision:

* Coarse graining: The PMF can be rebuild with fewer, wider bins and an extension of the range (e.g. by merging neighboring bins) until the model provides an answer to the predictive setting. This way the PMF is still purely based on the observed data, but the resulting predictive distributions will be more spread-out as they integrate more observations from a larger range of observed situations. Similar methods to avoid empty bins by adjusting the binning have been proposed for example by Darbellay and Vajda (1999), Knuth (2013) and Pechlivanidis et al. (2016).

* Gap-filling: A widely applied alternative is to maintain the binning and fill the empty bins with non-zero values based on a deemed-to-be-reasonable assumption on the shape of the PMF if more data were available. Approaches comprise adding one counter to each zero-probability bin of the sample histogram, adding a small probability to the sample PDF, smoothing methods such as Kernel-density smoothing (Blower et al. , 2002; Simonoff, 1996), or Bayesian approaches based on the Dirichlet and Multinomial distribution, or a Maximum-Entropy Method recently suggested by Darscheid et al. (2018). All of these methods are applicable both for the extrapolation case mentioned by the referee and the interpolation case. We suggest adding to a revised version of the manuscript a short discussion of this topic as we think it is of interest for the reader.

Comment 25: The heading for section 4.2 is vague (non-descriptive)

Response 25: We chose this section header in relation to that of section 4.1 (model performance for the full data set) to emphasize the difference between the two. In order to maintain this emphasis, we prefer to keep the header as it is.

Comment 26: 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?

Response 26: In our work it means one standard PC running for around 300 hours to analyze a 5- dimensional model (18 different sample sizes, 500 repetitions of each sample size). As it is well known that resampling-based analysis techniques tend to be computationally expensive, we suggest keeping the statement in the text as general as it is.

Comment 27: Conclusions: The first part of the conclusions is just a summary of the paper. I think this can be shortened.

Response 27: We agree that the summary can be shortened and suggest to do so in a revised version of the manuscript. However, we think it is important to give the reader a short wrap-up of what was done before the conclusions are drawn, so we will not abandon the summary altogether. Also, in order to better reflect the content of the section, we suggest changing the section header from 'Conclusions' to 'Summary and conclusions'.

References:

Blower, G.; Kelsall, J.E. Nonlinear Kernel Density Estimation for Binned Data: Convergence in Entropy. Bernoulli 2002, 8, 423–449.

Darbellay, G.A., Vajda, I. Estimation of the information by an adaptive partitioning of the observation space. IEEE Trans. Inf. Theory 1999, 45, 1315–1321.

Darscheid, P., Guthke, A., & Ehret, U. (2018). A Maximum-Entropy Method to Estimate Discrete Distributions from Samples Ensuring Nonzero Probabilities. Entropy, 20, 601. https://doi.org/10.3390/e20080601

[2] Hlavácková-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

Knuth, K.H. Optimal Data-Based Binning for Histograms. arXiv 2013, arXiv:physics/0605197v2.

Mei, Y., Anagnostou, E. N. (2015), A hydrograph separation method based on information from rainfall and runoff records, Journal of Hydrology, 523, 636-649.

Paninski, L. (2003): Estimation of entropy and mutual information. Neural Comput., 15(6): 1191-1253.

Pechlivanidis, I.G.; Jackson, B.; McMillan, H.; Gupta, H.V. Robust informational entropy-based descriptors of flow in catchment hydrology. Hydrol. Sci. J. 2016, 61, 1–18.

Simonoff, J.S. Smoothing Methods in Statistics; Springer: Berlin/Heidelberg, Germany, 1996.

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

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