Review of “A WRF-Hydro-based retrospective simulation of water resources for US integrated water availability assessment” by Rafieeinasab et al.

**General Comments**:

This paper documents the WRF-Hydro IWAA modeling application in terms of its configuration and performance. It is generally well-written, and provides a detailed and comprehensive documentation of the WRF-Hydro IWAA. However, it has several main issues that need to be addressed in relation to the three HESS Review Criteria (i. Scientific significance, ii. Scientific quality, iii. Presentation quality). I recommend “Major Revision”.

* **Scientific Significance is “Fair”** *(Does the manuscript represent a substantial contribution to scientific progress within the scope of HESS (substantial new concepts, ideas, methods, or data)?”* The paper is within the scope of HESS, but the scientific contribution needs to be much better contextualized. The paper states: “*This paper focuses on providing an in-depth account of the WRF-Hydro modeling effort within the IWAAs, specifically delving into the details of the WRF-Hydro model configuration and evaluating its performance”*. As stated, the paper does comprehensively document the single experiment, the WRF-Hydro IWAAs. In that respect, it covers the “what” and the “how”. However, as is, it doesn’t provide enough context to help the reader understand if this is a substantial scientific advance – i.e., it’s missing the “why”. Three suggestions to help with this contextualization:
  + 1. The paper needs to more clearly state the “why”, this includes the research gap (the need) and how what you are putting forth fills that (the advance). Why does the IWAA need hydro model simulations (rather than just observations?). Why did you need a new configuration of WRF-Hydro for the IWAAs? (i.e., why didn’t you just use an existing application, like the operational NWM? Why use a national-scale model over catchment-based models?). Why did the IWAA need to be forced with a new met product like a bias-adjusted CONUS404? (i.e., why not just use the AORC?). Once the need is more clearly identified (see specific comments for Introduction), you need to show how this new application (i.e., CONUS404BA + IWAAs WRF-Hydro) is better (or an advance) in terms of design and/or performance, or at least better suited to your IWAAs application than other options might have been (see next point):
  + 2. The results need to be contextualized with another benchmark or model baseline. Currently, the only benchmark is the default vs calibrated, calibration vs validation periods, which follow the expected relative performance. There is one mention of the streamflow performance being lower than in Cosgrove et al. (2024); is there a way to obtain this data for comparison? Or to compare with NWMv2.1 retro performance (Towler et al. 2023)? Or a climatological benchmark like in Knoben et al 2020? If this can’t be done quantitatively, then more qualitative comparison and justification for why for the performance is (or is not) acceptable for the purpose of the IWAA is needed (this discussion could go in the Conclusions).
  + 3. The methods need to be contextualized with another benchmark or model baseline. Throughout the manuscript, there’s a lot of details of the WRF-Hydro model configuration, and often they are in comparison to other NWMs (v2.1, v3.0). I suggest that the methodological parts in the manuscript should focus more on what’s different or novel about this application – and why these changes are well-suited to the IWAAs. I suggest maybe a table or two to highlight these differences against some baseline, like the operational NWM, and perhaps some of the in-depth details can go in a Supplemental.

Knoben, W. J. M., Freer, J. E., Peel, M. C., Fowler, K. J. A., and Woods, R. A.: A brief analysis of conceptual model structure uncertainty using 36 models and 559 catchments, Water Resourc. Res., 56, e2019WR025975, <https://doi.org/10.1029/2019WR025975>, 2020.

Towler, E., Foks, S. S., Staub, L. E., Dickinson, J. E., Dugger, A. L., Essaid, H. I., Gochis, D., Hodson, T. O., Viger, R. J., and Zhang, Y.: Daily streamflow performance benchmark defined by the standard statistical suite (v1.0) for the National Water Model Retrospective (v2.1) at benchmark streamflow locations for the conterminous United States (ver 3.0, March 2023), US Geological Survey data release [data set], <https://doi.org/10.5066/P9QT1KV7>, 2023.

* **Scientific quality is “good”: “***Are the scientific approach and applied methods valid? Are the results discussed in an appropriate and balanced way (consideration of related work, including appropriate references)?*” The approach and methods are valid, but the results should be discussed more in terms of related work/references (see previous three points about contextualization), and what conclusions can be drawn from the results. Related to this, the Conclusions is more of a summary, and could benefit from revision and additional discussion and what conclusions can be drawn (see Scientific Significance point 2 and specific comments).
* **Presentation quality is “fair”:** “*Are the scientific results and conclusions presented in a clear, concise, and well-structured way (number and quality of figures/tables, appropriate use of English language)?*” The paper is generally well-written, but the content and overall structure need to be tightened and made more cohesive. The majority of the paper is focused on streamflow, and the other water budget component results aren’t introduced or well-integrated in the paper (see Specific Comments on these sections). As previously mentioned, there are a lot of dense, methodological details included; these might be better summarized in a table and details in a Supplemental (see Scientific Significance point 3). Similarly, the performance metric results could be tightened up, and I suggest where some figures could be combined or go in a Supplemental. In addition, if possible, a new figure or comparison with a benchmark or model baseline would be a good addition (see Scientific Significance point 2).

**Specific comments**:

Abstract: lines 1-5. This is missing the research gap or the “why” of this article. For instance, it is not clear why hydrological model simulations (in this case from WRF-hydro) are need to characterize water availability for the IWAAs (why not just use observations?). Making this case to readers up front is going to be critical to showing the value of your paper.

Line 12. “*specific emphasis on temporal accuracy issues”* is not clear.

Abstract – this is written almost like two abstracts, where the first paragraph is about streamflow, and the second paragraph is about snow, ET, and soil moisture. These should be integrated. Would help to introduce earlier what is done in the paper (like around line 5, say in this paper we evaluate simulations of streamflow, snow, et, soil moisture, etc). Final line of the abstract is very vague, can these results be tied to the streamflow results more definitively? What are the conclusions drawn from the study?

Introduction: Similar to comments on abstract, the research gap is not clear, and the reader is missing the “why” we need hydrologic simulations for the IWAAs, or what the main contribution of this work. Why do we need a new IWAA application, rather than use the operational NWM for instance? Some of the pieces of this are in the Introduction, but I suggest reworking the Introduction to make the value clear to the reader.

Line 27: After this sentence: “*The inaugural cycle of this national water availability assessment has two primary objectives: firstly, to provide a status assessment of water availability for the period 2010 to 2020 on a national scale, and secondly to conduct a historical trend analysis exploring multi-decadal changes over time for the period 1980 to 2020.”* Prior to this usgs would do trend assessment on observations, right? It seems after that sentence you should help the reader understand why hydro modeling is needed for this (ie. observations are sparse/many ungauged basins). Depending on how you reorganize the paper, this might be where you introduce the unified national-scale framework of the NWM, which provides the continuous spatial/temp coverage needed for this type of assessment*.*

Line 34: Need to define acronym and describe to the reader what CONUS404 is, and why it is relevant to IWAA application.

Line 36. Would it help to give more background on the IWAA (what are the other modeling applications contributing?).

Line 39: “*This paper focuses on providing an in-depth account of the WRF-Hydro modeling effort within the IWAAs, specifically delving into the details of the WRF-Hydro model configuration and evaluating its performance.”* I agree that this is what the paper focuses on, and while this is worth documenting for the IWAA community, it is harder to identify the new contribution within this for HESS’ international readership.

Line 57: Increasing the forecast points seems like something that should be highlighted sooner (or at least the need for this).

Line 60: You mention it aligns with the hydrography of NWM3.0, are there other differences with the operational model? For the model, what (if anything) makes the IWAAs different than the operational model – why did you make those decisions? How do those decisions better support the purpose of the IWAAs estimates?

Line 61: This paragraph doesn’t seem provide information relevant to the background or motivation for the study. Suggest moving to the model description unless this is relevant to the research gap or why of the paper.

Line 70: This is abrupt. Seems like it would be useful to have a preceding paragraph, especially on the need to evaluate national-scale hydrologic fluxes. Many studies look at streamflow, but do many others look at snow, soil moisture, and ET? What additional insight does evaluating those provide? Can you tie it back to the why of the paper (like the IWAAs?).

Model Description:

This is well written, but it is very dense, and it might help if you can better differentiate between what is unique about the IWAAs application versus another benchmark application (say the operational NWM) and why you made those decisions (see previous comment on Introduction). If there are quite a few differences, it might be easier to see them in a table(?). As part of this table, you could also include comparisons of how the model is Calibrated/Regionalized with a benchmark, from that section later in the paper. Then you could focus the main text on describing the key differences and why you made them for the IWAA, and put detailed descriptions in a Supplemental.

Figure 1. Maybe bold or highlight somehow the variables that you evaluate in this paper in this figure. You could say that at line 78 when you introduce it.

Forcing:

Line 113-119. I wonder if some of this background information might be better in the Introduction, perhaps after the need for a unified continental scale hydro model is introduced, you could move to the forcings.

Line 127-129. Maybe some of this justification should go in the Intro as well, and it would be good if there was a reference to qualify how the CONUS404 performs.

Line 130 - This is a very detailed description of the bias-correction. I wasn’t clear if this bias-correction (CONUS404BA) is also publicly available – suggest explicitly mentioning that this is available as a data release reference here to clarify for the reader. Also, when I first read this, based on what was stated as the purpose of the paper in the Intro, I thought this was not necessary and should go in a Supplemental, but then after reading the entire paper I wasn’t sure if the CONUS404BA was one of the more important parts. This goes back to more clearly describing the scientific advance, and setting up the reader in the Intro to know if this is one of the main contributions of the paper.

Figure 4 is not well explained in the text (either provide the explanation in the text or caption and/or maybe put in Supplemental).

Line 183. Figure 5 shows the calibration basins in terms of the RFCs – it would be helpful to introduce/define RFCs here, and why you use them in your analysis. How do the RFCs fit into the IWAAs? Also, suggest editing Figure 5 caption (maybe remove: *as it will be used frequently in the results and discussions)*.

Table 1. Is WRF-Hydro always calibrated with these 17 model parameters? Is this the same/similar table as in Cosgrove 2024? If so, maybe it can go in the Supplemental (or just pull out parameters in a table highlighting what the differences are between IWAA WRF-Hydro and NWM). Is DDS always used? Again, it would be good to have a WRF-Hydro benchmark/comparison for context, and a table or tables could be a good way to show this.

Figure 6 & Regionalization. I wonder if it’s more important to document the differences between the regionalization approach used in IWAA versus in NWMv2.1 and NWMv3.0 (like in a table), than to provide the workflow in Figure 6 (which could maybe go in Supplemental?).

Section 4. Model Calibration: This is a very detailed and dense section. To repeat a previous Main comment, I suggest that the parts of this that go in the manuscript should focus more on what’s different or novel about this application – and why this is well-suited to the IWAAs. I suggest maybe a table or two to highlight these differences against some baseline, like the operational NWM, and perhaps some of the details that are already published can go in a Supplemental.

5. Results and Discussion

5.1. Add “Streamflow” to Evaluation of Calibration Basins

Figure 8. Can you put the calibration and validation on the same plot but have one of the lines (say validation) as dashed? Otherwise, it is hard to see the differences between the model runs for the different periods. This is expected relative performance, so perhaps some of the metrics could go in a Supplemental.

5.2. To be consistent with the 5.1, maybe make both “Evaluation” or both “Verification”

Figure 9. It is hard to see all the points… can you make all the points the same size to try to improve the visibility? %Bias and correlation coefficient are also quite small (since they are in the same row). Maybe put one in the supplemental so the other can be better seen like KGE?

Line 325 – this paragraph is abrupt/out of place as is, but I think you can tie it in if you put this in reference to Figure 9, which has Ref and Non-Ref gages. Do you see better/worse performance depending on if it is Ref or Non-Ref? Hard to tell now since there are so many points in Figure 9 (see previous point).

Line 344. It would be useful to contextualize the performance difference if you can quantify the comparison with Cosgrove 2024, which would be a great baseline. Can you obtain this data? See General Comments for other suggestions on benchmarking. Either way, this is a useful discussion point even if it stays as qualitative, but seems buried here (this might fit better in a discussion and conclusions section).

Line 339: I think you mean “Figure 10” here, not “9”. Also, suggest only showing 1 or 2 of these metrics unless the story is different or relevant to IWAAs, and putting the others in the Supplemental.

Figure 11. Caption is wrong. Interesting figure, but need to indicate why it was included (what is it telling us, how can it help us to interpret the other results?).

5.3. Snow Analysis

As stated previously, this is a bit abrupt, since there is no background/motivation in the Introduction for why this is useful for IWAAs, or diagnosing streamflow errors, etc. Or will IWAA be doing trend analysis on these variables too? It almost feels like we are starting a new paper at this point in the manuscript.

Line 401. Here you mention potential reasons for the snow biases, but if the paper is focused on streamflow (which the majority of it was), can you say anything about how these biases relate to the streamflow results we’ve seen?

Is there a baseline to compare the snow results to, even if it is qualitative? Seems like you could reference Cosgrove et al. 2024 here, but maybe others too.

5.4 ET Analysis

Again, it would be great if this could be better integrated into the paper. As with snow: Are there other efforts/studies that have evaluated ET using WRF-Hydro (to compare baseline or benchmark)? Or how do these analyses help to explain the streamflow results we’ve seen?

5.5. Soil Moisture Analysis

Similar to snow and ET: Are there other efforts/studies that have evaluated soil moisture using WRF-Hydro (to compare baseline or benchmark)? Or how do these analyses help to explain the streamflow results we’ve seen?

Conclusions (suggest removing potential model enhancements from section name)

Line 492. No need to redefine, just say WRF-Hydro IWAA.

The “Conclusions…” section is mostly a summary and needs to be revisited once the paper is revised. Suggest adding more discussion and conclusions based on results (see General Comments on this). A few thoughts (you don’t need to answer all of these, they are just to get you thinking):

* How will places where there is poor model performance be handled in the IWAA? Will trend analyses still be done?
* Can you discuss how do these results compare to other related work and studies? If this can’t be done quantitatively, then more qualitative comparison and justification for why for the performance is acceptable for the purpose of the IWAA is needed.
* What parts of this study (methods or results) might you recommend that others in the hydro community adopt or compare with? What are the “lessons learned” for using CONUS404 and a national-scale hydro model for the IWAA? What caveats would you provide in terms of using these results for water availability and trend assessment?
* Can you use some of the other water budget results to better understand the streamflow results? This could help to flesh out more of the potential model enhancements.