Responses to short comment #1

We thank John Ding for his comments on our manuscript. In the following, we reply to all comments one by one. Comments by John Ding are in blue.

I enjoy reading this Discussion paper from a hydrologic model performance perspective, and suggest the authors consider expanding the list of candidate models (Table 1) and their training methodology as follows:

1. A second–order autoregressive process as a baseline model

The (almost) ignorant model (Model-00) is a baseline model in the popular Nash–Sutcliffe model efficiency (NSE) criterion (e.g., Knoben et al., 2019, and SC1 therein for my comment).

As an alternative to it, I've suggested a simple(st) autoregressive model of order 2, AR(2):

$$Q(t) = Q(t-1) + [Q(t-1) - Q(t-2)] - 2Q(t-1) - Q(t-2) \qquad (1)$$

This and their top–rated Model-07 (AR(3)) belong to the class of autoregressive processes. Their Equation (1) reads, omitting the subscript HOST to the discharge variable Q:

$$Q(t) = 0.0549 + 1.9266Q(t-1) - 1.2071Q(t-2) + 0.2685Q(t-3) \qquad (2)$$

Similarity in terms of the coefficient between the two is striking. In an integer form, both are identical.

It would be instructive to score the performance of all the candidate models by the NSE criterion, both in its original and the newly suggested AR(2) form.

The AR(2)–based NSE will score Model-07 (AR(3)) again as a best performing model. It may differentiate more clearly the one–bucket (Model-02) and the two–buckets (Model-05) one. As expected, the latter performs better than the former (Lines 273-278), but the two are indistinguishable from each other on the authors' proposed model performance scale (Figure 3).

We thank John Ding for suggesting an alternative baseline model, and for suggesting alternative performance scores. We agree that adding more models, and NSE as an additional performance score, would be valuable if the focus of the paper would be about finding an optimal hydrological model for the Dornbinerach watershed. However, the main purpose of the manuscript is about introducing the bit-by-bit method. In this context, the reason behind applying a broad range of model types is to demonstrate the general applicability of the method. The performance of the models themselves is not a central element of the study. We think that giving more room to a discussion of model performance would distract from the key message of the paper rather than strengthening it, therefore we prefer to keep the range of models and performance criteria as is.

Nevertheless, the questions raised by John Ding are interesting, and we are happy to provide some further details in the following:

AR-models

There is indeed strong similarity between the AR-3 model we used and the AR-2 model suggested by John Ding. We also fit an AR-2 model to our data by solving the Yule-Walker equations. The resulting model is $Q(t) = 1.727\ Q(t-1) – 0.743\ Q(t-2) + 0.0751$. Again, the coefficients are very similar to those suggested by John Ding, which suggests that the discharge time series the models were fit to are very similar in terms of their autoregressive properties.

Model training and model performance

Wherever applicable, we fit our models by minimizing mean absolute error (MAE). MAE is a popular performance score in hydrological modeling if good overall performance is sought. If good reproduction of high flow is important, NSE is the better alternative. MAE-optimization was done for model-02 and model-05. For model-00 and model-01, no calibration was required. Models -03, -04 and -06 apply the parameters of model-02. The AR-coefficients of model-07 were found by solving the Yule-Walker equations, and the coefficients of model-08 were found by minimizing the mean squared error, the standard loss function in the related Neural Network software package.

Table 1 contains the performance of all models used in the study in terms of Conditional Entropy (Hc), Nash-Sutcliffe efficiency (NSE) and Mean Absolute Error (MAE). We can see that in general, the model ranking is similar for the different performance criteria: Model-01 is the best for all criteria; model-02, model-03 and model-06 perform identical (as to be expected). Differences occur with respect to the worst performing model: According to Hc, it is model-00, for NSE and MAE it is model-04. Further, we can see that overall performance of the single-bucket model-02 and the two-bucket model-05 is not very good in terms of NSE, which indicates that the chosen model structure does not well reflect the complex hydrological behavior of the catchment (snow processes, occasional overland flow, seasonal patterns of evapotranspiration). However, the two-bucket model-05 outperforms the simpler single-bucket model-02 in terms of their calibration objective MAE , which is in accordance with expectations.

The bucket models were chosen for their simplicity, and clearly their performance leaves plenty of room for improvement. But as they just serve as a demonstration cases for the bit-by-bit method, we are convinced they are nevertheless useful and serve their purpose.

Table 1

ID	type	Hc	NSE	MAE
Model-00	Mean	3.46	0	4.19
Model-01	Perfect	0	1	0
Model-02	1 bucket	2.89	0.08	4.78
Model-03	02+dt 1 min	2.89	0.08	4.78
Model-04	02+integer	3.23	-1.8	7.99
Model-05	2 bucket	2.85	-0.11	4.54
Model-06	02+iterative	2.89	0.08	4.78
Model-07	AR-3	0.66	0.99	0.18
Model-08	ANN	3.37	0.12	3.50

2. Catchments as a quadratic reservoir

The linear storage–discharge equation, Q = S/K (Figure 1a) can be extended to a quadratic one below:

$$Q = (CS)^2 \tag{3}$$

In the absence of precipitation, P(t) = 0 for Δt >>1 d, the recession hydrograph is linearized below:

$$\frac{-1}{\sqrt{Q_t}} = \frac{-1}{\sqrt{Q_0}} - C(t - t_0) \tag{4}$$

This has been called a negative inverse square root (NISR)–transformed recession flow model (Pelletier and Andreássian, 2020, and SC3 therein for my comment). This is in contrast to the universal logarithmic transformed one,

$$logQ_t = logQ_0 - \left(\frac{1}{K}\right)(t - t_0) \tag{5}$$

both having a single scale parameter K or C.

We agree that extending the range of reservoir candidates from linear to higher-order relations adds flexibility to models and potentially allows better calibration to a given catchment. However, the main purpose of the manuscript is about introducing the bit-by-bit method. In this context, the reason behind applying a broad range of model types is to demonstrate the general applicability of the method, and the performance of the models themselves is not a central element of the study. We therefore prefer to keep the range of models as is (please also see our reply to comment 1).

3. Training on transformed streamflow space

As a consequence of data linearization described above, some prior transformation of the observed streamflow time series data may help reducing the model computational complexity (i.e. number of computing steps) as opposed to improving model performance (i.e. a consequence of applying hydrologic law, formulas, and equations) (Lines 167-171).

As the case maybe, this can be the log or the NISR transformation of both a single reservoir (Figure 1a) and a two-parallel-reservoirs (Figure 1b) model, linear (Model-02 and 05) or quadratic as in Equation (3) above.

We agree that the suggestions by John Ding can help to better fit the models to the given data. But for the reasons already given in the replies to the previous two comments, we prefer to keep the models as they are.

Yours sincerely,

Uwe Ehret, on behalf of all co-authors