Hydrol. Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/hess-2018-606-RC1, 2019

Hydrology and Earth System Sciences
Discussions

Interactive comment on "Quantitative precipitation estimation with weather radar using a data- and information-based approach" *by* M. Neuper and U. Ehret

Anonymous Referee #1

Received and published: 13 March 2019

The paper, in general, is quite well written and well structured, and for sure will be interesting to the readers of HESS. Measures of model quality based on information theory add an additional facet to the problem of model assessment and model selection (which traditionally are based mainly on estimates of mean (squared) model error assuming deterministic models). The conducted experiments are comprehensive, very well designed and presented. Illustrative material is adequate. Conclusions are based on evidence from experiments. It can be considered for publication, however there are some issues outlined below which require attention before publication can be recommended. —-

On the references to earlier literature. Immediately after the authors pose the objectives of this work they write: "Comparable approaches have been suggested by Sharma and Mehrotra (2014) and Thiesen et al. (2018)", but in the rest of the paper you do not cite these authors again, and don't mention what are the differences of your approach and the one taken in the papers mentioned. It is suggested to present briefly the essence of the approaches already published, and to explain the advances made in this paper, and formulate its novelty w.r.t. earlier work.

The same can be said about a citation of Yang et al. 2017 or Kirstetter et al. 2015 which is said to be "a similar approach". (It also remains unclear, what this this approach is similar to. Have these researchers also use probabilistic Z-R relationship instead of the deterministic one?)

A general comment: it is suggested, if any reference to earlier work is given, to specify what was done in that work, its main conclusion, and what does this mean for this work.

P 2, L 27: the aim of the paper is formulated: "aim of this paper to suggest and apply a framework which comprises expressing relations among data directly by empirical discrete probability distributions (dpd's), and measuring the strength of relations and remaining uncertainties with measures from Information Theory." However the paper title says that you want to estimate precipitation (QPE). In my view these are two different objective (albeit related). It is also unclear what is "estimation of precipitation" exactly - is it deterministic, or probabilistic? It is therefore suggested to formulate the objectives clearer, and to relate them to the title, and to the section 1.1. "design of experiments".

In Sec 1 there is a block of text, before 1.1, with a number. Suggest to give it a title and number it as well.

P4, L11-13: Eq 1 uses log2 but in the text you mention other options. Please coordinate better. L18: definition of information was about an "event", and on L18 you switch to "signal" (indeed this latter is what the most work on information theory traditionally use).

I suggest to think of a more consistent terminology for this paper, or explain how event and signal relate to each other.

P5 L5: terms expected information and expected uncertainty — here I would use quotes around the terms. L10: why Entropy starts with is the capital letter, and information not? Suggest to be consistent. L15: While –> while

L27: please introduce what is set Y. Are you guessing X, or you are guessing a realisation of X, i.e. xi?

L32: you write Y=yi, but in Eq 3 it us just y

Overall comment on section 2.1: this is a brief introduction to information and entropy, and it uses terminology common for I.T. textbooks. It will be clear to those who know it, but I am not sure all is clear to those who have not used information theory before. It is also somewhat different form the terminology used in the rest of the paper. What is "event"? What is a set X={x1...xn}? Is X a sample from the (same?) distribution? Is its pdf known? Is X a time series as well? An example from hydrometeorology would have helped a lot - assuming, if I understand correctly, that X is some random variable related to rainfall (is its distribution assumed to be known?), and xi are its realisations (is this right?).

P6 L26: "true distribution and a model thereof" - it is first time these terms are used, and a problem of building a model of the true distribution is mentioned. If indeed it is the problem to be solved in this paper, I would suggest to present it earlier.

P7 L3: 2.2. is Methods, but 2.1 was also methods, right? Of Information theory does not belong to Methods? Subsection is entitled: "Data-based models and predictions, information-based model evaluation", so it is also about measures (since "evaluation" is based on measures) - but 2.1 had also "measures for information theory". Suggestion: to coordinate the titles of various sections to prevent overlaps.

Reading this subsection, I was expecting to find the "data-based models", but could

not... It is suggested to clarify what is meant by these and to present them, or not to use this term. (When seeing the term "data-based, or "data-driven" models, I would expect to see a model build on data, using statistical or machine learning techniques and able to make a prediction of a variable (predictant) based on several other variables (predictors).)

L17: "we lose the information about the absolute and relative position of the data tuples in the data set" - unclear. (Do you mean we lose the time stamp of each data tuple (since this are time series)? If so, this is of course always true when a time series is represented as a pdf.)

P8 L1: "statement about the target value" - I don't see a statement about a "value" (value meaning a real valued estimate of the target), just pdf. What is presented on (a very useful and informative) Fig 1 is estimation of predictive uncertainty conditional on the model output (being the range [-2, -2.3]. This uncertainty is estimated based on ALL data across the whole time domain. (Please see e.g. papers by Todini on predictive uncertainty, and I would suggest to add at least one of them.) This is not a model that predicts the target value *for a given time moment*. (But such model what would nice to have of course.)

P17 L25: "to build better models by simply adding more predictors, which according to the Information Inequality (Eq. 4) never hurts." – well, in theory... If we assume that we are building a predictive model (e.g. predicting R on the basis of radar data), in practice more predictors typically means more complex models (more parameters to calibrate/train), and this could be a problem due to the following. (1) You may break the balance between the amount of available data and number of parameters to train, and (2) more model parameters means increasing the dimension of the search space, and it could mean that during training there is a higher chance to be stuck in a local minimum (e.g. if MLP neural networks are used). So there are good practical reasons to avoid having too many predictors, and (3) more complex model may overfit and not be accurate on cross-validation sets.

P18 L1: I am not sure I understand how overfitting relates to curse of dimensionality by Bellman.

— To summarize the most important comments (and there are more above):

1. The title promises “Quantitative precipitation estimation using a data-based approach”. Perhaps I missed something, but I have not found a data-based model that makes a quantitative estimation of rainfall using radar data. Yes, analysis of distributions of radar data help to provide informative uncertainty estimates of precipitation (and assess the overall information gain) but in my view it is not what is typically understood by QPE (deterministic, and possibly with uncertainty estimates on top).

2. Suggested to present clearly innovation w.r.t. to earlier work.

3. Introduction: suggested to formulate the objectives clearer, and to relate them to the six experiments.

4. On the ways to measure the quality of predictive models. What does real practice need? In essence, the probability distribution-based measures (including the ones using the information theory) give an estimate how far are distributions from each other, rather than measuring the distance between time series, i.e. how far are the deterministic model predictions from (ground truth) observations (on average, or for particular important time moments). In my view it is important to stress this difference. So decision on what measures to use depends on the needs of practitioners and the corresponding goal setting. I would suggest that the best is use both since they measure different things.

5. As far as I understand, the authors used All data to estimate the information gain (and hence to “build” a model allowing for QPE). In other words, there is no data left for cross validation and for testing. If so, this does not correspond to the theory of data-based model building. Is this right?

Editorial comments:

Minor editing of English may be desirable, e.g.: Besides rain gauges with its own limitations: its –> their

Radar data have among other been used for urban hydrology –> Among other data sources, radar data have been used in urban hydrology

its use relies on some sometimes more, sometimes less justified assumptions –> its use relies on some assumptions, which are sometimes justified and sometimes not,

Much work has since then been done –> Much work has since been done

framework which comprises expressing relations among data directly by –> framework which would use relationships between data expressed as the

I suggest shorten some long sentences, and to split long paragraphs (e.g. the first one in the Introduction).

—- The authors are kindly invited to address these comments, revise the paper (or provide a rebuttal), and to resubmit it.

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-606, 2019.