Response to Reviewer 2

General Response: We thank Reviewer 1 and Reviewer 2 for their detailed reviews, and for sharing their constructive and insightful comments. Before our point-by-point response to Reviewer 2, we note several major changes in the manuscript organization and content based on comments from both reviewers. We have addressed all review comments in our responses as well as in a substantial revision to the manuscript. The resulting manuscript is clearer and much improved. We note that for us to properly address and understand the reviews, it was necessary to actively revise the manuscript. Although the HESS process does not allow us to share our revised manuscript at this stage of the review process, we provide excerpts throughout the response to help illustrate the changes, and will provide the revised manuscript when invited.

In terms of manuscript organization, we have consolidated the methodology to be one section. Section 3 now combines previous sections 2, 3, and 4.2. This improves the clarity of the method and reduces redundancies. In terms of content, we have carefully re-evaluated what use cases, figures, and tables to include in the paper. For example, we have removed one of the locations, Bad River, SD, from the Results entirely (Section 4). In doing so, we have culled our figures and tables such that all of the figures and tables are in the main manuscript. We currently have the same number of figures as the last draft (12 Figures), but we no longer have any figures or tables in the Supplemental document. This streamlining has helped to simplify the paper and to improve its clarity. One notable example of this is that we now focus the methodology and results on cluster maxs, whereas in the previous version we included methods/results from cluster maxs and cluster means. By focusing on the cluster max analysis, we clarify how the (dis)agreement between the events in the observed WT and modeled XWT can be used to quantify “hits/misses”. In summary, we thank the reviewers for bringing these points to our attention. As a result of their careful reviews, we have added numerous clarifications to the text and figures in an effort to improve overall understanding and interpretation. We hope that the editor and reviewers find these changes helpful and we look forward to sharing the revised manuscript.

Below, we provide a point-by-point response, where we address each of Reviewer 2’s concerns.

Reviewer 2.
Cedric David (Referee) cedric.david@jpl.nasa.gov
Received and published: 24 October 2020

General comments

The following is a review of manuscript hess-2020-323 entitled “A Wavelet-Based Approach to Streamflow Event Identification and Modeled Timing Error Evaluation” by E. Towler and J. L. McCreight being considered by Hydrology and Earth System Sciences.

This manuscripts describes a methodology for evaluating the timing of simulated river discharge hydrographs when compared with in situ observations. The approach first uses wavelet

transforms (WTs) to expand observed one-dimensional hydrographs (discharge vs time) into two-dimensional WTs (power vs timescale and power vs time) as a means for event detection. From those events detected in the observations, the methodology then uses cross wavelet transforms (XWTs) to evaluate the difference in timing and duration of events (at multiple power levels) between observations and simulations. The new approach is specifically designed to compare subsequent versions of a given river model and can be used both for evaluation and for diagnosis. The paper uses simulations from subsequent versions of NOAA's National Water Model and observations from the USGS at four selected locations.

I really enjoyed reading this paper and I learnt a lot from it. I agree that the evaluation of hydrograph timing is an important aspect of river model calibration/validation, and one that is often overlooked. I am guilty of that myself! This subject matter is timely given the ongoing explosion of continental scale river models such as the NWM or similar global applications. In my opinion, the authors make a strong case for the value of their work given the complexity of hydrograph shapes and describe clear guidelines for the implementation of their methodology, while also acknowledging the multiple ways in which it could be adapted. This research ought to be of interest to the readership of HESS. I elected to write this review without reading any of the other community comments so that my opinion could be relatively unbiased. My recommendation is to return the manuscript to the authors for minor revisions. My comments are outlined below, in decreasing order of importance. I hope that the authors will find some value in my suggestions. Thank you for the opportunity to review this work.

Thank you very much for this helpful global assessment. Your specific suggestions are extremely valuable and have helped strengthen the presentation and content. Thank you for your careful review. To refer to the different responses, we have numbered each Reviewer response.

Specific comments

First, I really want to highlight that I think the authors did a commendable effort in the clarity of their explanations. I specifically enjoyed the inclusion of a "Conceptual Overview" (Section 2) and the use of the simple prescribed "+5 hours" example (Section 4.2). I think the general descriptive approach used really serves the manuscript very well and makes it accessible to many readers, including those (such as myself) who are not familiar with wavelet transform-based event detection methods. I'd like to make two suggestions that may further help in this manner. First, I understand from Supplemental Table 1 that the units of "power" are mˆ6/sˆ2, hence they are squared units compared to discharge. It may be valuable to some readers to specify these units in the manuscript and also to perhaps suggest a hydrologic meaning to this quantity. For example, if one was to take the square root of the power value, would this in any way represent the amplitude of the peaks in the figures? If so, this may be a valuable explanation to add, which could also be graphically illustrated in Figure 1a. A rough estimate from Figure 1 suggests that the maximum power is 60,000 mˆ6/sˆ2 leading to a square root of approximately 245 mˆ3/s which is strikingly close to the amplitude of the hydrograph.

R.2.1. Before we respond to this specific inquiry, we want to point out that as noted in our General Response, we have restructured the methodology (now Section 3) by merging draft sections 2, 3, and 4.2. This allows our description of the methodology to progress logically, reduce redundancies, and allows us to clarify points of the method, such as the one the reviewer brings up here.
The point brought up here is worthy of discussion generally: additional interpretation of the wavelet transforms and associated quantities (WT, WT power, XWT, XWT power, phase, and significance). Reviewer 1 also made a similar suggestion and this is a point echoed later in this review as well. We have substantially bolstered the description of the methodology and its details in multiple places. With regards to the interpretation of the wavelet power raised here, we have inserted the following text in the methodology section to improve the paper:

> *We make several additional notes on the wavelet power and its representation in the figures. The units of the wavelet power are those of the timeseries variance (m6/s2 for streamflow) and it is natural to want to cast the power in a physical light or relate it to the timeseries variance. Indeed, the power is often normalized by the timeseries variance when presented graphically. However, it must be noted that the wavelet convolved with the timeseries frames the resulting power in terms of itself at a given scale. Wavelet power is a (normalized) measure of how well the wavelet and the timeseries match at a given time and scale. The power can only be compared to other values of power resulting from a similarly constructed WT. There are various transforms that can be applied to aid graphical interpretation of the power (log, variance scaling), but the utility of these often depends on the nature of the individual timeseries analyzed. For simplicity, we plot the raw bias-rectified wavelet power in this paper.*

Second, I could have used some further hand-holding in understanding Figure 2c. I guess I expected the entirety of the color scale to correspond to “+5 hours” (i.e. all red). I don’t really understand why there appears to be a 10-hour minimum to accurately catching errors and why this makes sense because “the [time]scale must be at least double the error”. It would be valuable to expand on this concept around Lines 322-325.

R2.2. We fully agree that this should be better explained. In fact, the “phase aliasing” that can be seen at timescales shorter than 10 hrs (2x the synthetic timing error) was something we hoped the reviewers would ask about and that we could expand our discussion in that context. We reworked and expanded the description of the cross-wavelet phase difference with discussion in context of the synthetic example as follows in our revised manuscript:

> *3.2.2. Step 2b. Calculate the cross-wavelet timing errors*
> *For complex wavelets, such as the Morlet used in this paper, the individual WTs include an imaginary component of the convolution. Together, the real and imaginary parts of the convolution describe the phase of each timeseries with respect to the wavelet. The cross wavelet transform combines the WTs in conjugate, allowing the calculation of a phase difference or angle (radians) which can be computed as:*

$$\phi_n^{XY}(s) = tan^{-1}\left[\frac{\Im(\langle s^{-1}W_n^{XY}(s)\rangle)}{\Re(\langle s^{-1}W_n^{XY}(s)\rangle)}\right],$$

Where I is the imaginary and R is the real component of W_n^XY (s). The arrows in Figure 3b indicate the phase difference for our example case, which are used to calculate the timing errors. Note that these are calculated at all points in the wavelet domain.
The distance around the phase circle at each timescale is the Fourier period (hours). We convert the phase angle into the timing errors (hours) as in Liu et al. (2011):

$$\Delta t_n^{XY}(s) = \phi_n^{XY}(s) * T/2\pi$$

*where T is the equivalent Fourier period of the wavelet. Note that the maximum timing error which can be represented at each timescale is half the Fourier period because the phase angle is in the interval (-pi, pi). In other words, only timescales greater than 2E can accurately represent a timing error E. Because of the phase discontinuity at ±pi, true phase angles outside this range alias to angles inside this range. (For example, the phase angles 1.05 * pi and -.95 * pi are both represented by the latter). When the wavelet transforms are approximately antiphase, the computed phase differences and timing errors produce bimodal distributions due to noise in the data. Figure 3c shows phase aliasing in the negative timing errors at timescales less than 10 hours, double the 5 hour synthetic timing error we introduced. The bimodality of the phase and timing are also seen at the 10hr timescale when the timing errors abruptly change sign (or phase by 2pi). We note the convention used is that the XWT produces timing errors that are interpreted as "modeled minus observed", i.e., positive values mean the model occurs after the observed. Positive 5 hour timing errors in Figure 3c describe that the model is "late" compared to the observations as seen in the hydrographs in the top panel (a).*

I wonder if there could be a good graphical way to explain how the timing error is computed in Figure 2c. The equations provided in Section 3.2.2 are not exactly straightforward, and I assume that making a graphic from those would be challenging, but it seems to be a key component of the study and some readers might benefit from such addition.

R2.3. We see the reviewer's point, and note that in the previous draft, we had aspects describing and demonstrating methodology spread across several sections (Section 2, Section 3, and Section 4.2); whereas in this new version, we have consolidated those sections into a single section (now Section 3). We hope that by having the equations and the illustrative example (i.e., Onion Creek isolated peak) all in one place, we will help to better guide the reader. We concur that an illustration would be challenging and have elected to more comprehensively describe the complex algebra involved in the calculation of the phase angle and the timing error, as per the response to the previous comment (R2.2). We have also

bolstered the description of the XWT power and how it is used to diagnose if timing errors are "hits"; this is further described in our final response, R2.6.

The manuscript tends to rely a bit heavily on supplemental figures throughout the text, which makes the reader jump from document to document. I suggest that the authors go through their figures carefully and evaluate whether some of supplemental figures could be combined with main manuscript figures.
R2.4 We agree, and as mentioned, we have carefully reviewed the content of the manuscript to address this. As a result, we have culled our figures and tables, so that now all of the figures and tables are now in the manuscript. This required removing one of the use cases: the Bad River, SD, as well as additional streamlining. We currently have the same number of figures as the last draft (12 Figures), but we no longer have any figures or tables in the Supplemental.

For the benefit of the reviewer, here is a summary table mapping the old figures to the new figures, as well as a description of the figures now included:

NEW	OLD	Description
Figure 1	Supplemental Figure 1	Flow chart of methodology
Figure 2	Figure 1	Isolated Peak Onion Creek, Step 1 (WT)
Figure 3	Figure 2	Isolated Peak Onion Creek, Step 2 (XWT), synthetic +5 hour offset
Figure 4	Figure 3	3-month Pemigewasset River, Step 1 (WT)
Figure 5	Figure 4	3-month Pemigewasset River, Step 2 (XWT), NWM v1.2
Figure 6	Figure 5	3-month Pemigewasset River, Cluster timing errors by characteristic scale
Figure 7	Supplemental Figure 4	1-year Taylor River, Step 1 (WT)
Figure 8	NA (new figure)	1-year Taylor River, Step 2 (XWT), NWM v1.2
Figure 9	Figure 8	1-year Taylor River, zoom in of spring runoff obs and NWM v1.2 time series
Figure 10	Figure 11	5-year Onion Creek, NWM version comparison, cluster max timing errors
Figure 11	Figure 12	5-year Pemigewasset River, NWM version comparison, cluster max timing errors
Figure 12	Supplemental Figure 7	5-year Taylor River, peak streamflows by characteristic scale

Further, here are the figures that have been removed from the manuscript, along with the reason:

Cut figures (OLD)	Description	Reason cut
Supplmental Figure 2	Isolated Peak Onion Creek, Step 2 (XWT), NWM v1.2	Need to reduce use cases and was repetative; got cut when we consolidated methods section in re-structuring.
Supplmental Fig 3	5-year Pemigewasset River, NWM v1.2, compared cluster max with cluster mean timing errors	Need to reduce use cases; showing mean vs max is unnecessary (now just mention in Discussion)
Figure 9	Spring runoff for Taylor River, Cluster timing errors by characteristic scale	Need to reduce use cases, already see clusters once for Pemi (Fig 6)
Figure 10	1-year Taylor River, Cluster timing errors by characteristic scale	Need to reduce use cases, already see clusters once for Pemi (Fig 6); Table 4 now summarizes this.
Sup Fig 5	5-year Taylor River, NWM version comparison, cluster max timing errors.	Need to reduce figures, already show this for Onion and Pemi, Table 7 describes this sufficiently.
Sup Fig 6	1-year Bad River, NWM v1.2	Need to reduce use cases
Figure 6	2-months Bad River, Step 1 (WT)	Need to reduce use cases
Figure 7	2-months Bad River, Step 2 (XWT), NWM v1.2	Need to reduce use cases

The authors make a clear argument that some of the traditional error metrics (e.g. RMSE, NSE) implicitly include errors in timing but don't shed much light on them. Yet, the strength of these metrics is also the simplicity of their computation. In an effort to increase the broad acceptance and use of timing metrics that are less subjective than peak-over-threshold, it might be helpful to

for the authors to attempt a recommendation for the simplest possible form of their methodology. I understand that different powers are related to independent peaks and that there is value in looking at them all. I wonder if using powers computed as the square of discharge values from traditional occurrence probability thresholds could help reconcile (connect?) the WT approach and the threshold approach. I am not suggesting more analysis here, more so a paragraph in the discussion (Section 6, around Lines 478-487) where the authors might further expand on the simplest way for others to apply their approach.

R2.5. We agree that presenting the simplest form is appealing. Some of this is related to our previous response, where we described how we evaluated what was included in the manuscript. For example, we removed the Bad River, SD, case to help focus our results on less use cases. Second, we note that we now limit our results to only looking at cluster maximums, whereas in our previous version we looked at both cluster max and cluster mean. We now note this in the Discussion and Conclusions:

> *For instance, we focus on the cluster max, but one could also examine the cluster mean. Also, instead of finding the event of maximum power in each cluster (i.e., for a given timescale), it would be possible to identify the event with maximum power in "islands of significance", i.e., significant areas contiguous in time across timescales. However, this ignores that multiple frequencies can be important at once and defining the islands when connected is problematic. If one suspected non-stationarity in the characteristic timescales over the timeseries, then a different approach such as a moving average in timescale could be employed.*

I'm not sure I fully understand the True/False (dark grey/light grey) legend in figures 5, 9, and 10. Figure 10 seems to suggest that lighter colors are not statistically significant but it's not clear from the legend. Likewise, the "Avg XWT Signif" color bar in Figures 11 and 12 is a bit mysterious to me as I see no associated colors on the graph. Could the authors rework their legends in these figures?

R2.6. We note that we have eliminated figures 9 and 10 in an effort to simplify the analysis and presentation, as mentioned above. We have left the True/False legend in Figure 5 (now Figure 6), but have clarified how XWT significance is used to determine when timing errors are "hits" in the modeled timeseries. Although this was a minor point by Reviewer 2, we note that this was also picked up by Reviewer 1, and so to this end, we have also bolstered the description of the XWT power and how it is used to diagnose if timing errors are "hits", which we include below.

We agree that we need to enhance our description of the event (dis-)agreement of observations and simulations. We address this, but first we want to remind the reviewer of the context that, in our re-organization and streamlining, we simplified our methodology and results by focusing on cluster maxs. (In the previous version we presented results for both cluster max and cluster means approaches). The cluster max is a single point of maximum power per cluster, and so cluster maxs can be classified as either hits or misses. In the previous draft we calculated the cluster mean timing error and the corresponding percent of hits within the cluster. We used the % hits in that case as a confidence measure for the timing error of each cluster. We see now how this was not clear as the hit diagnostic was different for these approaches and hence our

decision to now focus the manuscript on cluster maxs. Focusing on cluster maxs simplifies our Step 2d, which was previously called “Quantify the confidence in the timing error”, which we now call “Quantify Percent Hits”. Percent hits now refers to a full-timeseries diagnostic of cluster maxs (instead of individual clusters). The edited section is included below:

3.2.4. Step 2d. Quantify Percent Hits

The premise of computing a timing error between the observed and modeled time series is that they share common events which can be meaningfully compared. In a two-way contingency analysis of events, a “hit” refers to when the modeled timeseries reproduces an observed event. When the modeled timeseries fails to reproduce an observed event, it is termed a “miss”. In the case of a miss, it does not make sense to include the timing error in the overall assessment. Because the timing errors are calculated from the XWT, we choose to diagnose hits and misses based on the significance of the XWT. Once cluster maxima are selected based on the characteristic timescales of the observed event spectrum and timing errors are obtained at these locations in the XWT, the significance of the XWT on the cluster maxima is used to decide if the model produced a hit or a miss for each point. For a single cluster max, such as shown in Figure 3c, the XWT significance is either True or False, the point is either a hit or a miss. Table 2 also displays the results of the timing error analysis for this synthetic example. We can see the prescribed 5 hour offset and that the cluster maximum was significant in the XWT. When calculating timing errors for a longer time series, a useful diagnostic is to calculate the percent hits over all the cluster maxima in a timescale. When summarizing timing errors statistics for a timescale, we drop misses from the calculation and the % hits indicates what portion of the timeseries was dropped (% misses = 100 - % hits). In our tables we provided timing error statistics this way as well as over all observed events to reveal the impact of dropping misses.

Because, in step 1, we constrain our analysis to observed events in the wavelet power spectrum, we do not consider either of the remaining categories in a 2-way analysis (false alarms and correct negatives). We note that a complete 2-way event analysis could be constructed in the wavelet domain based on the Venn diagram of the observed and modeled events without necessarily using the XWT.

The new Table 2 and Figure 3, referenced above, follow:

Table 2. Summary of timing error results for isolated peak and prescribed 5 hour offset from Onion Creek, TX, for cluster max.

Characteristic Timescale (hr)	Avg WT Power	Number of Clusters	Cluster Max		
			Timing Error (hr)	Time (hr)	Hit?
22	598,000	1	5	37	Yes

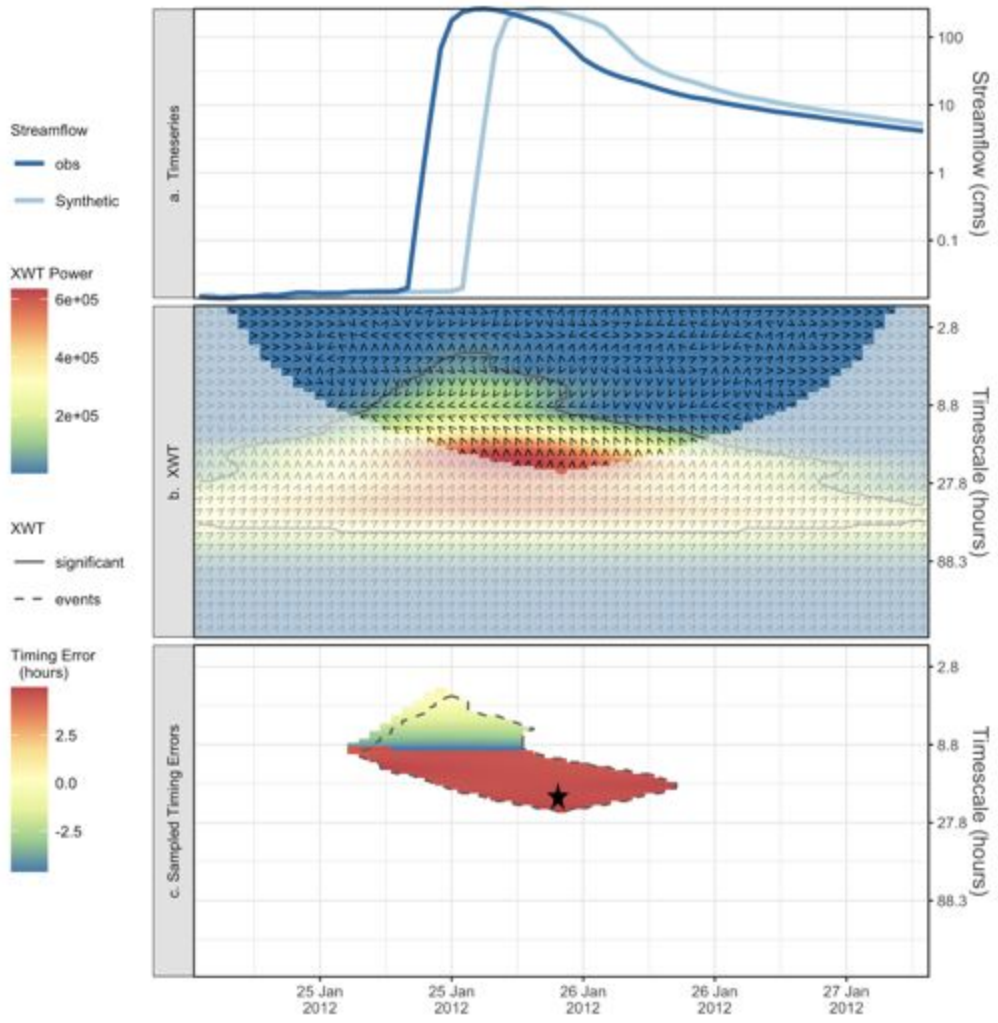

Figure 3. An isolated peak from Onion Creek, TX and a synthetic +5 hour offset: (a) observed and synthetic time series (note logged y-axis), (b) cross wavelet (XWT) power spectrum and phase angles (arrows), (c) sampled timing errors for observed events (dashed contour is XWT significant events) and star is cluster maximum.

Related to the decision to focus on cluster maxs and their diagnosis as hits or misses using the XWT, earlier in the manuscript (Step 2a) we have also edited the cross wavelet (XWT) figure panels to show this visually. This is seen in Figure 3 (previously Figure2, Onion Creek synthetic example) above and in its caption. Specifically, we have adjusted panel c, to show how the observed events (colors) don't exactly overlap with the XWT significant events (dashed contour). This is now explained in the methodology:

> *Similar to Step 1b of the WT, we can also calculate areas of significance for the XWT power as shown by the black contour in Figure 3b. For the XWT, significance is calculated with respect to the theoretical background wavelet spectra of each timeseries (Torrence and Compo, 1998). We define XWT events as points of significant XWT power outside the COI. XWT events indicate significant joint variability between the observed*

and modeled timeseries. In the next section, we employ XWT events as a basis for identifying hits and misses on observed events for which the timing errors are calculated. Described further in step 2d, timing errors are valid only for hits. Figure 3c shows the intersection of the observed events (colors) and the XWT events (dashed contour). This is a region of hits. Note that the early part of the observed events, particularly at shorter timescales, is not in the XWT events. This is because of the timing offset in the modeled timeseries, which misses the early part of the observed event.

For Figures 11 and 12 (now 10 and 11) we have renamed the color scale scale “Percent Hits” and clarified the caption to include: “... outline shading shows percent (%) hits in the cross wavelet transform (XWT).” This is shown for Figure 11, below:

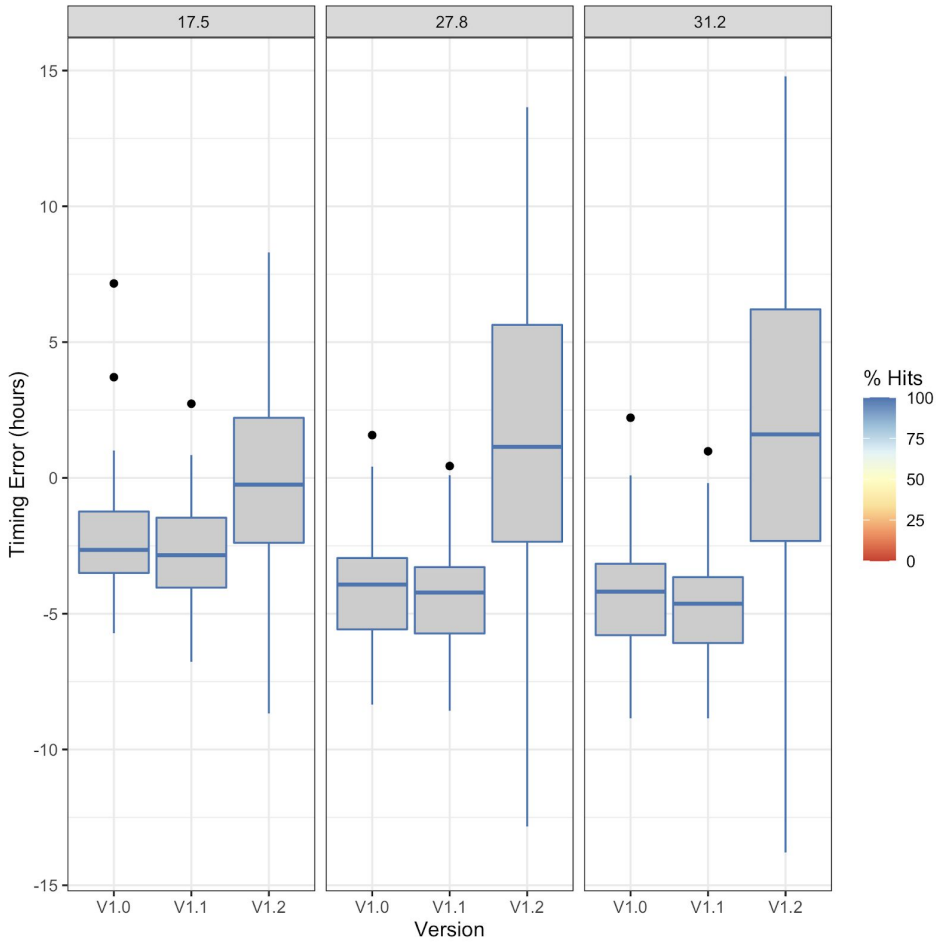

Figure 11. Five year run from Pemigewasset River, NH: Comparing cluster max timing error distributions for top three characteristic timescales (see panel title) across NWM versions; outline shading shows percent (%) hits in the cross wavelet transform (XWT).

In our original tables, timing errors were calculated over all observed events (i.e., both hits and misses) and we reported the percent hits (previously called “Avg % Significant in XWT”). We will calculate timing errors over only the hits for each timescale and add this as an additional column

to the tables. We may drop the original timing error statistic over all observed events (hits and misses), but for now we plan to compare the two.

Technical corrections

Line 296-297: missing “Land” in the acronym for NLDAS.

This has been fixed.