Review for “A likelihood framework for deterministic hydrological models and the importance of non-stationary autocorrelation”

Summary

The study introduces a new approach for treating non-stationary auto-correlation in residuals of hydrological models. A likelihood framework is used to evaluate different complexities in auto-correlation, and this is applied for different temporal resolution of data in two catchments. The improved representation of auto-correlation is shown to solve previously identified problems with joint calibration of hydrological and residual error models, and the authors demonstrate how auto-correlation is important for hydrological signatures such as flashiness index.

I found this study to be well written and insightful. The framework for comparing different complexities of the error models, and the experiments implemented via this framework, worked well. I believe this paper will be of clear interest for the HESS audience, and recommend it be published following the following minor comments being addressed.

Main comments

1. Title and focus: I found the title to be somewhat misleading, in that it refers to deterministic hydrological predictions, whereas this paper is largely about improving probabilistic predictions. I think the authors are under-selling their contribution by not including the term “probabilistic” or “uncertainty” in the title.
2. While results of this study are promising, the findings are somewhat limited by considering only 2 catchments. I feel the authors should recognize this in the discussion or conclusions, and emphasize that further research required on large datasets to determine which error model to use for different catchments, temporal resolutions, etc.
3. The authors state on page 3, lines 23-27 that “An explicit marginal distribution of streamflow (Krzysztofowicz, 2002) facilitates scientific communication and discussion, since hydrologists are generally more familiar with streamflow than with Box-Cox transformation parameters or distributions of the innovations of residuals.” I see where the authors are coming from here, but predictive performance is likely as important as scientific communication. I would like to see comparison with existing error models listed as an area of further research in the discussion or conclusions.
4. In equation (7), the conditional probabilities when $Q(t_{i-1})$ =0 do not take into account information about $\mathbf{Q}$ from any previous time steps. In theory, these probabilities should take into account the last uncensored observation, and the fact that all observations in between are censored. See (Zeger and Brookmeyer, 1986) and (Hannachi, 2014) for details. This limitation should be recognised in equation (7).

Minor comments

Pg 6, line 20: "Note that, if the distributional assumptions about D_Q hold at all points in time, $\eta(t_i)$ are a sample from a standard normal distribution, except for the lower tail, which can be lighter due to the truncation at zero at each individual time step. "

I agree there would be a lower proportion of $\eta(t_i)$ near the lower tail. But the pdf should have an area of 1, so I would have thought there would be an increase in samples of $\eta(t_i)$ close to the truncation point (which varies in time) to make up for this.

Pg 10, line 6: What does $\tau_{\min} = 0$ correspond to? Why was this chosen?

Table 2: It would be interesting to see the proportion of zero flows for these catchments, since zero flows are represented in the likelihood function.

Figure 3,4: It would also be useful to have arrows for "better" performance for all metrics (not just reliability and precision"

Figures 3-6: It would be useful to label all panels of figures to make it easier to refer to parts of figures.

Pg 18, line 5-6: " $\hat{I}_{\mathrm{F,det}}$ is often similar to I_F for E2 (Tables B1 and B2), indicating that the large part of the flashiness of the model output is due to the hydrological model response and only a small part is due to the stochastic variability added through the error model.

Isn't this true for all models?

Pg 20, line 29-31: "Figure 5 compares the predicted hydrographs of E1, E2 and E3a. In this case, allowing for different characteristic correlation times during precipitation events and dry periods (E3a) prevents the problematic behaviour encountered when making the constant correlation assumption."

It is not clear which problems the authors are referring to, and how E3a fixes them. Please provide more details.

Figure 5: It is difficult to see difference between models. Zooming in on a smaller time period may help emphasize differences.

Minor edits (typos, etc)

Pg 5, line 17: Move "e.g." to start of brackets

Pg 10, line 28: Change "10'000" to "10,000"

Pg 11, line 15: Change "empirical error models, as the ones" to "empirical error models, such as the ones "

Pg 13, line 18: Change "Streamflow data is a courtesy" to "Streamflow data is courtesy"

Pg 15, line 12-13: Change "prior believe" to "prior belief"

Pg 19, line 16: Change "the reliability measure shows a stable performance in," to "the reliability measure shows stable performance"

Pg 29, line 11: Change "catchments storage" to "catchment's storage"

References

HANNACHI, A. 2014. Intermittency, autoregression and censoring: a first-order AR model for daily precipitation. *Meteorological Applications,* 21**,** 384-397.

ZEGER, S. L. & BROOKMEYER, R. 1986. Regression Analsis with Censored Autocorrelated Data. *Journal of the American Statistical Association,* 81**,** 722-729.