First, we would like to thank both reviewers for the thorough and incisive reviews, and the editor for allowing this paper to go through with minor revisions.

- Both reviewers raised **concerns with the title of the paper.** We recognize the challenge with executing a truly comprehensive evaluation as conceptualized by the reviewers. As such the title has been changed (per D. Blodgett's suggestion) to: **An Integrated Evaluation of the National Water Model (NWM) Height Above Nearest Drainage (HAND) Flood Mapping Methodology.**

- Moreover, per the suggestions of Dr. Tarboton, a more exhaustive look at HAND performance in gaged catchments has been added. In some ways this new analysis changed the way we discuss our outcomes and the format/flow of the paper. **Because of this, not all new additions / changes can be highlighted in this response. However, we have made every effort to specifically address each specific concern.**

- The largest structural change induced by reviewer comments was that all methods were moved into their own section allowing results to simply communicate what we found.

Attached to this submission please find the revised manuscript, with continuous page number and in-text figures and captions.

David Tarboton

Dr. Tarboton suggested changing the error statistic used to better eliminate arbitrary factors.

Thank you for this comment and pointing out the issues with the arbitrary convex hull and how the inclusion of matching dry regions may bias our results. To address this, we have adopted a new comparison that calculates accuracy, as well as overprediction and underprediction. These new values guide the remainder of the analysis. These are calculated by classifying the observed and simulated rasters cell-wise as WW, WD, DW, DD where W refers to wet and D refers to dry. The first character in the classification references the cell state in the observed flood map while the second refers to the state of the cell in the simulation.

Accuracy = WW / (WW + WD + DW) (fit index used Zheng 2018, and Sangwan, 2015)
Over = DW / (WW + WD + DW)
Under = WD / (WW + WD + DW)

These can be found in the revised manuscript as equations 3-5. This new metric did not change the overall conclusions of the tendency of NWM-HAND to under predict floodplain level inundation but did provide a more robust discussion and analysis that have improved the paper.

Dr. Tarboton requested that we report the matching and non-matching area between observed and modeled floods as well as total area.

The agreement of *total area* (Total Simulated Wet Cells / Total Observed Wet Cells) can be seen in new figures for the flood plain analysis (Fig. 2) and for the catchment level analysis (Fig. 4). The *matching and non-matching areas* are represented via the Accuracy (matching), Over (non-matching) and Under (non-matching) statistics and visualized in figure 3 as a stacked bar plot and reported in table 2. These images were added for clarity and to address this point.

Generate flood rasters for all NHD catchments that have a USGS gage. Compare these to the NWM ones to better separate out errors.

A new section (4.2 – 4.4) was added in the revised manuscript addressing this concern for the 54 available catchments that were completely contained in a USFIMR bounding box and had a recorded NWIS and NWM-reanalysis flow values. Overall, we found that the uncertainties in the NWM forecasts have a limited influence on the accuracy of the simulated flood extent and have documented these findings in the new sections.

Better articulate the issues with raster resolution. Make figure 5A more compelling. Problems with roughness (Manning's n), slope, and the synthetic rating curve are all potential causes.

Thank you for this comment and pointing out where our prior analysis was unclear. While testing the sensitivity of the SRC Manning Equations to roughness and wetted perimeter we discovered that our previous inclinations towards wetted perimeter being a driving factor were incorrect. In text (lines 376-380) we state:

>> Keeping slope (NHD attribute) and the cross-sectional area required to generate a stage of 3.8 m constant, we independently varied the roughness coefficient (N) and the hydraulic radius (via the wetted perimeter), solving for a Q of 80 m^3/s. In doing so we found that the SRC relationships are generally insensitive to changes in hydraulic radius (needed to be increased by a factor of 10), but were sensitive to changes in Manning's N. <<

In fact, the geometries that we tested could all generate proper discharge values when varying N between 0.001 and 0.2. Instead the most sensitive factor is that of roughness which is discussed at multiple points throughout the revised manuscript.

The reviewer requested a more thorough examination of what went wrong in figure 5B:

To really understand what was going on in this instance we needed a gaged reach to better dissect whether the previous large stage resulted from poor NWM prediction or a poor SRC curve. As such we changed our analysis to look at gaged reach upstream of our last example. This new reach can be seen in Figure 7A and is discussed in lines 398-404.

Explicitly state which NHD versions are used:

Thank you for this comment. The NHD version used is the medium resolution. This is now stated in line 112 - 114.

>>In 2017, HAND raster's and SRCs were generated for CONUS using the 10-meter NED and medium resolution NHD datasets on the ROGER supercomputing system at the University of Illinois Urbana Champaign (Liu et al., 2018; Zheng et al., 2017). <<

Remove comment on velocity or expand

Thank you for identifying the isolated nature of this comment. The idea of integrating the NWM velocity has been expanded on in lines 427-440. In text:

"A second possible alternative to refactoring is to make use of the NWM velocity and flow estimates to define cross sectional areas from the NWM forecast (equation 9). The intention would be to allow the physical model (NWM) and routing-routines (WRF-Hydro) to deal with issues of volume preservation. The resulting cross-sectional areas could be used as an Area-Stage rather than Q-Stage look up within the existing SRCs. This would work around some of the issues with roughness (outsourcing to the NWM) while capitalizing on the observed accuracies in the floodplain cross sections. Moreover, by controlling for the volume of water in the channel instead of the height, low lying areas will be less prone to exaggeration. Such a change would require (A) an understanding of how the NWM is handling hydraulics and thus velocity and (B) a test of how variations in velocity impact volume estimation. Both are interesting pursuits in their own right but out of scope for this paper."

Move discussion of software ect from collusion to discussion

Thank you for this suggestion. We have moved this section to the discussion and drastically reduced the detail. Please see lines 539-547.

Add a citation of how the methodology has been added to the NWC operational framework:

Unfortunately, we are unaware of any official citation for this. Instead we have cited the HydroShare resource for Hurricane Harvey (line 47-48).

NOAA National Water Center, E. Boghici, D. Arctur (2018). NOAA NWC - Harvey NWM-HAND Flood Extents, HydroShare, https://doi.org/10.4211/hs.fe85a680d0144e79b39e8c483dc1e5aa

Remove comments of 'first extensive evaluation' comparison

Thank you for the comment. We have noted the comment and removed all references to first extensive evaluation. Nevertheless, our analysis is novel in that it looks solely at the performance of the integrated NWM-HAND approach for a large sample of locations.

State how relief between cells is calculated:

We made use of the precomputed HAND rasters and have included the TauDEM distance down function reference you provided. This is now explained in line 113-114.

In the pre-computed HAND rasters, relief was calculated via the TauDEM distance down function (Tesfa et al., 2011).

Identify the "appropriate NWM output"

Thank you for identifying this sloppy sentence. The product used was the NWM version 1.2 reanalysis product which is now explicit stated in lines 164-166

The timestamp of each USFIMR satellite image was used to query the needed NWM v1.2 reanalysis values by COMID and generate an inundation map using the HAND methodology (section 2.1).

Add some info on USFIMR development and how rasters are aligned.

Thank you for the interest in the USFIMR products. We have pointed to the documentation for the shapefile development (lines 143-145)

"The USFIMR web portal provides more information on each flood, the specific sensor, as well as supplementary data including NED elevation and upstream NWIS hyperlinks (http://sdml.ua.edu/usfimr)."

and have described how rasters were created and aligned section 3.1.

Technical corrections:

Thank you for your detailed look at our paper, all suggested technical corrections have been accepted and incorporated in the revised manuscript including grammatical correction, subjective statements, the description of red/pink.

David Blodgett

Dr. Blodgett asked us to be more upfront with our choice of using a 2D fit statistic and the implications of not treating floods as a 4D event.

Thank you for this comment, it prompted some thoughts about what our analysis truly entails, the choice of methods, and why it was important.

The choice for a 2D statistic (XY) is driven by the limitations of remote sensing imagery that only offers a snapshot at a single time point (T). Analysis of time-space outside of this snapshot is doable for streamflow and simulated events but not for our observed 'truth' reference dataset. As for the Depth dimension, while there are some new methods for looking at flood water depths from RS imagery (see Cohen et al, 2018), doing so would have added a new source of uncertainty into an analysis where we were already tried to isolate and attribute errors from multiple sources.

In the revised manuscript we now explicitly state that we are implementing a 2D analysis of the flooded area coinciding with the timing of Aerial imagery and that it should not be read that we are analyzing peak flooded areas (lines 218-223).

The choice of a 2D fitness statistic (examining only the extent of flood, as opposed to depth and timing of flood propagation) is governed by the aerial imagery products available (which only captures the extent of the flood, at a singular point in time). By electing this form of evaluation, we only analyze the strengths of NWM-HAND simulations at the given time-step coinciding with the time of image capture (not necessarily peak flooding).

Bring discussion of limitations and realistic potential up from the discussion to the introduction/abstract.

Thanks for this important point suggestion. We have moved the ideas as suggested. Please see lines 48-51.

The current objective of the NWM-HAND approach is rapid flood prediction for the purposes of disaster warning and guidance. Model accuracy should therefore be viewed in this context and expectations should be tempered while recognizing the importance of having an operational, continental scale flood forecasting system.

Dr. Blodgett asked us to include more information regarding the NWM forcing data, parameterization, and routing.

Thank you for this suggestion. To be upfront, deciding on the level of detail to include with respect to NWM, HAND, SRC, and USFIMR background has proven to be a difficult task to making this paper both complete and concise. As such we have done two things. (1) made an explicit data section in the background. (2) We listed the attributes of the model you suggest but have avoided discussing any implications. Further, we point readers to a presentation talking about the model in detail (please see lines 120-128)

Dr. Blodgett noted that the introduction is lacking a general overview of the NWM's objectives which could/should be used to temper the expectations and focus the aims of an evaluation.

Thank you for pointing this out. We have added a caution of sorts to the introduction as well as a statement reflecting the current goals and objectives of the model (please see lines 48-51; see above)

Noting that the NWM-HAND system is not used for official forecasts and is to be considered for guidance only at this stage in its development. Given these kinds of caveats, the evaluation presented in this manuscript is of great value as it demonstrates that the NWM-HAND system is producing flood inundation products that would be generally useful for the intended purpose.

Thank you for this pointer and notes on caveats. By noting your suggested caveats, we think our discussion has become better focused. That said with the re-evaluation we now find the simulated inundation products are limited in accuracy when it comes to pointing out flooded extents.

Dr. Blodgett requested more information on the nature of the retrospective model run noting that it is only calibrated in some locations and should not be expected to produce realistic flow volumes. Additionally, he observed that the retrospective does not assimilate observed streamflow and suggested including such a remark.

Thank you for the comment, both of these have been noted (please see lines 130-134).

Complimenting the operational products are 23-year reanalysis studies for NWM versions 1.0, 1.2, and 2.0. These products use downscaled NLDAS-2 climate forcing's with the standard NWM configuration. Unlike the operational Analysis and Assimilation product however, the reanalysis products do not assimilate observed streamflow and have been calibrated in limited number of basins (Gochis, 2016)

Dr. Blodgett asked why we had not included NHD Areas in our masked-out regions.

Thank you for this suggestion. We were unaware of the NHD area product and have now included it in our mask. Moreover, the NHD Fcodes (for both water bodies and areas) have been listed in text to increase transparency (see lines 168-169).

For each event, a waterbody mask was created by combining the perennial NHD water bodies (NHD Fcode 39004, 39009) and NHDAreas (NHD FCode 40300, 40307, 40308, 40309) in each extent.

Dr. Blodgett aske to re consider our binning by stream order?

Despite variable density of streams across the country there was clear evidence in our evaluations that lower order reaches underpredicted flood extents while higher order reaches preformed better. We attribute this to the use of a single default Manning n coefficient in the SRC generation and discuss the implication of this in a few spots throughout the manuscript. Most relevantly, we show how stream order is a driving factor resulting in competing results seen at the floodplain and catchment analysis resolution.

Dr. Blodgett suggested generating a driving hypothesis and provided the example that "Given that HAND is not a physically based model in that it does not route flow over the landscape or preserve mass, we would expect small errors in stage to produce large errors in inundated areas in low-relief landscapes."

Thank you for this succinct explanation of the phenomenon we were trying to describe as "volume control" in regions of low-relief. This wording has been added to lines 396-398 and help clarify our point.

Since HAND is not a physical model, it is unable to conserve volume through space or time. In areas of low relief, where many cells have similar if not equal HAND values, small errors in stage can have disproportionate errors in inundation extent at the 10m grid cell resolution.

Dr. Blodgett asked us to re-think the distinction of 'Errors in the NHD' as a section heading

This point is greatly appreciated. Paragraph 2 in section 4.3.1 now starts:

"*With respect to the streamlines it is important to recognize that the NHD was developed as a cartographic representation of the nation's waterways and using a cartographic toolset for hydrologic modelling and routing applications has inherent limitations.*

And discusses the previously listed issues in this context. A new paragraph about the challenges with using a cartographic data as a modelling geofabric has been made in lines 364-377 with specific references to issues of refactoring catchment delineations to more compact and consistent modelling units. More over the section heading has been changed to better represent this and add a brief discussion of DEM resolution.

"Data Models: Use, Limitations, and Adaptions"

Again, thank you to both reviewers for helping make this paper substantially better than its original submission,

Sincerely,

Mike Johnson, Dinuke Munasinghe, Dami Eyelade, Sagy Cohen

References:

Zheng, X., Maidment, D. R., Tarboton, D. G., Liu, Y. Y., & Passalacqua, P. (2018). GeoFlood: Large-Scale Flood Inundation Mapping Based on High-Resolution Terrain Analysis. Water Resources Research, 54(12), 10-013.

Cohen, S., G. R. Brakenridge, A. Kettner, B. Bates, J. Nelson, R. McDonald, Y. Huang, D. Munasinghe, and J. Zhang (2018), Estimating Floodwater Depths from Flood Inundation Maps and Topography, *Journal of the American Water Resources Association*, 54 (4), 847–858.