Review: Estimation of predictive hydrological uncertainty using quantile regression: examples from the national flood forecasting system (England and Wales)
Authors: Weerts, Winsemius, Verkade

The paper demonstrates and argues for the efficacy of quantile regression (QR) as an approach for quantifying uncertainties associated with deterministic predictions. Essentially, the QR approach is a way of estimating error quantiles that have been derived from past verification of hydrological forecasts and applying them to real-time forecasts. Overall, the paper gives a clear and useful presentation of the approach and it may be beneficial to other hydrologic forecasting groups who are limited to deterministic forecast approaches -- thus I recommend publication, subject to the authors addressing a number of questions and suggestions detailed in specific comments below. The primary suggestions are the following:

- The authors choose to apply a transformation (NQT) to the forecast and observational data before quantifying error quantiles in standard normal space, then transform back to flow space for the final representation of flows and their uncertainty. One of the strengths of QR is that it does not require assumptions about the normality of variables or residuals as more traditional linear regression techniques – hence one would think the transformation step is unnecessary, and in fact exactly the kind of effort that is avoided by using QR. It appears that the main benefit is that it allows the authors to derive linear error quantiles, yet achieve non-linear uncertainty plumes in flow space. The alternative would be to apply QR in flow space using non-linear conditional flow estimation functions (polynomials or splines, or locally linear piecewise functions – all of which are possible in the quantreg package they describe). Although I do not suggest that the authors should alter their approach, I would like to see a savvy discussion of this issue.
- The authors claim that the approach is very simple, though it may be unfamiliar to many readers. It would be useful to see a comparison of the QR approach with a more familiar standard linear regression approach in which predictive quantiles are derived from the standard error. This could still take place in std. normal space if desired, but a logical question is whether anything is gained from the application of QR relative to more common practice.
- Though QR's 'father' works in the area of economics (ie, Roger Koenker), QR has been applied operationally in other areas, in particular as an approach for draping errors on deterministic wind power forecasting, with one of the leading authors on the techniques being John Bjørnar Bremnes. It would strengthen the paper to have a more comprehensive discussion of other applications of QR such as these – thus would be practitioners would be advised of other sources of insight.
- The thresholding of the flow domain for QR application to exclude low flows at which the quantile functions could cross is a reasonable approach, as long as the main area for concern is not the low flow area or the area of the transition. I'd like the authors to discuss both the rationale for choosing the threshold, and also other alternatives to the selection of an arbitrary threshold if known. Other researchers who have applied QR (Tom Hopson of UCAR/RAL, James Brown of NOAA OHD) have developed other approaches for addressing this issue, though this reviewer does not know if they are published. Remarkably, the abrupt quantile slope changes in std.

normal space appear relatively smooth when transformed back to flow space, so in practice this decision may not make a substantial difference in the application.

- Lastly, one would hope that the error draping technique would result in statistically reliable uncertainty estimates (in which obs frequencies match forecast quantiles) – in fact, this is one of the main criteria of success in estimating uncertainty bound! The authors mostly attribute this to sampling error or regulation effects, but I think they should examine this problem in more detail. If standard linear regression is applied as a benchmark, I wonder if it would also show similar reliability problems? To evaluate the sampling issue, the authors could reverse their training and testing sample periods (hoping for the opposite bias in reliability).

Additional comments of a more specific or stylistic scope follow:

Abstract: "extremely simple" is a subjective/relative qualifier, and I'm not sure "robust" has been demonstrated by the paper (given the reliability issues). To some readers, QR may not be simple – perhaps best to describe it as computationally non-demanding (or something better worded), since a more involved manner of estimating predictive uncertainty may be generating ensembles.

p 5549, ln 1: "…Agency **should** shift…"

p 5547, ln 21: I don't agree that deterministic forecast imply certainty or accuracy. I think users construct a mental, qualitative model of their uncertainty, and forecasters have their own model as well. The authors should rephrase this, perhaps to say that deterministic forecasts do not provide an explicit uncertainty model to go with the forecasts.

p 5550, ln 15: The authors may wish to note that the QR approach as applied is a form of the "second option" described in the previous paragraph.

p 5550, ln 22: this would be a good place to describe other applications of QR, eg in wind power forecasting and elsewhere.

p 5551, ln 25: as described earlier, this choice (to transform to std. normal space) is not required for application of QR. It's not clear that the authors understand this distinction (an important one), hence they may want to revisit the theory. There may be other reasons to take this step, as describe above, but the authors should revamp this section, the better to justify the NQT step and elaborate on other alternatives (that would achieve flow-dependent quantiles, e.g., spline based or other non-linear ways of applying QR) for the reader to be aware of. Also, in contrast to 'stationary', the meaning of the word "ergodic" in this context is not clear – why is an assumption of ergodicity required?

p 5553, ln 24: Transforming forecasts and observations to a Gaussian domain does not guarantee that their residuals will also be Gaussian, although the authors make this

assumption – they should probably check it, though as noted before this condition is not required for applying QR.

p 5554, ln 5 – the authors here recognize this point, calling again into question why the NQT step has been taken.

p 5558, ln 9 – either here or elsewhere, the authors should comment on the adequacy of the training period for estimation the error relationships.

p 5558, ln 21 – "evident": please elaborate a little further on the effect of NQT, rather than leaving it with "evident".

p 5560: Fig 4-7 do not present the shaded uncertainty areas with a dark enough shading -- particularly for the outer quartiles. These need replotting.

p 5560, ln 2-8: The language of this paragraph indicates that the authors will take a qualitative approach to evaluating the results: "reasonably accurate", "more or less provides", etc. I tend to agree with these assessments. Compared to a deterministic forecast alone, the QR-based uncertainty bounds will convey more information about the uncertainty of the predictions, and the bounds are reflective of past errors. Nonetheless, I would ask the authors to be more rigorous in this part of the assessment. Can you assign a confidence to the ability of the predictive uncertainty functions to describe observed error? ie, construct rank histograms (deciles, quintiles, etc.) and evaluate their significance with a goodness of fit test, perhaps a Chi Squared statistic. Does the technique, in fact, provide predictive uncertainty ranges that validate a reliability hypothesis? It's not much work to answer this question. What may be harder is explaining why they don't, if that's the case. The authors allude to this kind of assessment on p 5560 (describing QQ plots, not shown).

p 5561, ln 4: "maybe somewhat too wide" is pretty casual language for a technical paper. How about "wider than expected for reliable distributions". The idea that this is due to differences in test & train periods could be easily tested by reversing them to see if the error distributions end up being narrower. All in all, I would like the authors to tackle this question of whether the resulting distributions are reliable or not with a bit more vigor. Perhaps focusing on the worst case (the highly managed case) could lead to quick insights into the problem, and hopefully result in suggestions for would-be practitioners faced with similar performance problems.

Final comment – although the writing is generally clear, there are a number of awkward diction choices and phrasings that could be eliminated by having a native English speaker/technical writer proofread the document for such errors before resubmitting it.