Nat. Hazards Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/nhess-2019-82-RC1, 2019

Natural Hazards and Earth System Sciences
Open Access
Discussions

Interactive comment on "A Comprehensive Evaluation of the National Water Model (NWM) – Height Above Nearest Drainage (HAND) Flood Mapping Methodology" *by* J. Michael Johnson et al.

David G. Tarboton (Referee)

dtarb@usu.edu

Received and published: 18 May 2019

General comments

This paper evaluates the National Water Model (NWM) Height above Nearest Drainage (HAND) flood mapping methodology by comparing modeled inundation area with mapped inundation derived from satellite products for 30 flood events. This is a worthwhile comparison, as, given the importance of the NWM and its potential use for flood

warning, emergency response, and planning, it is important to know the uncertainty associated with its predictions. The comparisons presented in this paper are helpful in this regard, though there are limitations to the analysis, that merit caution in not over-interpreting the results. While the analysis is helpful, I feel that its claims of being "comprehensive" in the title and "first detailed" in the abstract are not justified.

The first caution is that the statistic (AFI, equation 1) used to quantify accuracy is generous and can introduce bias and arbitrariness. It scores the match between model and observed as the fraction of area modeled and observed to be flooded, plus the fraction of area modeled and observed to not be flooded. This weights area not flooded heavily in the calculation. Bias and arbitrariness are introduced as the non-flooded area depends on the extent of the convex hull used for the area analyzed. If the geometry is such that a large part of non-flooded area is included in the convex hull, then AFI will be inflated. I am concerned that the conclusion that the method can be used "quite confidently" is biased because of this, and the ranking of how well the different floods are modeled is uncertain due to the arbitrariness of the convex hull. See specific comments below for expanded discussion of this point.

Additional cautions are that errors can be due to errors in the NWM modeled discharge or errors in the HAND methodology. The paper mentions these and presents a table where the dominant cause is stated for some cases with poor AFI. However, the analysis does not separate the effect of the different errors, except anecdotally for some of the cases. Where there are discharge differences (such as illustrated in Fig 5 and 6) it should be possible to model inundation using the observed discharge and thus interpret which part of the error is due to discharge error, and which part of the error is due to HAND methodology error. I think that such separation is important in assessing limitations of the HAND methodology.

It would also be helpful, if in pointing out error (e.g. for small streams or flat areas) for the authors to offer ideas or suggestions towards correcting the problems. The detailed comments below offer a few suggestions.

Given these general concerns, I find myself taking with caution the conclusion that NWM-HAND can be used "quite confidently" (Page 5, line 23, or P5 L23) or "general level of agreement" (P9 L28). Don't get me wrong. I do not want to dismiss this study. I believe that the HAND approach has considerable merit and that comparison studies such as this paper are important, but the analysis needs to be presented in a more complete, balanced and objective way to document where the approach is effective and to support the conclusions. I think that changes to address the arbitrariness problems of the statistic and to separate errors due to discharge versus HAND methodology, should be made before this manuscript is finally published. In the reviewer form, I characterized these as “major” revisions, but they may actually fall between major and minor and could, I think, be done fairly quickly.

Specific Comments

I do not think “Comprehensive” in the title is justified. 30 flood events are certainly a worthwhile study, but the study really only ended up reporting errors and did not dig into the causes in a way that merits use of the term comprehensive. E.g. the separation of problems due to discharge errors vs HAND errors was limited.

Error statistic. The authors acknowledge in part that the weighting of non-flood area in AFI is a limitation (discussion P9 L20-24) but did not do anything about it, claiming it is suitable for the purposes of this paper. I tend to disagree. For example, suppose that 10% of the area is observed to be flooded (Figure 2 suggests this to be about right), 90% of the area is not observed to be flooded. Say that the model predicts 10% of the area to be flooded, but mostly in the wrong location. Perhaps 1% of the area is observed and modeled to be flooded, with 9% of the area modeled to be flooded but not observed, and 9% of the area observed to be flooded but not modeled. This would seem like poor prediction, only getting 10% of the flooding in the right place. But with

these numbers 81% of the area that is not flooded is modeled to be not flooded, so the AFI score is 82% (81% non-flooded match + 1% flooded match) which would be interpreted as good. Another part of this problem is that the results are sensitive to the area of the convex hull, because that dictates the amount of non-flooded area in any evaluation of AFI. If the geometry is such that a large part of non-flooded area is included in the convex hull, then AFI will be inflated. For example, if given the numbers above the convex hull had been smaller so that 20% of the area was flooded, 20% modeled as flooded and 2% observed and modeled as flooded. This could be exactly the same comparison as above but removing half of the overall area from the non-flooded part of the convex hull. With these numbers the AFI score would be 64% (62% non-flooded match + 2% flooded match) falling towards the lower end of moderate, rather than good. I think there is a need to consider the potential distortion of the results and their interpretations due to this effect. This could be done by reporting the matching and non-matching areas for each case, as well as the total areas. Further, it may be better to use a statistic that does not consider non flooded area such as the fit statistic from Sangwan and Merwade (2015).

$$Fit(\%) = 100 \times \frac{F_{pred} \cap F_{obs}}{F_{pred} \cup F_{obs}}$$

Separate errors due to discharge errors from errors due to HAND methodology. Where observed discharge is available it should be used to generate a flood inundation map and differences between this and observed flooding examined as they are due to the HAND methodology alone. The paper states as its goal (P4 L17) “exploration of how errors in the methodology and its individual components may compound and propagate. . .”. Nowhere does the paper separate out individual error components. Errors reported always combine effects of discharge error and HAND methodology error. This is an important shortcoming that should be addressed. Within the 30 study cases, which are due to poor discharge predictions, not a problem of the HAND methodology per se, but of the discharge forecasts. For example, in Fig 5, and the text (P6 L9) the paper indicates that NWM simulated discharge was 50 m3/s, while the gage recorded 80 m3/s. This is a 60% difference and one should not expect a flood inundation mapping model to do well given such differences. An inundation map produced with the observed discharge would examine this. The paper could correct this by evaluating NWM discharges where there are streamflow gages. Section 4.2 suggests that some of this was done (P7 L20 "all NWIS stations within each flooded region ..."), but where are the results? The paper only presents a few discharge examples.

I do not think that Figure 5A and section 4.1.1 make a compelling case that raster resolution systematically suppresses stage in lower order reaches. Figure 4 is a bit of an oversimplification. The authors acknowledge this. While it is OK to make the point, it is not really consistent with Fig 5A. The observed inundation depth is stated to be 3.96 m (It would be good to state where this came from). Placing a stage of 3.96 m on cross section A-A' the river is notably over the banks and about 200 m wide. The cross section is thus not that poorly generalized by a 10 m DEM and a narrow channel such as suggested in Fig 4A, 5 m deep, seems inconsistent with the cross section for the observed stage and section A-A'. I would look to some other problem for this case. Discharge is part of it. But also Manning's n, slope, and general representation of the hydraulics involved (uniform flow assumption) may be problematic and warrant investigation. An additional point with respect to low order reaches, is that, being smaller, with smaller contributing areas, they likely produce less damaging floods.

Overprediction in areas of low-relief. Fig 5B. It would be helpful to diagnose what has gone wrong here. Was the large area mapped as inundated part of the NHD catchment used to compute the rating curve? A stage of 9.75 m is huge, and evidently an overestimate. Some analysis of why would be helpful rather than just generally saying this can be a problem. I think that overlaying NHD catchments used may be helpful, to see the areas used in calculating SRCs and whether the flood is extending

across catchments. Problems with roughness (Manning's n), slope, and the synthetic rating curve are all potential causes. For Fig 5C, this is a case where high resolution NHD streams may offer an improvement.

Data representation (section 4.1.3). When "NHD" is stated (P2 L27) I think that it is important to specifically indicate that it is the medium resolution NHDPlus dataset (I think that is what was used). There are high resolution NHD products becoming available that may improve matters, though there is work to be done to manage the scale of catchments used. NHDPlus medium resolution is nominally 1:100,000 scale from https://www.epa.gov/waterdata/get-nhdplus-national-hydrography-dataset-plus-data, while NHD High Resolution from https://www.usgs.gov/core-science-systems/ngp/national-hydrography is being used to create NHDPlus high resolution https://www.usgs.gov/core-science-systems/ngp/national-hydrography/nhdplus-high-resolution.

P9 L17. "Perhaps there is an opportunity to use NWM velocity forecasts ..." This seems speculative and is not followed up on. Consider either deleting it or building out the idea you had in mind.

The paragraph starting P9 L35 to the end of the paper is not conclusions drawn from results presented in the paper. As such, it should not be in the conclusions. It may be appropriate for discussion, but while certainly software considerations are important (OpenDAP, THREDDS, GUI etc), the paper has not said anything about them up to this point and tacking this discussion on at the end is a digression and distraction from the results of the study.

P2, L10-12. A citation describing how the methodology has been added to the NWC operational framework would be good.

P2, L13 (and P1, L15). Avoid claiming this paper is "the first extensive evaluation".

Some may say that there have been earlier evaluations (e.g. Zheng et al., 2018a; Zheng et al., 2018b; Godbout, 2018)

P2, L33-35. Please state how the relief between each cell and the nearest stream is calculated. If you are using the HAND layers from Liu et al. (2018) then this is computed using the TauDEM distance down function (Tesfa et al., 2011).

P3 L14-39. Section 2.2. A lot of the details here seem unnecessary. For example, the details about four forecast configurations, the products being made available to RFC's and on NOMAD do not seem relevant to the analysis reported. Further P4 L29 indicates "query the appropriate NWM output". What was the appropriate NWM output for this paper. Was it a forecast or one of the analysis and assimilation products?

P4 L4-13. Section 2.3. In contrast to section 2.2, the details about USFIMR are quite limited. It may be helpful to say a bit about the spatial resolution from the different satellite sensors and how these were rescaled for comparison with HAND inundation.

Technical corrections

P1 L17. I suggest delete "both" and "and". How is a quantitative comparison different from a detailed evaluation? I think this should read "These comparisons are made quantitatively through a detailed evaluation ..."

P2 L19. This sentence seems incomplete. What does stand for?

P3 L24. In the review draft I received the dot is a light pink, not red.

P3 L30. "most valuable models for prediction". This is a subjective statement. The most valuable models depend on purpose.

P4 L10. Incorrect parentheses.

P4 L28. For specificity please state the URL used to download the HAND products (presumably https://web.corral.tacc.utexas.edu/nfiedata/).

P5 L16. Indicate that units on Fpred, NFpred and Fobs are area units. Just saying Fpred is predicted flood may create the impression that this is a discharge, which it is not.

P5 L34. In what sense are the HAND products "recyclable". I suggest rephrase.

P6 L7. It is conventional to introduce figures in order. Here Figure 5 is introduced before Figure 4.

P6 L33. "where" last word.

P9 L11. In areas "where" ...

References

Godbout, L. D., (2018), "Error assessment for Height Above the Nearest Drainage Inundation Mapping," Master of Science in Engineering Thesis, The University of Texas at Austin, https://repositories.lib.utexas.edu/handle/2152/68235.

Liu, Y. Y., D. R. Maidment, D. G. Tarboton, X. Zheng and S. Wang, (2018), "A CyberGIS Integration and Computation Framework for High-Resolution Continental-Scale Flood Inundation Mapping," JAWRA Journal of the American Water Resources Association, 54(4): 770-784, 10.1111/1752-1688.12660.

Sangwan, N. and V. Merwade, (2015), "A Faster and Economical Approach to Floodplain Mapping Using Soil Information," JAWRA Journal of the American Water Resources Association, 51(5): 1286-1304, 10.1111/1752-1688.12306.

Tesfa, T. K., D. G. Tarboton, D. W. Watson, K. A. T. Schreuders, M. E. Baker and R. M. Wallace, (2011), "Extraction of hydrological proximity measures from DEMs using parallel processing," Environmental Modelling Software, 26(12): 1696-1709, 10.1016/j.envsoft.2011.07.018.

Zheng, X., D. R. Maidment, D. G. Tarboton, Y. Y. Liu and P. Passalacqua, (2018a), "GeoFlood: Large-Scale Flood Inundation Mapping Based on High-Resolution Terrain Analysis," Water Resources Research, 54: 10013-10033, doi:10.1029/2018WR023457.

Zheng, X., D. G. Tarboton, D. R. Maidment, Y. Y. Liu and P. Passalacqua, (2018b), "River Channel Geometry and Rating Curve Estimation Using Height above the Nearest Drainage," JAWRA Journal of the American Water Resources Association, 54(4): 785-806, doi:10.1111/1752-1688.12661.

Interactive comment on Nat. Hazards Earth Syst. Sci. Discuss., https://doi.org/10.5194/nhess-2019-82, 2019.