

Dear Mrs. Schaefli,

The current document consolidates the referee comments, the author responses posted so far and the actions taken in the major revision of the manuscript. It is organized according to the following legend:

- Referee comments
- Authors' responses from the previous stage
- Authors' major revisions and responses for the new version of the manuscript

After this section of comments and responses, you can find the manuscript pdf with the tracked revision. Please note that the page numbers in our responses refer to the revised, track-free version of the manuscript.

Best regards,  
Stephanie Thiesen, Uwe Ehret and Paul Darscheid

## Referee #1 | Dr. Yiwen Mei

**Comment 1:** Results of the manuscript show that precipitation at the current time step is not an informative predictor for flood event. I think this makes sense because there is always a lag time for the excess rainfall to travel to the outlet of the catchment. This has also been pointed out by Mei and Anagnostou (2015) with an event time lag parameter. Therefore, could a cumulative precipitation quantity that is produced based on a window with some size before the current time step help?

**Response 1:** We agree that aggregated precipitation is a potentially very useful predictor for event detection and will likely improve the results. The choice of predictors is an interactive and incremental process, and besides aggregated precipitation there may be other predictors which may eventually improve the results obtained so far. Thus, since the main point of the paper is to introduce the method without necessarily finding the perfect predictive model, and since precipitation data are often not available for analysis, we consider that the presented application is sufficient to demonstrate the potential of the approach. We added to the new version of the manuscript a brief discussion on the potential of using aggregated precipitation as a predictor and snowmelt events (Section 4.1.3, p. 17, l. 4-8).

**Comment 2:** Fig. 4: clearly, the window sizes are different in the center scheme to the forward and backward one and I observed different patterns of the conditional entropy with number of time step involved. I wonder why the authors did not use  $2*n$  in the forward and backward scheme to make the window size consistent?

**Response 2:** We have designed this parameter in a way such as to explore all possible window sizes at the finest resolution, while avoiding problems such as centering with odd window sizes.

The “n” for each predictor was selected individually, and only displayed simultaneously on the same graph. The suggestion by the referee can be met by simply using window size as x-axis position instead of “n” in the graph. We suggest that neither of the two choices entails a particular advantage or disadvantage. Thus, considering that the results are clearly visible and that the results will remain the same, there were no changes in the new version of the manuscript.

**Comment 3:** Clearly, the selection of an optimal probability threshold is essential of the proposed method but the authors did not introduce a method of doing so. The only description is on P14 L10-15 and given the relatively short demonstration period, the authors arbitrarily select 75%. I wonder could such a threshold change with a) increasing number of time step and b) different basins?

**Response 3:** Indeed the threshold could be optimized, and it might be different for different datasets. We chose 75% as a threshold rather ad hoc, and only to demonstrate that it is possible to convert by choosing threshold the probabilistic prediction to a binary one if so desired. If there is no particular reason to do so, it will always be better to keep the probabilistic result as the binary transformation includes loss of information. That said, we agree with the referee that if a binary result is desired, the choice of threshold can and should be found by optimization. The binarization procedure was added as part of the benchmarking comparison between the proposed ITM (information theory method) and the CPM (characteristic point method, proposed by Mei and Anagnostou, 2015) in Section 3.3, and its application can be observed in Section 4.2.

**Comment 4:** It can be seen from Figure 7 that there exist time steps associated with probability lower than 75% within the manually-identified events. This means that event identified by a user-defined fixed probability threshold are different than the manually identified one (i.e. the automatic one will have more events due to the existence of more separation time steps). So, do the authors have a method to skip those very-short discontinuous time steps so as to form longer events in an automatic manner?

**Response 4:** We agree that in general it is desirable to have non-interrupted events of realistic length, but would like to mention that for some use-cases it is not relevant (e.g. if only the number of time steps classified as event is of interest).

However, we see the use-cases where this property is of interest. There are several ways to address this in our method: The first is to include a memory effect in the classification by applying a recursive predictor  $e_p(t-1)$ . Comparable to a Markov model, this helps the model to better ‘stick’ to a classification after a transition from event to no-event or vice versa. While we already present and discuss such a model in the manuscript, the memory effect could be further enhanced if required by adding more recursive predictors (t-2, t-3, etc.). An alternative option would be to increase event coherence in a post-processing step with an autoregressive model, with model parameters found by maximizing agreement with the observed events. We prefer the first method as it simply adds more predictors instead of adding another model component. A brief discussion of the alternatives was incorporated at the end of Section 4.1.3 (p. 17, l. 9-15), but not an application.

**Comment 5:** The authors show that the algorithm fell to detect the snowmelt event. I wonder is the inclusion of additional predictor can potentially help to extract the snowmelt event? For example, by adding in the in-situ or remote sensing-based observation of snow depth/extend, ground temperature and date of year, could the snowmelt event be identified?

**Response 5:** One of the strengths of the data-based approach we describe is that it accepts any kind of additional predictors such as air temperature, nitrate concentrations, etc. We agree with the referee that snow depth (or depth change) could be a potentially very useful predictor to identify snowmelt events. As mentioned in comment #1, we suggest that doing so would add another facet of application to the manuscript, but would not add to the method description as such. Same as response #1. We added to the new version of the manuscript a brief discussion on the potential of using aggregated precipitation as a predictor and snowmelt events (Section 4.1.3, p. 17, l. 4-8).

**Comment 6:** I think a comparison with the existing event identification methods could help to reveal the values of this newly-developed data-based method. If the authors would like to compare with Mei and Anagnostou (2015) method they may find the matlab codes on our GitHub profile ([https://github.com/YiwenMei/Hydro\\_Seper](https://github.com/YiwenMei/Hydro_Seper)). Another way to verify the method is to demonstrate the patterns of flood event parameters (e.g., runoff coefficient, time lag, baseflow) of events identified by the method.

**Response 6:** We agree that a comparison with existing method adds a valuable additional perspective to the study. Thus, two new sections were added to describe (Section 3.3) and apply (4.2) the binary transformation of the probabilistic prediction, holdout cross-validation and comparative analysis between ITM and CPM.

**Comment 7:** I did not see descriptions on potential usage of this method. Can this method be used to construct flood event database like for example the Shen et al. (2017, Comprehensive Database of Flood Events in the Contiguous United States from 2002 to 2013) constructed by the Characteristic Points Method introduced in Mei and Anagnostou (2015)?

**Response 7:** Given sufficient data to learning, we believe it is possible to construct such a database with our method. For each gauge in the database an individually optimized set of predictors could be identified. In addition, since we are dealing with a data-driven approach and avoiding parametrizations such as equations or indexes, the more event categories we aspire to classify, the more data will be required. The discussion of the model potential usage was updated in Sections 1 (p. 3, l. 27-30) and Section 5 (p. 20, l. 11-22).

## Referee #2 | Anonymous

**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.

Thus, as mentioned in the response #1 of referee #1, two new sections were added to describe (Section 3.3) and apply (4.2) the binary transformation of the probabilistic prediction, holdout cross-validation and comparative analysis between ITM and CPM. Furthermore, an more detailed explanation regarding Fig. 2 was added in Section 2.3.2 (p. 7, l. 18-27).

**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. As mentioned in the previous answer, two new sections were added to describe (Section 3.3) and apply (4.2) the binary transformation of the probabilistic prediction, holdout cross-validation and comparative analysis between ITM and CPM.

**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  $\frac{D_{\{k\}}}{H(X|Y)}[N=50]$  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. 2 and Fig. 5 were maintained and a more detailed explanation regarding Fig. 2 was added in Section 2.3.2 (p. 7, l. 18-27) and regarding Fig. 5 in Section 2.3.2 (p. 15, l. 16-18). 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  $D_{k|}/H(X|Y)$  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. In the new version of the manuscript, we removed the Fig. 6, and instead added to Table 7 (Column 5) the relative sample size given the full data set, 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 models, 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 reformulated the text in Section 2.4 (p. 8, l. 17-21) 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 adapted the writing style throughout the text 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. We included our reformulated sentence in a revised version of the manuscript (Abstract, p. 1, l. 22 and Section 5, p. 19, l. 25).

**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 included the correct quote in a revised version of the manuscript (Section 1, p. 2, l. 2).

**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, available 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.". We included the reformulated sentence in a revised version of the manuscript (Section 1, p. 2, l. 5-7).

**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. We added a brief overview of older methods in Section 1, p. 2, l. 21-27).

**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. We included the reformulated sentence with some minor adjusts in a revised version of the manuscript (Section 1, p. 3, l. 15-21).

**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 included a more elaborate explanation in a revised version of the manuscript (Section 2.1, p. 4, l. 10-11).

**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 non-stationary 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 reimpose the trend in a post-processing step. We added a non-stationarity discussion in Section 2.2 (p. 5, l. 3-6) in the revised version of the manuscript.

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

**Response 13:** Thank you. We corrected the errant dot in Eq. (2), Section 2.3.1.

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

**Response 14:** Thank you. We corrected this miswriting in a revised version of the manuscript (Section 2.3.2, p. 7, l. 7).

**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  $q$ ". We included the reformulated sentence in a revised version of the manuscript (Section 2.3.1, p. 6, l. 1).

**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.". We included the reformulated sentence in a revised version of the manuscript (Section 2.3.2, p. 7, l. 14-15), but not extra information.

**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. No action was taken regarding this comment.

**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). Same as response #4 and #10. We reformulated the text in Section 2.4 (p. 8, l. 17-21) and create a reference for “appropriate binning choice” in (Section 1, p. 3, l. 20-21).

**Comment 19:** p.8 l.2: the area of the catchment should be in  $\text{km}^2$ .

**Response 19:** Sorry, we do not understand this comment. The catchment area is already expressed in  $\text{km}^2$ . No action was taken regarding this comment.

**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. No action was taken regarding this comment.

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

**Response 21:** Sorry, we do not understand this comment. Discharge values are already expressed in  $\text{m}^3 \text{s}^{-1}$ . No action was taken regarding this comment.

**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. As mentioned in the responses #1 and #5 of referee #1. We added to the new version of the manuscript a brief discussion on the potential of using aggregated precipitation as a predictor and snowmelt events (Section 4.1.3, p. 17, l. 4-8).

**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. No action was taken regarding this comment. Note that since the structure of the revised manuscript was rethought to settle the comparative investigation between ITM and CPM, the mentioned Section 3.2.5 is currently the Section 3.2.1.5 in the revised version.

**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 rebuilt 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 included to the revised version of the manuscript a short discussion of this topic (Section 2.4, p. 8, l. 22-32).

**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. **Thus, no action was taken regarding this comment.**

**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. **Thus, no action was taken regarding this comment.**

**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. 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'. We revised the manuscript conclusion (Section 5), renaming the header to 'Summary and conclusions', removing few excerpts of the summary, and including the cross-validation and comparative investigation procedure and results.

## 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] 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>
- 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.

# Identifying rainfall-runoff events in discharge time series: A data-driven method based on **i**nformation **t**heory

Stephanie Thiesen<sup>1</sup>, Paul Darscheid, Uwe Ehret<sup>1</sup>

<sup>1</sup>Institute of Water Resources and River Basin Management, Karlsruhe Institute of Technology - KIT, Karlsruhe, Germany

5 *Correspondence to:* Stephanie Thiesen (stephanie.thiesen@kit.edu)

**Abstract.** In this study, we propose a data-driven approach to automatically identify rainfall-runoff events in discharge time series. The core of the concept is to construct and apply discrete multivariate probability distributions to obtain probabilistic predictions of each time step being part of an event. The approach permits any data to serve as predictors, and it is non-parametric in the sense that it can handle any kind of relation between the predictor(s) and the target. Each choice of a particular predictor data set is equivalent to formulating a model hypothesis. Among competing models, the best is found by comparing their predictive power in a training data set with user-classified events. For evaluation, we use measures from **Information Theory** [information theory](#) such as Shannon Entropy and Conditional Entropy to select the best predictors and models and, additionally, measure the risk of overfitting via Cross Entropy and Kullback-Leibler Divergence. As all these measures are expressed in “bit”, we can combine them to identify models with the best tradeoff between predictive power and robustness given the available data.

We applied the method to data from the Dornbirnerach catchment in Austria distinguishing three different model types: Models relying on discharge data, models using both discharge and precipitation data, and recursive models, i.e., models using their own predictions of a previous time step as an additional predictor. In the case study, the additional use of precipitation reduced predictive uncertainty only by a small amount, likely because the information provided by precipitation is already contained in the discharge data. More generally, we found that the robustness of a model quickly dropped with the increase in the number of predictors used (an effect well known as the **Curse of Dimensionality** [curse of dimensionality](#)), such that in the end, the best model was a recursive one applying four predictors (three standard and one recursive): discharge from two distinct time steps, the relative magnitude of discharge [compared to all discharge values in a surrounding in-a](#) 65-hour time window and event predictions from the previous time step. Applying the model reduced the uncertainty about event classification by 77.8 %, decreasing Conditional Entropy from 0.516 to 0.114 bits. [To assess the quality of the proposed method, its results were binarized and validated through a holdout method, and then compared to a physically-based approach. The comparison showed a similar behavior of both models \(both with accuracy near 90 %\) and the cross-validation reinforced the quality of the proposed model.](#)

Given enough data to build data-driven models, their potential lies in the way they learn and exploit relations between data unconstrained by functional or parametric assumptions and choices. And, beyond that, the use of these models to reproduce a hydrologist’s way to identify rainfall-runoff events is just one of many potential applications.

## 1 Introduction

Discharge time series are essential for various activities in Hydrology and Water resources management. In the words of Chow (1988) “[...] the hydrograph is an integral expression of the physiographic and climatic characteristics that govern the relations between rainfall and runoff of a particular drainage basin”. Discharge time series are a fundamental component of hydrological learning and prediction since they i) are relatively easy-to-obtain, available 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, and iv) are available from widespread and long-existing observation networks.

Beyond their value to provide long-term averages aiding water balance considerations, the information they contain about limited periods of elevated discharge can be exploited for baseflow separation, water power planning, sizing of reservoirs and retention ponds, design of hydraulic structures such as bridges, dams or urban storm drainage systems, risk assessment of floods and soil erosion. These periods, essentially characterized by a rising (start), peak and recession (ending) points (Mei and Anagnostou, 2015), will hereafter simply be referred to as “events”. They can have many causes (rainfall, snowmelt, upstream reservoir operation, etc.) and equally many characteristic durations, magnitudes, and shapes. Interestingly, while for a trained hydrologist with a particular purpose in mind, it is usually straightforward to identify such events in a time series, it is hard to identify them automatically based on a set of rigid criteria. One reason for this is that the set of criteria for discerning events from non-events typically comprises both global and local aspects, i.e., some aspects relate to properties of the entire time series, some to properties in time windows. And to make things worse, the relative importance of these criteria can vary over time, and they strongly depend on user-requirements, hydroclimate and catchment properties.

So why not stick to manual event detection? Its obvious drawbacks are that it is cumbersome, subject to handling errors and hard to reproduce especially when working with long-term data. As a consequence, many methods for objective and automatized event detection have been suggested. The baseflow separation, and consequently the event identification (since the separation allows the identification of the start and end time of the events), has a long history of development. Theoretical and empirical methods for determining baseflow are discussed since 1893 as presented in Hoyt et al. (1936). One of the oldest techniques according to Chow et al. (1988) date back from the early 1930s, with the normal depletion curve from Horton (1933). As stated by Hall (1968), fairly complete discussions of baseflow equations, mathematical derivations, and applications were already present in the 1960s. In the last two decades, more recent techniques embracing a multitude of approaches (graphical-, theoretical-, mathematical-, empirical-, physical- and data- based) aim to automate the separation.

Ehret and Zehe (2011) and Seibert et al. (2016) applied a simple discharge threshold approach with partly unsatisfactory results; Merz et al. (2006) introduced an iterative approach for event identification based on the comparison of direct runoff and a threshold. Merz and Blöschl (2009) expanded the concept to analyze runoff coefficients and applied it to a large set of catchments. Blume et al. (2007) developed the “Constant-k” method for baseflow separation employing a gradient-based search for the end of event discharge. Koskelo et al. (2012) presented the physically-based “Sliding Average with Rain Record”

- SARR - method for baseflow separation in small watersheds based on precipitation and quickflow response. Mei and Anagnostou (2015) suggested a physically-based approach for combined event detection and baseflow separation, which provides event start, peak, and end times.

While all of these methods have the advantage of being objective and automatable, they suffer from limited generality. The reason is that each of them contains some kind of conceptualized, fixed relation between input and output. Even though this relation can be customized to a particular application by adapting parameters, it remains to a certain degree invariant. In particular, each method requires an invariant set of input data and sometimes it is constrained to a specific scale, which limits its application to specific cases and to where these data are available.

With the rapidly increasing availability of observation data, computer storage, and processing power, data-based models have become increasingly popular as an addition or alternative to established modeling approaches in Hydrology and Hydraulics (Solomatine and Osfeld, 2009). According to Solomatine and Osfeld (2008, 2009), they have the advantage of not requiring detailed consideration of physical processes (or any kind of a priori known relation between model input and output); instead, they infer these relations from data, which however requires that there are enough data to learn from. Of course, including a priori known relations among data into models is an advantage as long as we can assure that they really apply. However, when

facing undetermined problems, i.e., for cases where system configuration, initial and boundary conditions are not well known, applying these relations may be over-constraining, which may lead to biased and/or overconfident predictions. ~~Predictions based on data-driven models are less bias-prone, and neither over- nor underconfident, at least if appropriate binning choices were made and the application remains within the domain of the data used for learning. We suggest that event detection is a good example of such an underdetermined problem, and hence that a data-based approach might be conducive to capture and reproduce a hydrologist's way of event identification. 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 (see the related discussion in Section 2.2), and the application remains within the domain of the data that were used for learning.~~

In the context of data-based modeling in Hydrology, concepts and measures from ~~Information Theory~~ information theory are becoming increasingly popular to describe and infer relations among data (Liu et al., 2016), quantify uncertainty and evaluate model performance (Chapman, 1986; Liu et al., 2016), estimate information flows (Weijs, 2011; Darscheid, 2017), analyze spatio-temporal variability of precipitation data (Mishra et al., 2009; Brunsell, 2010), describe catchment flow (Pechlivanidis et al., 2016), and measure quantity and quality of information in hydrological models (Nearing and Gupta, 2015).

In this study, we describe and test a data-driven approach for event detection formulated in terms of ~~Information Theory~~ information theory, showing that ~~-its potential goes beyond event classification, since it enables the identification of the drivers of the classification, to choose the most suitable model for an available data set, to quantify minimal data~~



requirements, to automatically reproduce classifications for database generation, to handle any kind of relation between the data. The method is presented in Section 2. In Section 3, we describe [atwo](#) test applications with data from the Dornbirnerach catchment in Austria. We present the results in Section 4 and draw conclusions in Section 5.

## 2 Method description

5 The core of the [information theory](#) method ([ITM](#)) is straightforward and generally applicable; its main steps are shown in Fig. 1 and will be explained in the following.

### 2.1 Model hypothesis step

The process starts by selecting the target (what we want to predict) and the predictor data (that potentially contain information about the target). Choosing the predictors constitutes the first and most important model hypothesis, and there are almost no  
10 restrictions to this choice: They can be any kind of observational or other data, transformed by the user or not; they can be part of the target data set themselves, e.g., time-lagged or space-shifted; they can even be the output of another model. The second choice and model hypothesis is the mapping between items in the target and the predictor data set, i.e., the relation hypothesis. It is important for the later construction of conditional histograms that a 1:1 mapping exists between target and predictor data,  
15 [i.e., one particular value of target is related to one particular value of predictor \(in contrast to 1:n or n:m relationships\)](#). Often, the mapping relation is established by equality in time.

### 2.2 Model building step

The next step is the first part of model building. It consists of choosing the value range and binning strategy for target and predictor data. These choices are important as they will frame the estimated multivariate probability mass functions (PMFs) constituting the model and directly influence the statistics we compute from them for evaluation. Generally, these choices are  
20 subjective and reflect user-specific requirements and should be made taking into consideration data precision and distribution, size of the available data sets and required resolution of the output. According to Gong et al. (2014), when constructing probability density functions (PDFs) from data via the simple bin-counting method, “[...] too small a bin width may lead to a histogram that is too rough an approximation of the underlying distribution, while an overly large bin width may result in a histogram that is overly smooth compared to the true PDF.”. Gong et al. (2014) also discussed the selection of an optimal bin  
25 width by balancing bias and variance of PDF estimation. Pechlivanidis et al. (2016) investigated the effect of bin resolution on the calculation of Shannon Entropy and recommended that bin width should not be less than the precision of the data. Also, while equidistant bins have the advantage of being simple and computationally efficient (Ruddell and Kumar 2009), hybrid alternatives can overcome weaknesses of conventional binning methods to achieve a better representation of the full range of data (Pechlivanidis et al., 2016).

With the binning strategy fixed, the last part of the model building is to construct a multivariate PMF from all predictors and related target data. The PMF dimension equals the number of predictors plus one (the target), and the way probability mass is distributed within is a direct representation of the nature and strength of the relationship between predictors and target as contained in the data. Application of this kind of model for a given set of predictor values is straightforward: We simply extract the related conditional PMF (or PDF) of the target, which, under the assumption of system stationarity, is a probabilistic prediction of the target value.

If the system is non-stationary, i.e., when 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 non-stationary 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 reimpose the trend in a post-processing step.

## 2.3 Model evaluation step

### 2.3.1 Information Theory – Measures

In order to evaluate the usefulness of a model, we apply concepts from Information Theory information theory to select the best predictors (the drivers of the classification) and validate the model. With this in view, this section provides a brief description of the Information Theory information theory concepts and measures applied in this study. The section is based on Cover and Thomas (2006), which we recommend for a more detailed introduction to the concepts of Information Theory information theory. Complementary, for specific applications to investigate hydrological data series, we refer the reader to Darscheid (2017).

Entropy can be seen as a measure of the uncertainty about a random variable; it is a measure of the amount of information required on average to describe a random variable (Cover and Thomas, 2006). Let  $X$  be a discrete random variable with alphabet  $\chi$  and probability mass function  $p(x)$ ,  $x \in \chi$ . Then, the Shannon Entropy  $H(X)$  of a discrete random variable  $X$  is defined by

$$H(X) = - \sum_{x \in \chi} p(x) \log_2 p(x) \quad (1)$$

If the logarithm is taken to base two, an intuitive interpretation of Entropy is: “Given prior knowledge of a distribution, how many binary (Yes/No) questions need to be asked on average until a value randomly drawn from this distribution is identified?”.

We can describe the Conditional Entropy as the Shannon Entropy of a random variable conditional on the (prior) knowledge of another random variable. The Conditional Entropy  $H(X|Y)$  of a pair of discrete random variables  $(X, Y)$  is defined as

$$H(X|Y) = - \sum_{y \in \mathcal{Y}} p(y) \sum_{x \in \chi} p(x|y) \cdot \log_2 p(x|y) \quad (2)$$

The reduction in uncertainty due to another random variable is called the Mutual Information  $I(X, Y)$ , which is equal to  $H(X) - H(X|Y)$ . In the study, both measures, Shannon Entropy and Conditional Entropy, are used to quantify the uncertainty of the models (uni- and multivariate probability distributions, respectively). The first is calculated as reference and measures the uncertainty of the target data set. The latter is applied to the probability distributions of the target conditional on predictor(s), and it corroborates to select the best-more informative predictors, i.e., the ones which lead to the most significant reduction of uncertainty about the target.

It is also possible to compare two probability distributions  $p$  and  $q$  over the same underlying set of events. For measuring the statistical “distance” between the distributions  $p$  and  $q$ , it is common to use Relative Entropy or Kullback-Leibler Divergence  $D_{KL}(p||q)$ , which is defined as

$$D_{KL}(p||q) = \sum_{x \in \mathcal{X}} p(x) \log_2 \frac{p(x)}{q(x)} \quad (3)$$

The Kullback-Leibler Divergence is also a measure of the inefficiency of assuming that the distribution is  $q$  when the true distribution is  $p$  (Cover and Thomas, 2006). The Shannon Entropy  $H(p)$  of the true distribution  $p$  plus the Kullback-Leibler Divergence  $D_{KL}(p||q)$  of  $p$  with respect to  $q$  is called Cross Entropy  $H_{pq}(X||Y)$ . In the study, we use these related measures to validate the models and to avoid overfitting, by measuring the additional uncertainty of a model if it is not based on the full data set  $p$  but only a sample  $q$  thereof.

Note that the uncertainty measured by Eq. (1) to Eq. (3) depends only on event probabilities, not on their values. This is convenient as it allows joint treatment of many different sources and types of data in a single framework.

### 2.3.2 Information Theory – Model evaluation

As a benchmark, we can start with the case where no predictor is available, but only the unconditional probability distribution of the target is known. As seen, the associated predictive uncertainty can be measured by the Shannon Entropy  $H(X)$  of the distribution (here  $X$  indicates the target). If we introduce a predictor and know its value in a particular situation a priori, predictive uncertainty is the entropy of the conditional probability function of the target given the particular predictor value. Conditional Entropy  $H(X|Y)$ , where  $Y$  indicates the predictor(s), is then simply the probability-weighted sum of entropies of all conditional PMFs. Mutual information  $I(X, Y)$  is the difference between Shannon Entropy and Conditional Entropy and, like Conditional Entropy, is a generic measure of statistical dependence between variables (Sharma and Mehrotra, 2014), which we can use to compare competing model hypotheses and select the best among them.

Obviously, advantages of setting up data-driven models in the described way are that it involves very few assumptions and it is straightforward to formulate a large number of alternative model hypotheses. However, there is an important aspect we need to consider: from the information inequality, we know that Conditional Entropy is always less than or equal to the Shannon Entropy of the target (Cover and Thomas, 2006). In other words, “information never hurts”, and consequently adding more predictors will always either improve or at the least not worsen results. In the extreme, given enough predictors and applying

a very refined binning scheme, a model can potentially yield perfect predictions if applied to the learning data set. However, besides the higher computational effort, in this situation, the “~~Curse of Dimensionality~~curse of dimensionality” (Bellman, 1957) occurs, which “covers various effects and difficulties arising from the increasing number of dimensions in a mathematical space for which only a limited number of data points are available (Darscheid, 2017)”. This means that with each predictor added to the model, the dimension of the conditional target-predictor PMF will increase by one, but its volume will increase exponentially. For example, if the target PMF is covered by two bins and each predictor by 100, then a single, double and triple predictor model will consist of 200; 20 000 and 2 000 000 bins, respectively. Clearly, we will need a much larger data set to populate the latter PMF than the first. This also means that increasing the number of predictors for a fixed number of available data increases the risk of creating an overfitted/non-robust model in the sense that it will become more and more sensitive to the absence or presence of each particular data point. Models overfitted to a particular data set are less likely to produce good results when applied to other data sets than robust models, which capture the essentials of the data relation without getting lost in detail.

We consider this effect with a resampling approach: From the available data set, we take samples of various sizes and construct the model from the sample (see repetition statement regarding  $N$  in Fig. 1). Obviously, as the model was built from just a sample, it will not reflect the target-predictor relation as ~~good-well~~ as a model constructed from the entire data set. It has been shown (Cover and Thomas, 2006; Darscheid, 2017) that the total uncertainty of such an imperfect model is the sum of two components: The Conditional Entropy  $H(X|Y)$  of the “perfect” model constructed from all data and the Kullback-Leibler Divergence  $D_{KL}$  between the sample-based and the perfect model.  $D_{KL}$  measures the statistical “distance” between the two distributions, in other words, it quantifies the additional uncertainty due to the use of an imperfect model. For a given model (selection of target and predictors), the first summand is independent of the sample size as it is calculated from the full data set, but the second varies: the smaller the sample, the higher  $D_{KL}$ . Another important aspect of  $D_{KL}$  is that for a fixed amount of data, it strongly increases with the dimension of the related PMFs, in other words:e., it is a measure of the impact of the curse of dimensionalityhow hard the Curse of Dimensionality hits.

In information terms, the sum of Conditional Entropy and Kullback-Leibler Divergence is referred to as Cross Entropy  $H_{pq}(X||Y)$ . A typical example of Cross Entropy as a function of sample size is, for a single model, shown in Fig. 2.

The curve represents the mean of several repetitions, which were randomly taken with replacement among these repetitions. Note that, comparable to the Monte Carlo cross-validation, the analysis presented in Fig.2 summarizes a large number of training and testing splits performed repeatedly, and, in addition, also in different split proportions (subsets of various sizes). The difference here is that, in contrast to a standard split where data sets for training and testing are mutually exclusive, we build the model in the training set and apply it in the full data set, where part of the data has not been seen yet, and part has. In other words, we use the training subsets for building the model (a supervised learning approach), and the resulting model is then applied to and evaluated on the full data set. If, on the one hand, the use of the full data set for the application includes data of the training set, on the other hand, the procedure favors the comparison of the results always with the same model.

Thus, the stated procedure allows a robust and holistic analysis, in the sense it works with the mean of  $W$  repetition for each subset and compares different sizes of training subset with a unique reference, the model built from the full data set.

Particularly, Fig. 2 shows that  $F$  for small sample sizes,  $D_{KL}$  is the main contributor to total uncertainty, but when the sample approaches the size of the full data set, it disappears, and total uncertainty equals Conditional Entropy.

5 From the shape of the curve in Fig. 2 we can also infer whether the available data are sufficient to support the model: When  $D_{KL}$  approaches zero (Cross Entropy approaches its minimum), this indicates that the model can be robustly estimated from the data, or, in other words, the sample size is enough to represent the full data set. In an objective manner, we can also do this kinda complementary of analysis in an objective manner by calculating the ratio  $D_{KL}/H(X|Y)$ , which is a measure of the relative contribution of  $D_{KL}$  to total uncertainty. We can then compare this ratio to a defined tolerance limit (e.g., 5 %) to find  
10 the minimally required sample size.

Another application for Fig. 2 is to~~We can also~~ use these kinds of plots to select the best among competing models with different numbers of predictors. Typically, for small sample sizes, simple models will outperform multi-predictor models as the latter will be hit harder by the Curse of Dimensionality~~curse of dimensionality~~; but with increasing data availability, this effect will vanish, and models incorporating more sources of information will be rewarded.

15 In order to reduce the effect of chance when taking random samples, we repeat the described resampling and evaluation procedure many times for each sample size (see repetition statement  $W$  in Fig. 1) and take the average of the resulting  $D_{KL}$ 's and  $H_{pq}$ 's ~~Cross Entropies~~. Based on these averaged results, we can identify the best model for a set of available data.

The proposed Cross Entropy curve contains 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, being a support tool to decide, for a given  
20 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). And, since it incorporates a sort of cross-validation in its construction, o~~ne~~ of the advantages of this approach is that it avoids splitting the available data ~~set~~ into a ~~calibration-training~~ and a ~~validation-testing~~ set. Instead, it makes use of all available data to learn and provides measures of model performance across a range of sample sizes.

## 25 2.4 Model application step

Once a model has been selected, ~~its~~the ITM application is straightforward: From the multivariate PMF that represents the model, we simply extract the conditional PMF of the target for a given set of predictor values. The model returns a probabilistic ~~prediction 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  
30 their single-value nature will surely be overconfident, which is neither biased, over or underconfident as long as the

~~assumption of stationarity between the learning and the application situation is met, and appropriate binning choices were made.~~

~~For application in a new time series, if its conditions are outside of the range of the empirical PMF or if they are within the range but have never been observed in the training data set, the predictive distribution of the target (event Yes/No) will be empty and the model will not provide a prediction. Several methods exist to guarantee a model answer, however at the cost of reduced precision. The solutions range from i) coarse-graining: where the PMF can be rebuilt with fewer, wider bins and an extension of the range until the model provides an answer to the predictive setting, as have been proposed by Darbellay and Vajda (1999), Knuth (2013) and Pechlivanidis et al. (2016); to ii) gap-filling: where the binning is maintained and the empty bins are filled with non-zero values based on a reasonable assumption. Gap-filling 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 and Kelsall, 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). The latter being applied in the present study.~~

### 15 **3 Design of a test application**

~~For demonstration purposes, we applied the method to a set of test data.~~ In this section, we describe the hydro-climatic properties of the ~~test data,~~ and the two performed applications. ~~For demonstration purposes, the first test application was developed according to the Section 2, in order to explain which additional predictors we derived from the raw data and related binning and other choices, and present our strategy for model setup, classification, and evaluation. For benchmarking purposes, the second application compares the proposed data-driven approach (ITM) with the physically-based approach proposed by Mei and Anagnostou (2015), the characteristic point method (CPM), and applies the holdout method (splitting the data set into training and testing set) for the cross-validation analysis.~~

#### **3.1 Data and site properties**

25 We used quality-controlled hourly discharge and precipitation observations from a 9-year period (Oct/31/1996–Nov/01/2005, 78 912 time steps). Discharge data are from gauge “Hoher Steg”, which is located at the outlet of the 113 km<sup>2</sup> Alpine catchment of the Dornbirnerach ~~R~~river in north-western Austria. Precipitation data are from station “Ebnet” located within the catchment. For the test period, we manually identified hydrological events by visual inspection of the discharge time series. To guide this process, we used a broad event definition, which can be summarized as follows: “An event is a coherent period of elevated discharge compared to the discharge immediately before and after, and/or a coherent period of high discharge compared to the data of the entire time series”. We suggest that this is a typical definition if the goal is to identify events for hydrological



process studies such as analysis of rainfall-runoff coefficients, baseflow separation, or recession analysis. Based on this definition, we classified each time step of the time series as either being part of an event (value “1”) or not (value “0”).

Altogether, we identified 177 individual events covering 9092 time steps, which is 11.5 % of the time series. [For the available 9-year period, the maximum precipitation is 28.5 mm h<sup>-1</sup>, and the maximum and minimum discharge values are 237.0 m<sup>3</sup> s<sup>-1</sup>](#)

5 [and 0.037 m<sup>3</sup> s<sup>-1</sup>, respectively.](#) A preliminary analysis revealed that all times with discharge exceeding 15.2 m<sup>3</sup> s<sup>-1</sup> were classified as event, all times with discharge below 0.287 m<sup>3</sup> s<sup>-1</sup> were always classified as non-event.

Both, the input data and the event classification are shown in Fig. 3.

### **3.2 Application I – ITM**

#### 10 **3.2.3.2.1 Predictor data and binning**

Since we wanted to build and test a large range of models, we did not only apply the raw observations of discharge and precipitation but also derived new data sets. The target and all predictor data sets with the related binning choices are listed in Table 1; additionally, the predictors are explained in the text below. For reasons of comparability, we applied uniform binning (fixed width interval partitions) to all data used in the study, except for discharge: Here we grouped all values exceeding 15.2 m<sup>3</sup> s<sup>-1</sup> (the threshold beyond which an event occurred for sure) into one group to increase computational efficiency. For each data type, we selected the bin range to cover the range of observed data and chose the number of bins with the objective to maintain the overall shape of the distributions with the least number of bins.

##### **3.2.3.2.1.1 Discharge $Q$ [m<sup>3</sup> s<sup>-1</sup>]**

20 This is the discharge as measured at Hoher Steg. In order to predict an event classification at time step  $t$ , we tested discharge at the same time step as a predictor –  $Q(t)$  –, and at time steps before –  $Q(t - 1)$ ,  $Q(t - 2)$  – and after –  $Q(t + 1)$ ,  $Q(t + 2)$ .

##### **3.2.3.2.1.2 Natural logarithm of discharge $\ln Q$ [ln(m<sup>3</sup> s<sup>-1</sup>)]**

25 We also used a log-transform of discharge to evaluate whether this non-linear transformation preserved more information in  $Q$  when mapped into the binning scheme than the raw values. Note that the same effect could also be achieved by a logarithmic binning strategy, but as mentioned we decided to maintain the same binning scheme for reasons of comparability. As for  $Q$ , we applied the log-transform also to time-shifted data.

Formatado: Título 3

Formatado: Título 4 HESS ST

Formatado: Título 4 HESS ST

### 3.2.33.2.1.3 Relative magnitude of discharge $Q_{RM}$ [-]

This is a local identifier of discharge magnitude at time  $t$  in relation to its neighbors within a time window. For each time-step, we normalized discharge into the range  $[0, 1]$  using Eq. (4), where  $Q_{max}$  is the largest value of  $Q$  within the window and  $Q_{min}$  is the smallest.

$$Q_{RM} = \frac{Q(t) - Q_{min}}{Q_{max} - Q_{min}} \quad (4)$$

- 5 A value of  $Q_{RM} = 0$  indicates that  $Q(t)$  is the smallest discharge within the analyzed window, a value of  $Q_{RM} = 1$  indicates that it is the largest. We calculated these values for many window sizes and for windows with the time step under consideration in the center ( $Q_{RMC}$ ), at the right end ( $Q_{RMR}$ ) and at the left end ( $Q_{RML}$ ) of the window. The best results were obtained for a time-centered window of 65 hours. For further details see Section [3.2.23.3](#).

### 3.2.43.2.1.4 Slope of discharge $Q_{slope}$ [ $m^3 s^{-1} h^{-1}$ ]

- 10 This is the local inclination of the hydrograph. This predictor was created to take into consideration the rate and direction of discharge changes. We calculated both the slope from the previous to the current time step applying Eq. (5) and the slope from the current to the next time step applying Eq. (6). Positive values always indicate rising discharge.

$$Q_{slope_{before}} = \frac{Q(t) - Q(t-1)}{t - (t-1)} \quad (5)$$

$$Q_{slope_{after}} = \frac{Q(t+1) - Q(t)}{(t+1) - t} \quad (6)$$

### 3.2.53.2.1.5 Precipitation $P$ [ $mm h^{-1}$ ]

This is the precipitation as measured at Ebnet.

- 15 **3.2.63.2.1.6 Model-based event probability  $e_p$  [-]**

- In general, information about a target of interest can be encoded in related data such as the predictors introduced above, but it can also be encoded in the ordering of data. This is the case if the processes that are shaping the target exhibit some kind of temporal memory or spatial coherence. For example, the chance of a particular time step to be classified as being part of an event increases if the discharge is on the rise, and it declines if the discharge declines. We can incorporate this information by adding to the predictors discharge from increasingly distant time steps, but this comes at the price of a rapidly increasing impact of the ~~Curse of Dimensionality~~[curse of dimensionality](#). To mitigate this effect, we can use sequential or recursive modeling approaches: In a first step, we build a model using a set of predictors and apply it to predict the target. In a next step, we use this prediction as a new, model-derived predictor, combine it with other predictors in a second model, use it to make a second prediction of the target and so forth. Each time we map information from the multi-dimensional set of predictors onto
- 20 the 1-dimensional model output, we compress data and reduce dimensionality while hoping to preserve most of the information
- 25

Formatado: Título 4 HESS ST

Formatado: Título 4 HESS ST

Formatado: Título 4 HESS ST

Formatado: Título 4 HESS ST

contained in the predictors. Of course, if we apply such a recursive scheme and want to avoid iterations, we need to avoid circular references, i.e., the output of the first model must not depend on the output of the second. In our application, we assured this by using the output from the first model at time step  $t - 1$  as a predictor in the second model to make a prediction at time step  $t$ . Comparable to a Markov model, this kind of predictor helps the model to better stick to a classification after a transition

5 from event to no-event or vice versa.

### 3.3.2.2 Selecting the optimal window size for the $Q_{RM}$ predictor

To select the most informative window size when using relative magnitude of discharge as a predictor, we calculated Conditional Entropy of the target given discharge and the  $Q_{RMC}$ ,  $Q_{RML}$  and  $Q_{RMR}$  predictors for a range of window sizes on the full data set. The definition of the window sizes for the different window types and the Conditional Entropies are shown in Fig. 4.

10

The best (lowest) value of Conditional Entropy was obtained for a time-centered window ( $Q_{RMC}$ ) with  $2 \cdot 32 + 1 = 65$  h of a total width. We used this value for all further analyses.

### 3.4.2.3 Model classification, selection, and evaluation

#### 3.4.2.3.1 Model classification

15 All the models we set up and tested in this study can be assigned to one of three distinct groups. The groups distinguish both typical situations of data availability and the use of recursive and non-recursive modeling approaches. Models in the *Q-based group* apply exclusively discharge-based predictor(s). For models in the *P-based group*, we assumed that in addition to discharge, precipitation data are also available. This distinction was made because in the literature exist two main groups of event detection methods: One relying solely on discharge data, the other using precipitation data additionally. Finally, models in the *Model-based group* all apply a 2-step recursive approach as discussed in Section 3.2.1.6. In this case, the first model is always from the Q- or P-based group. Later, event predictions at time step  $t - 1$  of the first model application are then, together with additional predictors from the Q- or P-based group, used as a predictor in the second model.

20

#### 3.4.2.3.2 Model selection

25 In order to streamline the model evaluation process, we applied an approach of supervised model selection and gradually increasing model complexity: We started by setting up and testing all possible 1-predictor models in the *Q-* and *P-*based group. From these, we selected the best performing model and combined it with each remaining predictor into a set of 2-predictor models. The best performing 2-predictor model was then expanded to a set of 3-predictor models using each remaining predictor and so forth. For the Model-based group, the strategy was to take the best-performing models from both the *Q-* and the *P-*based group as the first model and then combine it with an additional predictor as explained in Section 3.2.6. In the end, we stopped at 4-predictor models, as beyond the uncertainty contribution due to limited sample size became dominant.

30

Formatado: Título 3

Formatado: Título 3

Formatado: Título 4 HESS ST

### 3.4.3.3.3 Model evaluation

Among models with the same number of predictors, we compared model performance via the Conditional Entropy (target given the predictors), calculated from the full data set. However, when comparing models with different numbers of predictors, the influence of the Curse of Dimensionality curse of dimensionality needs to be taken into account. To this end, we calculated sample-based Cross Entropy and Kullback-Leibler Divergence as described in Section 2.3.22 for samples of size fifty up to the size of the full data set, using the following sizes [50; 100; 500; 1000; 1500; 2000; 2500; 5000; 7500; 10 000; 15 000; 20 000; 30 000; 40 000; 50 000; 60 000; 70 000; 78 912]. To eliminate effects of chance, we repeated the resampling 500 times for each sample size and took their averages. In Appendix A, the resampling strategy and the choice of repetitions are discussed in more detail.

### 3.3 Application II – CPM

The second application aims to compare the performances of ITM and another automatic event identification method in a more familiar perspective. The predictions were performed in a new data set, and, as a measure of diagnostic, concepts from the receiver operating characteristic (ROC) curve quantified the hits and misses of the predictions of both models according to a time series of user-classified events (considered the true value). More about the ROC analysis can be found in Fawcett (2005).

For the comparison, the characteristic point method (CPM) was chosen because, in contrast with the data-driven ITM, it is a physically-based approach for event identification, which is applicable and recommended to the characteristics of the available data set (hourly time scale data on catchment precipitation and discharge) and open source. The essence of the method is to characterize flow events with three points (start, peak(s) and end of the event) and then associate the event to a corresponding rainfall event (Mei and Anagnostou, 2015). For the event identification, a baseflow separation is previously needed and proposed by coupling the revised constant k method (Blume et al., 2007) and the recursive digital filter proposed by Eckhardt (2005). More about CPM can be attained in Mei and Anagnostou (2015).

Since the outcome of CPM is dichotomous, such as event and non-event, the probabilistic outcome of ITM must be converted into a binary solution. The binarization was reached in the study by choosing an optimum threshold of the probabilistic prediction ( $p_{\text{threshold}}$ ), where all time steps with probabilities equal or greater than it were classified as being part of an event.

The objective function of the optimization was based on the ROC curve and sought to minimize the distance to the top-left corner of the ROC curve, i.e., the Euclidean distance between the true positive rate ( $R_{\text{TP}}$ , proportion of events correctly identified in relation to the total of true events) and false positive rate ( $R_{\text{FP}}$ , proportion of false events in relation to the total of true non-events) to the perfect model (where  $R_{\text{TPperfect}} = 1$  and  $R_{\text{FPperfect}} = 0$ ), as expressed in the Eq. (7). A discussion about the cut-off values of ROC curve can be found in Habibzadeh et al. (2016).

$$\min \sqrt{(1 - R_{\text{TP}})^2 + (0 - R_{\text{FP}})^2} \quad (7)$$

Even though the physically-based CPM method theoretically does not require a calibration step, for avoiding misleading comparison, the parameter  $R_{\text{nc}}$  (rate of no-change, used to quantify null-change ratio in recession coefficient k) were optimized

by Eq. (7).  $R_{TP}$ ,  $R_{FP} \in [0,1]$  and are calculated as a function of the optimized parameter  $p_{\text{threshold}}$  (for ITM) and  $R_{nc}$  (for CPM).

Due to the  $p_{\text{threshold}}$  and  $R_{nc}$  optimization and to enable the cross-validation of the models in a new data set, the available data were divided into training and testing sets. And, since ITM model requires a minimum data set size to guarantee the model robustness, the holdout split was based on the data requirement of the selected ITM model obtained according to the application I, Section 3.2. Therefore, the training data set was used to build the ITM model of the target and selected predictors, and to calibrate the  $p_{\text{threshold}}$  (needed for the binarization) and  $R_{nc}$ .

After that, the calibrated models (ITM and CPM) were applied in a new data set (testing data set) and measures of quality based on the ROC curve were computed in order to evaluate and compare their performance, such as: i) the true positive rate ( $R_{TP}$ ), which represents the percentage of event classification hits (counting of events correctly classified by the model,  $P_T$ , divided by the amount of the true events in the testing data set,  $P$ ); ii) the false positive rate ( $R_{FP}$ ), which represents the percentage of false events identified by the model (counting of events misclassified by the model,  $P_F$ , divided by the amount of the true non-events in the testing data set,  $N$ ); iii) the accuracy, which reflects the total proportion of events ( $P_T$ ) and non-events (or true negative,  $N_T$ ) that were correctly predicted by the model; and iv) the distance to the perfect model given by the Eq. (7), which represents the norm between the results obtained by the method and a perfect prediction.

## 4 Results and discussion

### 4.1 Results for application I

#### 4.1.1.1 Model performance for the full data set

Here we present and discuss the model results when constructed and applied to the complete data set. As we stick to the complete data set, Kullback-Leibler divergence will always be zero, and model performance can be fully expressed by Conditional Entropy (see Section 3.2.3.3.4.3), with the (unconditional) Shannon Entropy of the target data  $H(e) = 0.516$  bits as an upper limit, which we use as a reference to calculate the relative uncertainty reduction for each model. In Table 2, Conditional Entropies and their relative uncertainty reductions are shown for each  $Q$ - and  $P$ -based 1-predictor model.

1-predictor models based on  $Q$  and  $\ln Q$  reduced uncertainty to about 50 % (models #1-10 in Table 2, column 4), with a slight advantage of  $Q$  over  $\ln Q$ . Interestingly, both show their best results for time offset  $t + 2$ , i.e., future discharge is a better predictor for event detection than discharge at the current time step. As we were not sure whether this also applies to 2-predictor models, we decided to test both the  $t + 2$  and  $t$  predictors of  $Q$  and  $\ln Q$  in the next step. Compared to  $Q$  and  $\ln Q$ , relative magnitude of discharge  $Q_{RMC}$  and discharge slope  $Q_{\text{slope}}$  performed poorly, and so did  $P$ , the only model in the  $P$ -based group. Most likely this is because for a certain time step, to be part of an event is not so much dependent on precipitation at this particular time step, but rather on the accumulated rainfall in a period preceding it. Despite its poor performance, we decided to use it in higher-order models to see whether it becomes more informative in combination with other predictors.

Formatado: Título 3

Based on these considerations and the model selection strategy described in Section 3.2.3.23.4.2, we built and evaluated all possible 2-predictor models. The models and results are shown in Table 3.

As could be expected from the information inequality, adding a predictor improved the results, and for some models (#16 and #20), the  $t$ -predictors outperformed their  $t + 2$  counterparts (#17 and #21, respectively). Once more,  $Q$  predictors performed slightly better than  $\ln Q$ , such that for all higher-order models, we only used  $Q(t)$  and ignored  $Q(t + 2)$ ,  $\ln Q(t)$  and  $\ln Q(t + 2)$ .

In the  $P$ -based group, adding any predictor greatly improved results by about 50 %, but not a single  $P$ -based model outperformed even the worst from the  $Q$ -based group.

Finally, from both the  $Q$ - and  $P$ -based group, we selected the best model (#16 and #23, respectively) and extended them to 3-predictor models with the remaining predictors. The models and results are shown in Table 4.

Again, for both models, the added predictor improved results considerably, and we used both of them to build a recursive 4-predictor model as described in Section 3.2.33.4. The new predictor,  $e_p(t - 1)$  is simply the probabilistic prediction of a model (#27 or #28, in this case) for time step  $t - 1$  of being part of an event, with value range  $[0, 1]$ . This means  $e_{p\#27}(t - 1)$  carries the memory from the previous predictions of model #27 (and  $e_{p\#28}(t - 1)$  from model #28, accordingly), and the new 4-

predictor models #29 and #30 as shown in Table 5 are simply copies of these models, extended by a memory term:  $e_p(t - 1)$ . Again, model performance improved, and model #29 was the best among all tested models. Though, so far the effect of sample size was not considered, which might have a strong impact on the model rankings. This is investigated in the next section.

#### 4.24.1.2 Model performance for samples

The sample-based model analysis is computationally expensive, so we restricted these tests to a subset of the models from the previous section. Our selection criteria were to i) include at least one model from each predictor group; ii) include at least one model from each dimension of predictors; and iii) choose the best performing model. Altogether we selected the seven models shown in Table 6. Please note that despite our selection criteria, we ignored the 1-predictor model using precipitation due to its poor performance.

For these models, we computed the Cross Entropies between the full data set and each sample size  $N$  for  $W$  repetitions and, in the end, for each sample size  $N$ , we took the average of the  $W$  repetitions thereof. This was done separately for each sample size. The results are shown in Fig. 5. For comparison, the Cross Entropies between the target data set and samples thereof are also included and labeled as model #0.

In Fig. 5, the Cross Entropies at the right end of the x-axis, where the sample contains the entire data set, equal the Conditional Entropies, as the effect of sample size is zero. However, with decreasing sample size, Cross Entropy grows in a non-linear fashion as  $D_{KL}$  starts to grow. If we walk through the space of sample sizes in the opposite direction, i.e., from left to right, we can see that as the samples grow, the rate of change of Cross Entropy decreases, the reason being that the rate of change of  $D_{KL}$  decreases, which means that the model learns less and less from new data points. Thus, by visually exploring these

Formatado: Título 3



“learning curves” of the models we can make two important statements related to the amount of data required to inform a particular model: We can state how large a training data set should be to sufficiently inform a model, and we can compare this size to the size of the actually available data set. If the first is much smaller than the latter, we gain confidence that we have a well-informed, robust model. If not, we know that it may be beneficial to gather more data, and if this is not possible, we should treat model predictions with caution.

As mentioned in Section 2.3.2, besides Fig. 5 informing the amount of data needed to have a robust model (implying that sample size is enough to represent the full data set), it allows comparing competing models with different dimensions and select the optimal number of predictors (taking advantage of the available information and avoiding overfitting). In this sense, in the  $P$ -based group and for sample sizes smaller 5000, the 1-predictor model #23 performs best, but for larger samples sizes, the 4-predictor model #30 takes the lead. Likewise, in the  $Q$ -based group and for sample sizes smaller than 2500, the single-predictor model #3 is the best but is outperformed by the 2-predictor model #16 from 2500 until 10000, which in turn is outperformed by the 4-predictor model #29 from 10000 to the end. Across all groups, models #3, #16 and #29 form the lower envelope curve in Fig. 5, which means that one of them is always the best model choice, depending on the sample size.

Interestingly, the best performing model for large sample sizes (#29) includes predictors which reflect the definition criteria that guided manual event detection (Section 03-1):  $Q(t)$  and  $Q(t + 2)$  contain information about the absolute magnitude of discharge,  $Q_{RMC}$  expresses the magnitude of discharge relative to its vicinity, and  $e_{p\#27}(t - 1)$  relates to the requirement of events to be coherent.

We also investigated the contribution of sample-size effects to total uncertainty by analyzing the ratio of  $D_{KL}$  and  $H(X|Y)$  as described in Section 2.3. In Fig. 6, these ratios are shown as a function of sample size for each model.

The values in Fig. 6 are given as percentages, e.g., a value of 200 % means that the contribution of sample-size effects to total uncertainty is twice as high as the contribution from Conditional Entropy of the full data set. As expected, for all models these the contribution of sample-size effects to total uncertainty percentages decreases with increasing sample size, but the absolute values and the rate of change strongly differ. For the 1-predictor model #3,  $D_{KL}$  contribution is small already for small sample sizes (circa 65 % for sample size equal 50), and it quickly drops to almost zero with increasing sample size. For multi-predictor models such as #29,  $D_{KL}$  contribution to uncertainty exceeds that of  $H(X|Y)$  by a factor of six-seven for small samples (circa 700 % for sample size equal 50), and it decreases only slowly with increasing sample size.

In Table 7 (column 5), we show for each model the minimum sample size to keep the  $D_{KL}$  contribution below a threshold of 5 %.

As expected, the models with few predictors require only small samples to meet the 5 % requirement (starting from a subset of 12.6 % of the full data set for 1-predictor model to 37.3 % for 2-predictor), but for multi-predictor models such as models #29 and #30, more than 60000 data points are required (87.6 % and 79.4 % of the full data set, respectively). This happens because the greater the number of predictors, the greater is the number of bins in the model. This means that we need a much larger data set to populate the PMF with the largest number of bins, for example: model #29 has 279752 bins and requests 7.9

years of data). Considering that the amount of data available in the study is limited, this also means that increasing the number of predictors/bins, the risk of creating an overfitted/non-robust model also increases. Thus, the ratio  $D_{KL}/H(X|Y)$  and visual inspection of the curve in Fig. 5 orientate the user when to stop adding new predictors to avoid overfitting. In this fashion, Table 7 shows that each of the models tested meets the 5 % requirement, claiming up to 87.6 % of the available data set (69 102 out of 78 912 data points for model #29), which indicates that all of them are robustly supported by the data. In this case, we can confidently choose the best-performing among them (#29, with uncertainty equal to 0.114 bits) for further use. Interestingly, with this analysis, it was also possible to identify the drivers of the user classification, which, in the case of model #29, were the predictors  $Q(t)$ ,  $Q_{RMC}$ ,  $Q(t + 2)$ , and  $e_p(t - 1)$ .

#### 4.34.1.3 Model application

In the previous sections, we developed, and compared and validated a range of models to reproduce subjective, manual identification of events in a discharge time series. Given the available data, the best model was a 4-predictor recursive model applying  $Q(t)$ ,  $Q_{RMC}$ ,  $Q(t + 2)$ , and  $e_p(t - 1)$  as predictors and built with the full data set (#29, Table 7). This model reduced the initial predictive uncertainty by 77.8 %, decreasing Conditional Entropy from 0.516 to 0.114 bits. This sounds reasonable, but what do the model predictions actually look like? As an illustration, we applied the model to a subset of the training data, April 26 to June 21, 2001. For this period, the observed discharge, the manual event classification by the user and the model-based prediction of event probability are shown in Fig. 76.

In the period from June 01 to 21, four distinct rainfall-runoff events occurred which were also classified as such by the user. During these events, the model-based predictions for event probability remained consistently high except for some times at the beginning and end of events, or in times of low flow during an event. Obviously, the model here agrees with the user classification, and if we wished to obtain a binary classification from the model, we could get it by introducing an appropriate probability threshold (as further described in Section 4.2). (e.g. 75 %).

Things look differently, though, in the period of April 26 to May 10, when snowmelt induced diurnal discharge patterns. During this time, the model identified several periods with reasonable (above 50 %) event probability, but the user classified none. Arguably, this is a difficult case for both manual and automated classification, as the overall discharge is elevated, but not much, and diurnal events can be distinguished, but are not pronounced. In such cases, both the user-based and the model-based classification are uncertain and may disagree.

To identify snowmelt events or potentially improve the information contained in the precipitation set, other predictors could have been used in the analysis (such as aggregated precipitation, snow depth, air temperature, nitrate concentrations, moving average of discharge, etc.) or the target could have been classified according to its type (rainfall, snowmelt, upstream reservoir operation, etc.), instead of having a dichotomous outcome, such as event and non-event. The choice of target and potential predictors occurs according to user interest and data availability.

Formatado: Título 3, Recuo: À esquerda: 0 cm, Primeira linha: 0 cm

Another point that may be of interest to the user is the improvement of the consistency of event duration. This can be reached by selection of predictors or through a post-processing step. As previously discussed in Section 3.2.1.6, by applying a recursive predictor  $e_p(t - 1)$  a memory effect is incorporated into the model, bringing some inertia for the transition from event to no-event or vice versa. If it is the user interest, the memory effect could be further enhanced by adding more recursive predictors, such as  $e_p(t - 2)$ ,  $e_p(t - 3)$ , and so on. An alternative option to clear very-short discontinuous time steps or very-short events would be to increase event coherence in a post-processing step with an autoregressive model, with model parameters found by maximizing agreement with the observed events.

Finally, in contrast to the evaluation approach presented, where the subsets are compared to the full data set (subset data plus data not seen during training), the next section will present the evaluation of ITM and CPM applied for mutually exclusives training and testing sets.

#### 4.2 Results for application II

Section 4.1 showed that, for the full data set, the best model was the recursive one with  $Q(t)$ ,  $Q_{RMC}$ ,  $Q(t + 2)$ , and  $e_p(t - 1)$  as the drivers of the user classification (model #29, Table 7), which could be robustly built with a sample size of 69 102. Thus, to assure its robustness for the second application, since we are creating a new PMF based only on the training data set, the split of the data sets (discharge, precipitation and user event classification) divided the 78 912 time steps in two periods composed by: i) 87.6 % of the full data set (69 102 time steps) forming the training data set (from Oct/31/1996 1:00 to Sep/18/2004 6:00); and the remaining 12.4 % (9810 time steps) the testing data set (from Sep/18/2004 7:00 to Nov/01/2005 00:00). The characteristics of the user event classification data set, used as the true classification for accounting the hits and misses of ITM and CPM, is presented in Table 8.

For model training, input data from both models, ITM and CPM, were smoothed. First, a 24-hour moving average was applied to the discharge of CPM (this was recommended by the first author of the method in personal communication) and, to avoid misleading comparison, it was then applied to the probabilities of ITM right before the binarization. The smoothing improved the results of both models and worked as a post-processing filter which removed some noise (very-short-duration events) and attenuated effects from snow melting. Note that this is a feature of our training data set, and therefore it is not necessarily applicable to other similar problems and neither is a required step.

Following the data smoothing, we proceeded with the optimization of parameters: threshold for the probability output of ITM and rate of no-change for CPM (Section 3.3). The results of the two models also improved with the optimization performed. The optimum parameters obtained were  $p_{\text{threshold}} = 0.26$  and  $R_{nc} = -6.6$ . For these values, the final distances in the training data set given by Eq. (7) were: 0.05 and 0.23 for ITM and CPM, respectively.

After the model training, the calibrated models were applied to the testing data set to predict binary events. The event predictions were then compared to the true classification (Table 8, training line) and their hits and misses were calculated in order to evaluate and compare their performance. The results are compiled in Table 9.

The quality parameter presented in Table 9 show that the ITM true positive rate equals 96.6 %, i.e., 12.2 % higher than CPM ( $R_{TP}$ ). On the other hand, CPM false negative rate is equal to 9.9 %, while ITM  $R_{FP}$  is equal to 11.7 % (1.8 % higher). These results indicate that ITM is more likely to predict events than the CPM but at the cost of increasing the false positive rate. Combining these two rates into a single success criterion according to Eq. (7) showed ITM to be slightly superior to CPM.

5 Considering only the hits of the models, both models performed similarly, reaching almost 90 % of accuracy, with CPM being slightly better than ITM. However, it should be emphasized that although the accuracy of the model gives a good notion of the model hits, it was not used as a criterion for success because it is a myopic criterion for the false event classifications. False positives are essential in the context of event prediction, since most of the data are non-events (88.2 % of the training data set, Table 8), and a blind classification of all time steps as being non-event, for example, would overcome the accuracy obtained  
10 by both models (90.4 % of the testing data set, Table 8), even though it is not a useful model.

As an illustration, in the context of the binary analysis, the observed discharge, the true event classification (manually made by an expert), ITM predicted events and CPM predicted events are shown in Fig. 7 for a subset of the testing data, April 22 to August 23, 2005.

For the analyzed subset, nine distinct rainfall-runoff events occurred and were identified as such by ITM and CPM. However,  
15 differently from the true identification, both models grouped some of these events (July 20, August 07 and 16) in events with longer duration. False events were also observed in both models, where three false events were identified by ITM (July 5, July 7 and July 26) and two (but contemplating the same period as ITM) by CPM. It should be noted that they are false in relation to the user classification; however, we can not exclude the possibility of false classification by the visual inspection process. A further criticism is that the holdout cross-validation involves a single run, which is not as robust as multiple runs.  
20 Nevertheless, the way that the split was proposed recognizes the logical order of obtaining the data. Thus, despite the subjectivity of event selection by a user and the application of a simplified method of cross-validation, it is possible to conclude that, overall, ITM and CPM behaved similarly and provided reasonable predictions, as seen numerically in Table 9 and qualitatively through Fig. 7.

An interesting conclusion is that ITM was able to overcome CPM requiring only discharge data and a training data set of  
25 classified events (also based on the discharge set), while CPM demanded precipitation, catchment area and discharge as inputs. On the other hand, since CPM is a physically-based approach, it does not require a training data set with identified events (although the optimization in the calibration step has representatively improved its results) and there are no limitations in terms of data set size, which disappears with the robustness analysis, being then a method more easily implemented. The binarization of the ITM predictions and parameter optimization in CPM are not included in the original methods, however, they were  
30 essential adaptations to allow the comparison of the models. Finally, the suitability or not of the existing event detection techniques depends mainly on the user's interest and the data available for application.

## 5 ~~Summary and C~~onclusions

Typically, ~~for a user with a specific purpose and a particular discharge time series,~~ it is easy to manually identify rainfall-runoff events, due to the high discriminative and integrative power of the brain-eye system. However, this is i) cumbersome for long time series; ii) subject to handling errors; and iii) hard to reproduce since it depends on acuity and knowledge of the event identifier. To mitigate these issues, this study has proposed an ~~Information Theory~~information theory approach to learn from data, and to choose the best predictors, via uncertainty reduction, to create predictive models for automatically identifying rainfall-runoff events in discharge time series.

The method was established in four main steps: model hypothesis, building, evaluation, and application. Each association of predictor(s) to the target is equivalent to formulating a model hypothesis. For the model building, non-parametric models were constructing discrete distributions via bin-counting. ~~To proceed with the rainfall-runoff event identification, the predictive model required requiring~~ at least a discharge time series, and a training data set containing a Yes/No event identification as a target. In the evaluation step, we used Shannon Entropy and Conditional Entropy to select the ~~best~~more informative predictors, and Kullback-Leibler Divergence and Cross Entropy to analyze the model in terms of overfitting and ~~Curse of Dimensionality~~curse of dimensionality. Finally, the best model was applied to its original data set to compare the predictability of the events. ~~For benchmarking purpose, a holdout cross-validation a comparison of the proposed data-driven method with an alternative physically-based approach was performed.~~

The approach was applied to discharge and precipitation data from the Dornbirnerach catchment in Austria. In this case study, 30 models based on 14 predictors were built and tested. Among these, seven predictive models with a number of predictors varying from one to four were selected: ~~four models built purely arising from discharge-based predictor(s) and three from discharge and precipitation data, of which one in each group was a recursive model, i.e., they used their own predictions of a previous time step as an additional predictor.~~ Interestingly, across these models, the three best performing ones were obtained using only discharge-based predictors. The overall best model was a recursive one applying four predictors: discharge from two different time steps, the relative magnitude of discharge ~~compared to all discharge values in a surrounding-in-a~~ 65-hour time window and event predictions from the previous time step. Applying the best model, the uncertainty about event classification was reduced by 77.8 %, decreasing Conditional Entropy from 0.516 to 0.114 bits. Since the Conditional Entropy reduction of the models with precipitation was not higher than the ones exclusively based on discharge information, it was possible to infer that: i) the information coming from precipitation was likely already contained in the discharge data series; and ii) the event classification is not so much dependent on precipitation at a particular time step, but rather on the accumulated rainfall in the period preceding it. ~~Furthermore, precipitation data are often not available for analysis, which makes the model exclusively based on discharge data even more attractive.~~

~~A robustness~~Further analysis using Cross Entropy and Kullback-Leibler Divergence showed that the robustness of a model quickly dropped with the number of predictors used (an effect known as the ~~Curse of Dimensionality~~curse of dimensionality) and that the relation between number of predictors and sample size was crucial to avoid overfitting. Thus, the model choice is

a tradeoff between predictive power and robustness given the available data. For our case, the minimum amount of data to build a robust model varied from 9952 data points (1-predictor model with 0.260 bits of uncertainty) to 69 102 data points (4-predictor model with 0.114 bits of uncertainty). Complementarily, the quality of the model was verified in a more traditional way, by a cross-validation analysis (where the model was built and validated in a training data set), and a comparative investigation between our data-driven approach and a physically-based model. As a result, in general, both models presented reasonable predictions and reached similar quality parameters, with almost 90 % of accuracy. In the end, the comparative analysis and cross-validation reinforced the quality of the method, previously validated in terms of robustness using measures from information theory.

Since a rainfall runoff event does not have a clear definition, and its identification is dependent on the catchment and the user's interest, the final objective of the predictive model was to reproduce the users' Yes/No event decision in a probabilistic manner for application to further data sets. Considering the subjectivity of the event classification process and the high dependence of the user's background knowledge, a learning from data approach was applied to understand and replicate the user event classification. We used measures from Information Theory to quantify the minimum amount of data needed to build a robust model, or to choose the most suitable model for an available data set. The resulting automatic event classification also provides a probabilistic event prediction, which compared to a binary classification additionally provides a measure of the confidence of the classification.

In the end, the data-driven approach based on ~~Information Theory~~ information theory is a consolidation of descriptive and experimental investigations, since it allows to describe the drivers of the model through predictors and it investigates the similarity of the model hypothesis with respect to the true classification. In summary, it presents advantages such as: i) It is a general method that involves a minimum of additional assumptions or parametrizations; ii) Due to its non-parametric approach, it preserves as much as possible the full information of the data, which might get lost when expressing the data-relations by functional relationships; iii) It obtains data-relations from the data itself; iv) It is flexible in terms of data requirement and model building; v) It allows to measure the amount of uncertainty reduction via predictors; vi) It is a direct way to account for uncertainty; vii) It permits explicitly comparing information from various sources in a single currency, "bit"; viii) It allows to quantify minimal data requirements; ~~ix)iii~~ It enables to investigate the ~~Curse of Dimensionality~~curse of dimensionality; x) It is a way of understanding the drivers (predictors) of the model (also useful in machine learning, for example); xi) It permits to choose the most suitable model for an available data set; and xii) The predictions are probabilistic, which compared to a binary classification additionally provides a measure of the confidence of the classification.

Although the procedure was employed to identify events from a discharge time series, which for our case were mainly triggered by rainfall and snowmelt, the method can be applied to reproduce user classification of any kind of event (rainfall, snowmelt, upstream reservoir operation, etc.) and even identify them separately. Moreover, one of the strengths of the data-based approach is that it potentially accepts the approach permits any data to serve as predictors, and it can handle any kind of relation between the predictor(s) and the target. Thus, the proposed approach can be conveniently adapted to another practical application.

## 6 Data availability

The Event Detection program, containing the functions to develop multivariate histograms and calculate ~~Information Theory~~information theory measures, is published alongside this manuscript via GitHub <https://github.com/KIT-HYD/EventDetection>. The repository also includes scripts to exemplify the use of the functions, and the data set of identified event, discharge and precipitation time series from the Dornbirnerach catchment in Austria used in the case study.

## 7 Author contribution

UE and PD developed the model program (calculation of ~~Information Theory~~information theory measures, multivariate histograms operations, event detection) and developed a method to avoid infinitely large values of  $D_{KL}$  (Darscheid [et al., 2018](#)); ~~Guthke and Ehret, 2018~~). ST performed the simulations, cross-validation, parameter optimization, comparative analysis to a second model and the validation justification of the number of repetitions of the resampling stage. ST and UE directly contributed to the design of the method and test application, to the analysis of the performed simulations, and wrote the manuscript.

## 8 Acknowledgments

The first and third author acknowledge support by Deutsche Forschungsgemeinschaft DFG and Open Access Publishing Fund of Karlsruhe Institute of Technology (KIT). We thank Clemens Mathis from Wasserwirtschaft Vorarlberg, Austria for providing the case study data.

## Appendix A: Resampling strategy and number of repetitions

In the study, samples of size  $N$  from the data set were obtained through bootstrapping, i.e., they were taken randomly, but continuously in time, with replacement among the  $W$  repetitions. For each sample size, we repeated draws  $W$  times and took the average Cross Entropy and  $D_{KL}$  to eliminate effects of chance (see repetition statements  $N$  and  $W$  in Fig. 1).

Thus, in order to find the value of  $W$  which balances statistical accuracy and computational efforts, we did a dispersion analysis through calculating the Shannon Entropy (as a measure of dispersion) of the Cross Entropy distribution of the (unconditional) target model (~~without predictors~~, model #0 in Table 7). Sixty one bins ranging from 0 to 6 in steps of 0.1 bits were used, which contemplates the range of all possible Cross Entropy values among the tested pairs of  $N$  and  $W$ . Fig. A1 presents the Shannon Entropy applied as a dispersion parameter to analyze the effect of the number of repetitions  $W$  for different sample sizes  $N$ .

Considering the graph in Fig. A1, in general, the behavior of the Shannon Entropy among the repetitions is similar for each sample size analyzed, indicating the dispersion of the results according to the number of repetitions does not vary too much, i.e., the bins are similarly filled. However, it is possible to see that, as the sample size increases, the Shannon Entropy for the

different number of repetitions approaches that for the 50 000 repetitions. For sample sizes up to 7500, the bars from 50, 100 and 300 repetitions present some peaks and troughs, indicating some dispersion in filling the bins. Thus, in this case study, the minimum of 500 repetitions was assumed as a reasonable number of repetitions to compute the mean of the Cross Entropy in the sample size investigation. This number of repetitions was also validated considering the smoothness and logical behavior of the curves obtained during the data size validation and [Curse of Dimensionality](#) [curse of dimensionality](#) analyzes (Fig. 5 in Section [4.1.24.2](#)).

## References

Bellman, R.: Dynamic Programming. Princeton University Press, 1957.

[Blower, G. and Kelsall, J. E.: Nonlinear Kernel Density Estimation for Binned Data: Convergence in Entropy, Bernoulli 8, 423–449, 2002.](#)

Blume, T., Zehe, E., and Bronstert, A.: Rainfall-runoff response, event-based runoff coefficients and hydrograph separation, Hydrol. Sci. J., 52(5), 843–862, doi:10.1623/hysj.52.5.843, 2007.

Brunsell, N. A.: A multiscale information theory approach to assess spatial-temporal variability of daily precipitation, J. Hydrol., 385(1–4), 165–172, doi:10.1016/j.jhydrol.2010.02.016, 2010.

Chapman, T. G.: Entropy as a measure of hydrologic data uncertainty and model performance, J. Hydrol., 85(1–2), 111–126, doi:10.1016/0022-1694(86)90079-X, 1986.

Chow, V. T., Maidment, D. R., and Mays, L. W.: Applied Hydrology, McGraw-Hill, New York, 1988.

Cover, T. M. and Thomas, J. A.: Elements of Information Theory, 2nd ed., John Wiley & Sons, New Jersey, 2006.

Darscheid, P.: Quantitative analysis of information flow in hydrological modelling using Shannon information measures, Karlsruhe Institute of Technology, 73 pp., 2017.

[Darbellay, G. A. and Vajda, I.: Estimation of the information by an adaptive partitioning of the observation space. IEEE Trans. Inf. Theory, 45, 1315–1321, 1999.](#)

Darscheid, P., Guthke, A., and Ehret, U.: [A Maximum-Entropy Method to Estimate Discrete Distributions from Samples Ensuring Nonzero Probabilities, Entropy, 20, 601, doi:10.3390/e20080601, 2018.](#) [A maximum Entropy method to estimate discrete distributions from samples ensuring non-zero probabilities. Under review in Entropy, 2018.](#)

Eckhardt, K.: How to construct recursive digital filters for baseflow separation. Hydrol. Process. 19, 507–515, doi:10.1002/hyp.5675, 2005.

[Fawcett, T.: An introduction to ROC analysis Tom. Irbm, 35\(6\), 299–309, doi:10.1016/j.patrec.2005.10.010, 2005.](#)

Gong, W., Yang, D., Gupta, H. V., and Nearing, G.: Estimating information entropy for hydrological data: One dimensional case, Water Resour. Res., (1), 5003–5018, doi:10.1002/2014WR015874, 2014.

[Habibzadeh, F., Habibzadeh, P., and Yadollahie, M.: On determining the most appropriate test cut-off value: The case of tests with continuous results, Biochem. Medica, 26\(3\), 297–307, doi:10.11613/BM.2016.034, 2016.](#)



- Hall, F. R.: Base-Flow Recessions - A Review, *Water Resour. Res.*, 4, 973–983, 1968.
- Hoyt, W. G. and others: Rainfall and run-off in the United States, *Water-Supply Paper 772*, Washington, 1936.
- Koskelo, A. I., Fisher, T. R., Utz, R. M., and Jordan, T. E.: A new precipitation-based method of baseflow separation and event identification for small watersheds (<50km<sup>2</sup>), *J. Hydrol.*, 450–451, 267–278, doi:10.1016/j.jhydrol.2012.04.055, 2012.
- 5 Knuth, K. H.: Optimal Data-Based Binning for Histograms. *arXiv 2013*, [arXiv:physics/0605197v2](https://arxiv.org/abs/physics/0605197v2), 2013.
- Liu, D., Wang, D., Wang, Y., Wu, J., Singh, V. P., Zeng, X., Wang, L., Chen, Y., Chen, X., Zhang, L., and Gu, S.: Entropy of hydrological systems under small samples: Uncertainty and variability, *J. Hydrol.*, doi:10.1016/j.jhydrol.2015.11.019, 2016.
- Lott, D. A., and Stewart, M. T.: Base flow separation: A comparison of analytical and mass balance methods, *J. Hydrol.*, 535, 525–533, doi:10.1016/j.jhydrol.2016.01.063, 2016.
- 10 Mei, Y. and Anagnostou, E. N.: A hydrograph separation method based on information from rainfall and runoff records, *J. Hydrol.*, 523, 636–649, doi:10.1016/j.jhydrol.2015.01.083, 2015.
- Merz, R., Blöschl, G., and Parajka, J.: Spatio-temporal variability of event runoff coefficients, *J. Hydrol.*, 331(3–4), 591–604, doi:10.1016/j.jhydrol.2006.06.008, 2006.
- Merz, R. and Blöschl, G.: A regional analysis of event runoff coefficients with respect to climate and catchment characteristics
- 15 in Austria, *Water Resour. Res.*, 45(1), 1–19, doi:10.1029/2008WR007163, 2009.
- Mishra, A. K., Özger, M. and Singh, V. P.: An entropy-based investigation into the variability of precipitation, *J. Hydrol.*, 370(1–4), 139–154, doi:10.1016/j.jhydrol.2009.03.006, 2009.
- Nearing, G. S. and Gupta, H. V.: Information vs. Uncertainty as the Foundation for a Science of Environmental Modeling, eprint [arXiv:1704.07512](https://arxiv.org/abs/1704.07512), 1–23 [online] Available from: <http://arxiv.org/abs/1704.07512>, 2017.
- 20 Pechlivanidis, I. G., Jackson, B., Mcmillan, H., and Gupta, H. V.: Robust informational entropy-based descriptors of flow in catchment hydrology, *Hydrol. Sci. J.*, 61(1), 1–18, doi:10.1080/02626667.2014.983516, 2016.
- Ruddell, B. L. and Kumar, P.: Ecohydrologic process networks: 1. Identification, *Water Resour. Res.*, 45(3), 1–23, doi:10.1029/2008WR007279, 2009.
- Seibert, S. P., U-Ehret, U., and E. Zehe: Disentangling timing and amplitude errors in streamflow simulations. *Hydrol. Earth*
- 25 *Syst. Sci.* 20 (9), 3745–3763, doi:10.5194/hess-20-3745-2016, 2016.
- Simonoff, J. S.: *Smoothing Methods in Statistics*, Springer, Berlin/Heidelberg, Germany, 1996.
- Solomatine, D. P. and Ostfeld, A.: Data-driven modelling: some past experiences and new approaches, *J. Hydroinformatics*, 10(1), 3, doi:10.2166/hydro.2008.015, 2008.
- Solomatine, D., See, L. M., and Abrahart, R. J.: Data-Driven Modelling: Concepts, Approaches and Experiences, in *Practical*
- 30 *hydroinformatics*, edited by Springer, pp. 17–31, Berlin, Heidelberg., 2009.
- Thiesen, S., Darscheid, P., and Ehret, U.: *Event Detection Method Based on Information Theory*, [Zenodo/GitHub](https://zenodo.org/record/1404638), <http://doi.org/10.5281/zenodo.1404638>, 2018.
- Ujfalussy, G. V.: Comparative analysis of conceptual and data-based hydrological models in the Attert catchment, Luxembourg, Karlsruhe Institute of Technology., 45 pp., 2017.

Weijs, S. V.: Information Theory for Risk-based Water System Operation, Technische Universiteit Delft., 210 pp., 2011.

**Table 1: Target and predictors – Characterization and binning strategy.**

<b>Target (<math>X</math>)</b>	<b>Symbol</b>	<b>Unit</b>	<b>Bins* [start:end]</b>	<b>Number of bins</b>
User-based event classification at time $t$	e	[-]	[0:1]	2
<b>Predictors (<math>Y</math>)</b>	<b>Symbol</b>	<b>Unit</b>	<b>Bins* [start:step:end] &amp; [start:end]</b>	<b>Number of bins</b>
Discharge at times $t-2, t-1, t, t+1, t+2$	$Q(t-2), Q(t-1), Q(t), Q(t+1), Q(t+2)$	$[m^3 s^{-1}]$	[0:0.5:16] & [16:end]	34
Natural logarithm of discharge at times $t-2, t-1, t, t+1, t+2$	$\ln Q(t-2), \ln Q(t-1), \ln Q(t), \ln Q(t+1), \ln Q(t+2)$	$[\ln(m^3 s^{-1})]$	[-3.5:0.2:2.9] & [2.9:end]	34
Relative Magnitude of discharge in time windows (centered, left-ended, right- ended)	$Q_{RMC}, Q_{RML}, Q_{RMR}$	[-]	[0:0.1:1]	11
Discharge slope between time steps $t-1$ and $t$ & time steps $t$ and $t+1$	$Q_{\text{slope}_{\text{before}}}$ $Q_{\text{slope}_{\text{after}}}$	$[m^3 s^{-1} h^{-1}]$	[-50:5:90]	29
Precipitation at time $t$	$P$	$[mm h^{-1}]$	[0:1:30]	31
Model-based event probability at time step $t-1$	$e_p(t-1)$	[-]	[0:0.1:1]	11

\* Bins identified by their central values [leftmost center value : step : rightmost center value]

**Table 2: Conditional Entropy and relative uncertainty reduction of 1-predictor models.**

#	Predictive model ( $X Y$ )	$H(X Y)$ [bit]	$H(X Y)/H(X)^*$
<i>Q</i> -based group			
1	$e   Q(t-2)$	0.269	52.1 %
2	$e   Q(t-1)$	0.264	51.3 %
3	$e   Q(t)$	0.260	50.3 %
4	$e   Q(t+1)$	0.255	49.4 %
5	$e   Q(t+2)$	0.250	48.6 %
6	$e   \ln Q(t-2)$	0.269	52.2 %
7	$e   \ln Q(t-1)$	0.265	51.3 %
8	$e   \ln Q(t)$	0.260	50.4 %
9	$e   \ln Q(t+1)$	0.255	49.4 %
10	$e   \ln Q(t+2)$	0.251	48.6 %
11	$e   Q_{\text{RMC}}$	0.505	97.9 %
12	$e   Q_{\text{slope}_{\text{before}}}$	0.473	91.8 %
13	$e   Q_{\text{slope}_{\text{after}}}$	0.473	91.8 %
<i>P</i> -based group			
14	$e   P$	0.472	91.6 %

\*  $H(X) = H(e) = 0.516$  bits

**Table 3: Conditional Entropy and relative uncertainty reduction of 2-predictor models.**

#	Predictive model $(X Y)$	$H(X Y)$ [bit]	$H(X Y)/H(X)^*$
<i>Q</i> -based group			
15	$e   Q(t+2), Q(t)$	0.226	43.9 %
16	$e   Q(t), Q_{\text{RMC}}$	0.182	35.3 %
17	$e   Q(t+2), Q_{\text{RMC}}$	0.191	37.1 %
18	$e   Q(t), Q_{\text{slopeafter}}$	0.254	49.3 %
19	$e   \ln Q(t+2), \ln Q(t)$	0.233	45.1 %
20	$e   \ln Q(t), Q_{\text{RMC}}$	0.185	35.8 %
21	$e   \ln Q(t+2), Q_{\text{RMC}}$	0.194	37.5 %
22	$e   \ln Q(t), Q_{\text{slopeafter}}$	0.254	49.3 %
<i>P</i> -based group			
23	$e   Q(t), P$	0.248	48.2 %
24	$e   Q(t+2), P$	0.247	48.0 %
25	$e   \ln Q(t), P$	0.249	48.2 %
26	$e   \ln Q(t+2), P$	0.249	48.2 %

\*  $H(X) = H(e) = 0.516$  bits

**Table 4: Conditional Entropy and relative uncertainty reduction of 3-predictor models.**

#	Predictive model $(X Y)$	$H(X Y)$ [bit]	$H(X Y)/H(X)^*$
<i>Q</i> -based group			
27	$e   Q(t), Q_{\text{RMC}}, Q(t+2)$	0.144	28.0 %
<i>P</i> -based group			
28	$e   Q(t), P, Q_{\text{RMC}}$	0.167	32.5 %

\*  $H(X) = H(e) = 0.516$  bits

**Table 5: Conditional Entropy and relative uncertainty reduction of recursive 4-predictor models.**

#	Predictive model $(X Y)$	$H(X Y)$ [bit]	$H(X Y)/H(X)^*$
Model-based group			
29	$e   Q(t), Q_{\text{RMC}}, Q(t+2), e_{p_{\#27}}(t-1)$	0.114	22.2 %
30	$e   Q(t), P, Q_{\text{RMC}}, e_{p_{\#28}}(t-1)$	0.142	27.6 %

\*  $H(X) = H(e) = 0.516$  bits

**Table 6: Models selected for sample-based tests.**

<b>Model group</b>	<b>1 predictor</b>	<b>2 predictors</b>	<b>3 predictors</b>	<b>4 predictors</b>
<i>Q</i> -based group	$Q(t)$ #3	$Q(t), Q_{\text{RMC}}$ #16	$Q(t), Q_{\text{RMC}}, Q(t+2)$ #27	-
<i>P</i> -based group	-	$Q(t), P$ #23	$Q(t), P, Q_{\text{RMC}}$ #28	-
Model-based group with <i>Q</i> -based predictors	-	-	-	$Q(t), Q_{\text{RMC}}, Q(t+2), e_{p_{\#27}}(t-1)$ #29
Model-based group with <i>P</i> -based predictors	-	-	-	$Q(t), P, Q_{\text{RMC}}, e_{p_{\#28}}(t-1)$ #30



Table 7: **Application I** – Curse of Dimensionality and data size validation for models in Table 6.

#	Predictive model	H(X) [bit]	H(X)/H(X)*	Sample size where $D_{KL}/H(X) \leq 5\%$ <b>and % of the full data set**</b>	Sample size [a]	Number of bins
0	$e$	0.516	100 %	$\geq 4398$ (5.6 %)	0.5	2
#	Predictive model	H(X Y) [bit]	H(X Y)/H(X)*	Sample size where $D_{KL}/H(X Y) \leq 5\%$ <b>and % of the full data set**</b>	Sample size [a]	Number of bins
3	$e   Q(t)$	0.260	50.4 %	$\geq 9952$ (12.6 %)	1.1	68
16	$e   Q(t), Q_{RMC}$	0.182	35.3 %	$\geq 29460$ (37.3 %)	3.4	748
23	$e   Q(t), P$	0.248	48.2 %	$\geq 18880$ (23.9 %)	2.2	2108
27	$e   Q(t), Q_{RMC}, Q(t+2)$	0.144	28.0 %	$\geq 60178$ (76.3 %)	6.9	25432
28	$e   Q(t), P, Q_{RMC}$	0.167	32.5 %	$\geq 50377$ (63.8 %)	5.8	23188
29	$e   Q(t), Q_{RMC}, Q(t+2), e_{P\#27}(t-1)$	0.114	22.2 %	$\geq 69102$ (87.6 %)	7.9	279752
30	$e   Q(t), P, Q_{RMC}, e_{P\#28}(t-1)$	0.142	27.6 %	$\geq 62667$ (79.4 %)	7.2	255068

\*  $H(X) = H(e) = 0.516$  bits

\*\* size of the full data set: 78 912 data points (9 years)

**Table 8: Cross-validation data set – Characteristics of the user event classification set.**

<u>Data set</u>	<u>Time steps classified as positive events (<math>P</math>)</u>	<u>Time steps classified as non-events (<math>N</math>)</u>	<u>Percentage of events (<math>P/T</math>)</u>	<u>Percentage of non-events (<math>N/T</math>)</u>	<u>Total (<math>T</math>)</u>
Training	8150	60 952	11.8 %	88.2 %	69 102
Testing	942	8868	9.6 %	90.4 %	9810
Sum	9092	69 820	11.5 %	88.5 %	78 912

**Table 9: Application II – ITM and CPM performance.**

<b>Event detection method</b>	<b>True Positive (<math>P_T</math>)</b>	<b><math>R_{TP}</math> (<math>P_T/P^*</math>)</b>	<b>False Positive (<math>P_F</math>)</b>	<b><math>R_{FP}</math> (<math>P_F/N^*</math>)</b>	<b>Accuracy % (<math>(P_T + N_T^{**}) / (P^* + N^*)</math>)</b>	<b>Eq. (7) distance***</b>
<u>ITM</u>	<u>911</u>	<u>96.7 %</u>	<u>1038</u>	<u>11.7 %</u>	<u>89.1 %</u>	<u>0.12</u>
<u>CPM</u>	<u>796</u>	<u>84.5 %</u>	<u>877</u>	<u>9.9 %</u>	<u>89.6 %</u>	<u>0.19</u>

\*  $P = 942, N = 8868$  (Table 8)

\*\*  $N_T = N - P_F$

\*\*\* Distance to the perfect model of the ROC curve

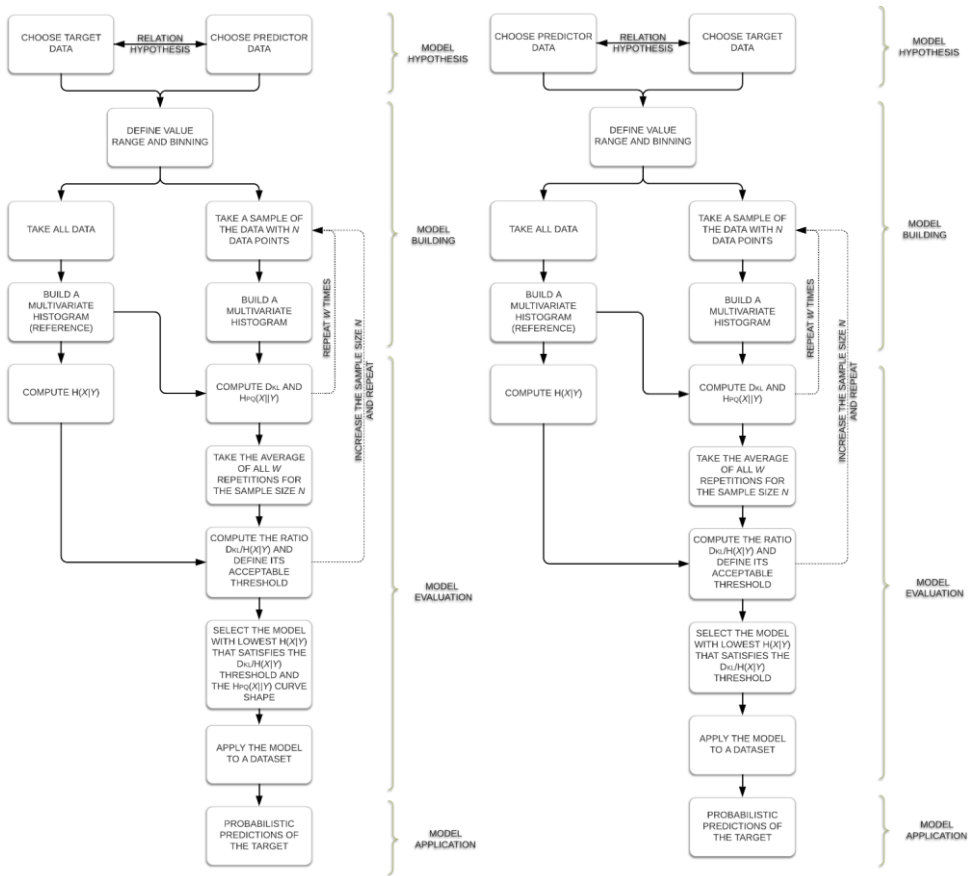
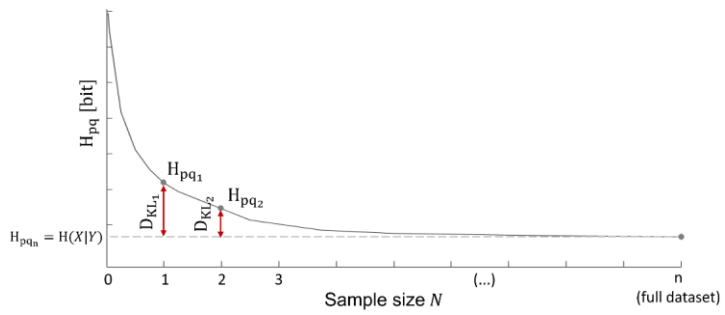


Figure 1: Main steps of the ITM data-driven model based on Information Theory.



**Figure 2: Investigating the effect of sample size through Cross Entropy and Kullback-Leibler Divergence.**

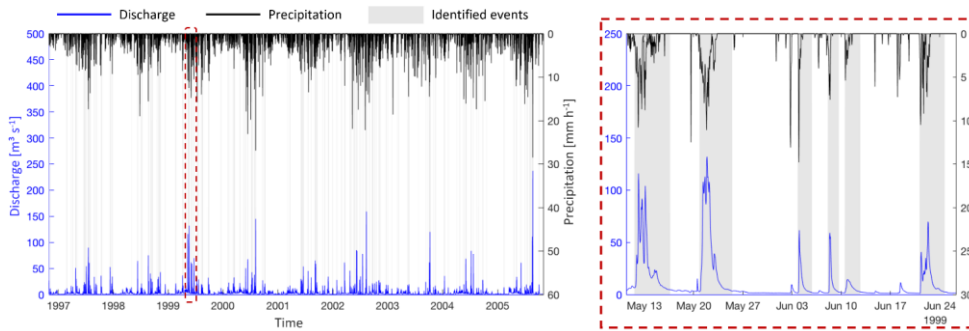


Figure 3: Input data – Discharge (blue), precipitation (black), user-based event classification (grey bars). Overview of the time series on the left and detailed example on the right.

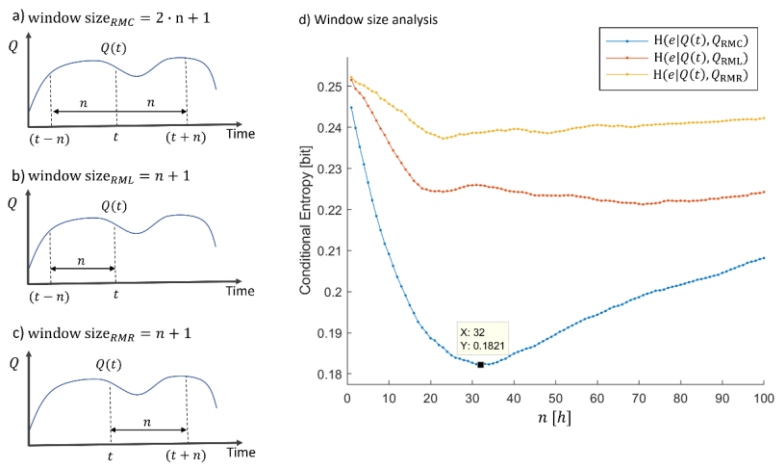


Figure 4: Window size definitions for window types. a)  $Q_{RMC}$ , b)  $Q_{RML}$ , c)  $Q_{RMR}$  window definitions, and d) window size analysis.

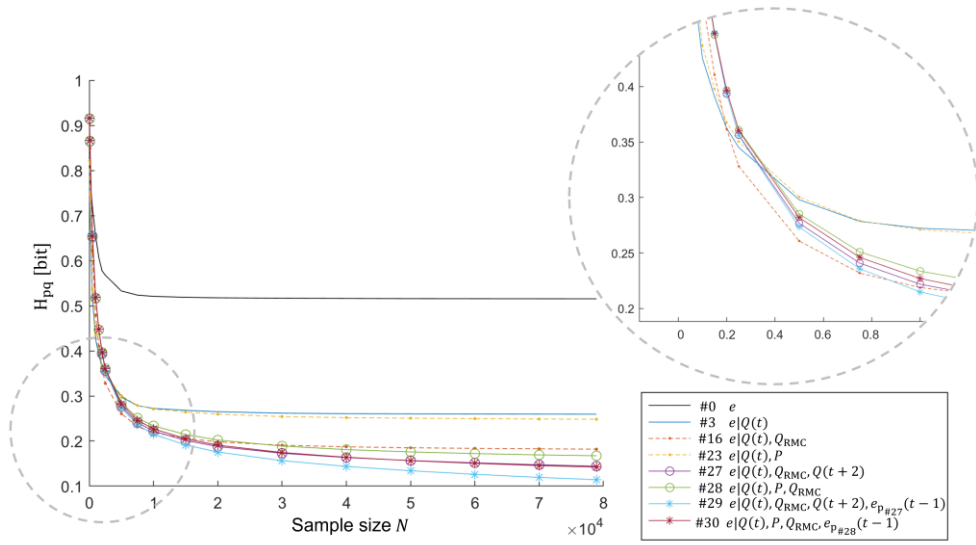


Figure 5: Cross Entropy for models in Table 6 as a function of sample size.



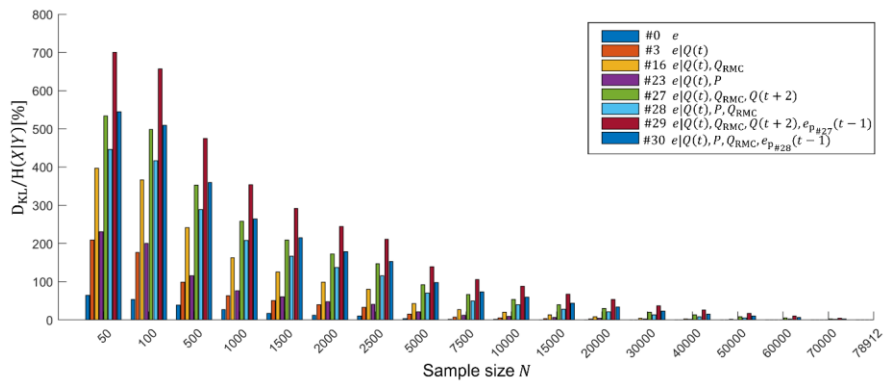


Figure 6: Curse of Dimensionality for models in Table 6 measured by the ratio of  $D_{KL}$  and  $H(X|Y)$ .

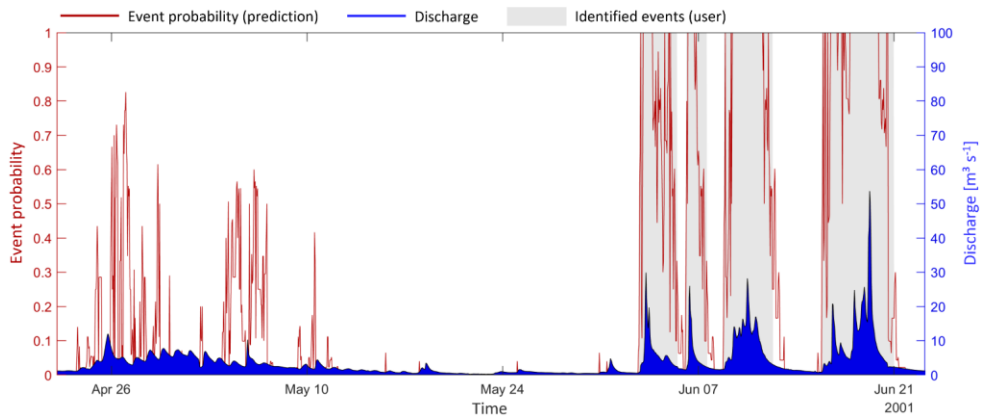
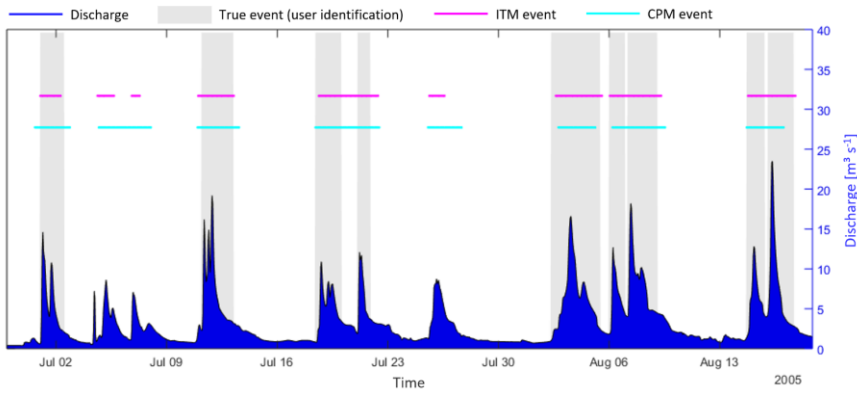


Figure 76: Application 1 – Probabilistic prediction of 4-predictor model #29 (Table 5) to a subset of the training data. Observed discharge (blue), user-based binary event classification (grey bars) and model-based prediction of event probability (red).



**Figure 7: Application II – Binary prediction of ITM and CPM to a subset of the testing set.**

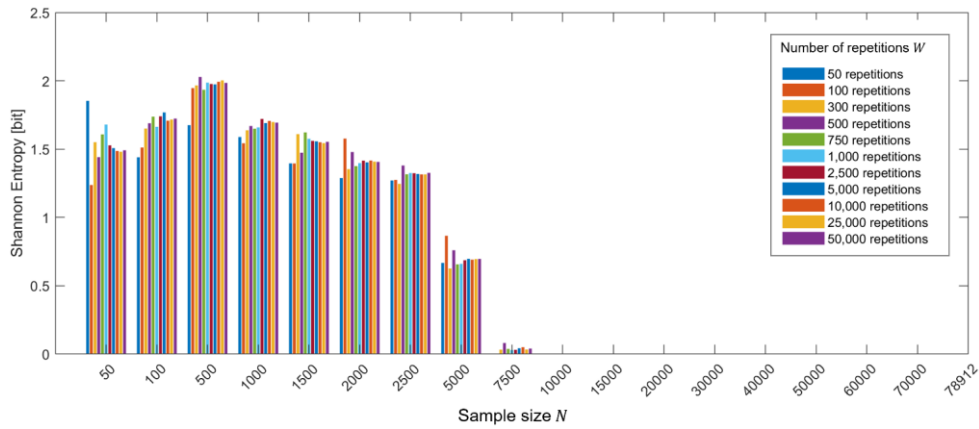


Figure A1: Dispersion analysis of the Cross Entropy. The effect of the number of repetitions in the target model ( $\#$  in Table 7).