Hydrol. Earth Syst. Sci. Discuss., 11, C5580–C5589, 2014
www.hydrol-earth-syst-sci-discuss.net/11/C5580/2014/

Hydrology and Earth System Sciences
Discussions

Interactive comment on "Performance and robustness of probabilistic river forecasts computed with quantile regression based on multiple independent variables in the North Central USA" *by* F. Hoss and P. S. Fischbeck

Anonymous Referee #1

Received and published: 9 December 2014

Review of: "Performance and robustness of probabilistic river forecasts computed with quantile regression based on multiple independent variables in the North Central USA" Authors: Hoss & Fischbeck Recommendation: Accept subject to major revision

General Comments

The paper is a fairly exhaustive study of the application of QR to post-process and provide exceedence probabilities for various thresholds from otherwise deterministic forecasts. It clearly should be published, as it is a comprehensive study of a useful

method. I strongly recommend, however, that the authors revise the paper in several ways – without which I find it both inaccurate and lacking in some regards:

1) The authors are apparently unaware of the first presentation of QR, which pertained to an American river, and predated the Weerts et al (2011) paper by several years. Wood et al (2009) is a citable conference presentation and is available online through the Amer. Met. Soc. (note the paper currently cites one conf. presentation). It is notable because the presentation also presents the rationale for using river rise as a predictor in QR, and demonstrates the application to operational river forecasts. This paper claims repeatedly to be the first application in an American context, which it is not given the earlier work, and also claims to introduce the concept of the additional predictors. I recommend that the paper recognize both Wood et al (2009) and Weerts et al (2011) as introducing the QR method for streamflow post-processing (until another earlier ref. can be found!), recognize the Wood et al inclusion of predictors such as river rise, and remove the framing of Weerts et al as the 'original' method versus this papers 'new additions'. The authors make a substantial contribution in their detailed examination of river rise together with the new predictor – trailing error – and the use of QR to estimate exc. probs.

2) Though the authors highlight several interesting characteristics about the varying performance of predictor combinations, they currently offer little physical explanation for outcomes such as the forecast itself being a poor predictor in some cases, or multiple predictors faring worse at high thresholds. Physical reasoning would help dispel the possibility of simple overtraining, or perhaps mis-aligned training given the sample. I think the paper needs a stronger physical or at least statistical discussion to provide insight into the cause of such findings.

3) The paper argues in several places that the exc. prob. forecasts are somehow 'more useful' for decisions than confidence intervals on forecasts (a widely used output). This arguably depends on the user. The position is taken to bolster the author's claim of an 'advance', but it's unnecessary – both are useful, and the author's can simply note that

they have taken a different tack than in earlier uses.

4) The results section is somewhat long, and I think the paper could still be effective if the figures and tables were trimmed somewhat – but I leave this to the author to decide.

Specific Comments

282,2 – awkward first sentence: 'further develops [QR]'? or just 'applies', or perhaps 'further develops an application of QR'. I don't think QR itself is being further developed. also, suggest rephrasing "…to predict flood stage exceedence probabilities based on post-processing single-value flood stage forecasts."

282,5 – it was not the first, actually – see comment below for 285,6.

282,8 – suggest avoiding references in the abstract. Also, this statement is not correct – see comment on 285,6 below – the first implementation did use additional variables. The Weerts implementation was far more comprehensive, leading to an article, and also added the nice feature of flow normalization as an innovation to the approach.

282,17 – I suggest adding one more sentence to the abstract to state the value of the approach – ie, that it helps quantify forecast uncertainty for the outputs of a deterministic forecasting process, which is currently common practice in many national flood forecasting services.

283,3 – "quantify 'forecast' uncertainty"

283,13 – perhaps mention that the HEFS system described in Demarge also includes a method for post-processing total uncertainty.

283, 23 – 'serves as' – perhaps, but who knows? It's never been verified. Better to say 'may serve as'

284,3 – this is true in the eastern US – in the west, ensemble forecasts go out as long as 2 years. This figure could be trimmed to reduce paper length.

284,10 – NWS also has a technique called HMOS which is applicable to post-processing single value forecasts. HEFS also includes the EnsPost module, which post-processes total forecast uncertainty, and these both should be mentioned.

284,13 – again, 'further developed'? What does this mean exactly? perhaps just use 'applied' or clarify what aspect of R. Koenker's method is being 'further' developed. 285,12 – this view is a bit narrow; certainly many users are concerned with low flow thresholds as well, and in any case, confidence bounds on forecasts are directly relatable to risk of threshold crossing (high or low). Suggest

285,6 – QR for streamflow post-processing was introduced both by Wood et al (2009) and Weerts et al (2011). The former reference described what was likely the first application of QR to streamflow in the 'US American context', and possibly anywhere:

Wood, AW, M Wiley and B Nijssen, 2009, Use of quantile regression for calibration of hydrologic forecasts, 23rd Conf. on Hydrology, Phoenix, AZ, Amer. Meteor. Soc., 11.3 [available online at: http://ams.confex.com/ams/89annual/wrfredirect.cgi?id=10049]

Wood et al. described using QR to provide confidence limits for deterministic forecasts of the Lewis River in Washington State (e.g., Figure 1). The work emphasized the need for determining the QR error models as a function of the rise rate of the river as well as lead time (e.g., Figure 2), and then demonstrated the application. An earlier version of this presentation had been given by the same author at the 2008 HEPEX workshop in Delft, NL on Hydrological Ensemble Post-processing Methods, and this was acknowledged as the inspiration for Weerts et al (2011). It is likely that the work was not submitted to a journal because the authors worked in the private sector, where publication is typically less encouraged than conference presentation. Incidentally, the Wood et al streamflow QR work had in turn been inspired by the application of QR for calibrating temperature forecasts, as described by Hopson and Hacker (2008), as well as by applications in the wind forecasting industry.

Hopson, TM and JP Hacker, 2008, Combined approaches for ensemble post-processing,19th Conference on Probability and Statistics, New Orleans, LA, Amer. Meteor. Soc., 3.1 [available online at: http://ams.confex.com/ams/88Annual/wrfredirect.cgi?id=7501]

285, 23 – given the previous comment, this statement is incorrect and should be removed. The paper should recognize the earlier work and related ideas therein.

285,25 – this paragraph summarizes results, and seems out of place. Better to state that QR is conditioned on several factors in the study, and say what those are and why they are considered, than to the tell the outcome (here) of doing so.

286,10 – having established earlier that Weerts, and I suggest also Wood, introduced QR, it is not necessary to return to it repeatedly in the paper (eg 286, 15, 19 etc). Overall, I think the paper should de-emphasize the verbiage about 'additions' and 'further development' in contrast to an 'original method', especially since the rise-conditioned error approach actually was the first method introduced at a national scientific meeting. Instead, just emphasize what has been done, as it is good work, and the paper can stand on its efforts alone, without requiring the label of being 'new' or 'first'.

286,20 – I would just write here that the work combines elements of Weerts et al and Wood et al, and also does [b] and [c] (though take out the word 'more' – not needed, and perhaps debatable).

287, 21 – Here and throughout the rest of the paper, please reframe the presentation of Weerts et al (2011) as the 'original' implementation focusing only on the forecast as predictor, with an 'addition' being the use of other predictors or conditioning factors – as this addition is quite clearly described in the earlier Wood et al (2009). Both works should be recognized, as they are citable/viewable by the field, and assigning the term 'original' to the second reference is misleading. Your paper, as noted above, makes other valuable contributions in addition to exploring these ideas, and does not need to work so hard to distinguish itself. Perhaps call the Weerts version the 'forecast-based' or 'W11' approach, versus multiple predictor approaches, or any other labeling that

seems better.

289, 16 – again, I object to the characterization that exceedence prob. is 'more useful' for decisionmaking than confidence intervals. This really depends on the decision, and I have actually more often, in forecast office settings, heard users ask about confidence than risk of exceedence, though again, it depends on the use. There is no reason to argue this point in the paper. Both uses of the uncertainty are valuable, and I support the authors focusing on the risk of exceedence predictand, and stating that is 'also important' or even 'more useful for some users'. But the assertion that it is somehow categorically more useful is needlessly provincial, and can be removed.

Section 2.3 – as per earlier comments, suggest retitling this 'Inclusion of additional independent variables'. Please reference Wood et al (2009) as described earlier in recognizing the value of including rise rate and lead time as variables (this can be done obliquely, eg, "...as noted earlier, rise rate and lead time have been previously shown to be informative independent variables. We assess these factors as well as ..." etc. Also, please give more detail (ie, an update of equations 1 &/or 2) to show mathematically how the additional predictors were included.

293,16 – again, this is a needlessly narrow view, as what aspects of the forecast PDF are required entirely depends on the decision model to which the forecasts may be input. For hydropower optimization, for instance, the full PDF of the forecast would be desired, and is 'decision relevant' input. I think all but the last sentence of this paragraph should be removed, and the remaining sentence added to the preceding paragraph.

294,7 – this is clearly quite a lot of work (which would lessen its operational applicability), and a somewhat brute force approach to determining the best functions. Can the authors suggest any more expedient alternatives to the more or less 'trial and error' search for the best predictor combinations? Is there an analogue to 'stepwise regression' here, perhaps? In stepwise fit approaches to MLR, there is typically a stopping criterion that discourages the addition of new predictor variables – would any similar measure be useable here? Later comments in Section 3.2.2 suggest that there may be overfitting with larger predictor sets.

Overall, the results presentation is quite long, although the figures do support a range of conclusions of the paper, and most are of interest. I suggest the authors look for chances to remove a few of the figures and/or tables, which may be overkill, especially if they jointly support a conclusions, but I leave it up to them.

298, 7 – Please provide greater insight into why the inclusion of the forecast itself might degrade the performance of the post-processed forecasts. Elsewhere, findings that, eg, more variables lead to worse performance at higher stages, also bear more physical explanation. What aspect of the variables could make them damaging to the high threshold models?

Also, it's not entirely clear that figures 11-12 support the assertion that "Without a transformation into the normal domain, the forecast does not provide a lot of information for the QR model" – giving metrics of these relationships (rˆ2 for instance) may help show that in fact, they are significantly different with normalization. There is a lot of scatter in both figs 11 & 12.

300,7 – I may be misinterpreting the figures (19,20), but it appears that length of record does matter (longer is better) somewhat more than the authors suggest, and more for the lower thresholds, which is surprising – I'd think those were better represented in any length record than high extremes, given the typical skew of flow distributions. Please comment or provide a more nuanced assessment.

301,22 – as per earlier comments, Wood et al (2009) preceded this study in the American context, and further argued for and demonstrated the use of the 'additional' variables of both river rise and lead time. Please adjust text appropriately.

301,26 – Instead: "This work confirms a prior finding that including additional predic-

tors such as rise rates in the past 24 and 48 h benefits the resolution of the resulting probabilistic forecasts. In the first comprehensive assessment of various combinations of. . ., we found that . . . "

302,10 –It's inaccurate to call these 'the new independent variables' as rise rate was used earlier.

302,14 – it's not clear why these variables do not lend themselves to transformation – please be more specific and speculate as to why you are finding this. Are they distributed such that the transformation reduces their correlation with the predictand? It's an interesting result, but not intuitive why it should be.

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., 11, 11281, 2014.

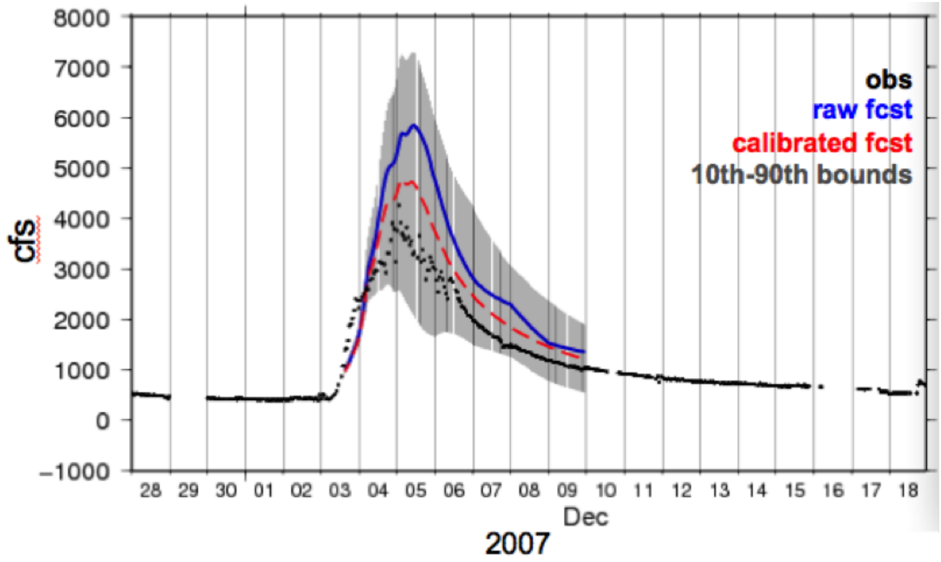
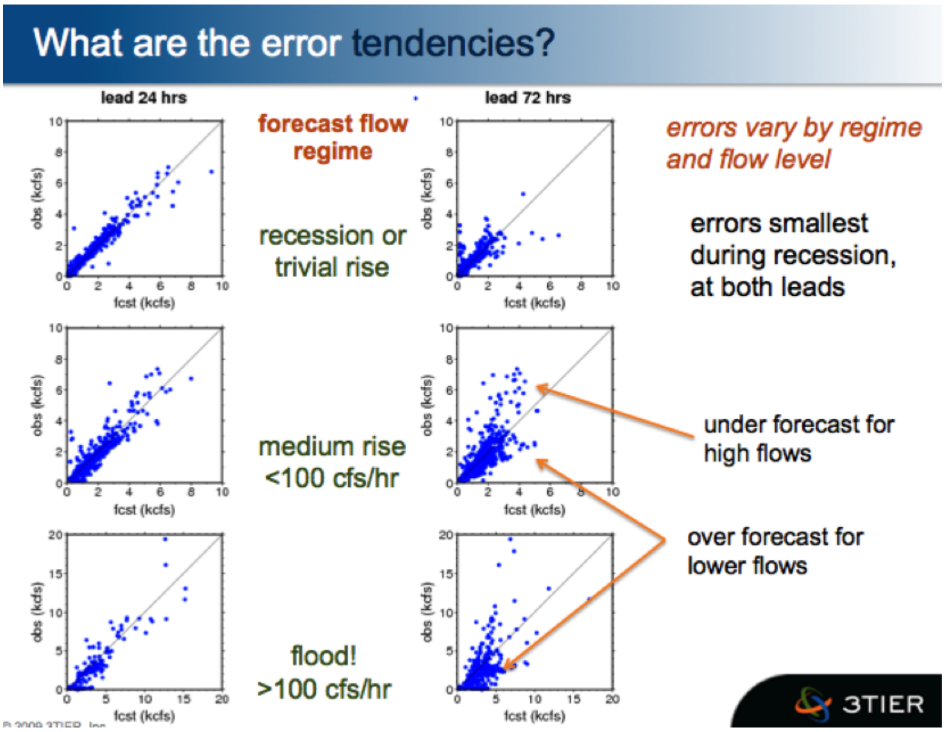


Fig. 1. Figure 1, QR streamflow application example from Wood et al (2009)


What are the error tendencies?

forecast flow regime

recession or trivial rise

medium rise <100 cfs/hr

flood! >100 cfs/hr

errors vary by regime and flow level

errors smallest during recession, at both leads

under forecast for high flows

over forecast for lower flows

3TIER

© 2009 3TIER, Inc.

Fig. 2. Figure 2, Description of streamflow rise and lead time conditioning factor from Wood et al (2009)