Natural Hazards
and Earth System
Sciences

Open Access

EGU

An integrated evaluation of the National Water Model (NWM)–Height Above Nearest Drainage (HAND) flood mapping methodology

J. Michael Johnson[1], Dinuke Munasinghe[2], Damilola Eyelade[1], and Sagy Cohen[2]

[1]University of California, Santa Barbara, Santa Barbara, CA 93106, USA
[2]University of Alabama, Tuscaloosa, AL 35487, USA

Correspondence: J. Michael Johnson (jmj00@ucsb.edu)

Abstract. Flood maps are needed for emergency response, research, and planning. The Height Above Nearest Drainage (HAND) technique is a low-complexity, terrain-based approach for inundation mapping using elevation data, discharge–height relationships, and streamflow inputs. The recent operational capacities of the NOAA National Water Model (NWM) and preprocessed HAND products from the University of Texas offer an operational framework for real-time and forecast flood guidance across the US. In this study, we evaluate the integrated National Water Model – Height Above Nearest Drainage (NWM–HAND) flood mapping approach using 28 remotely sensed inundation maps and 54 reach-level catchments. The results show the NWM–HAND method tends to underpredict inundated cells in 4th-order and lower-order reaches but does better with a slight tendency to overpredict in high-order reaches. An evaluation of the roughness coefficient used in the production of synthetic rating curves suggests it is the most important parameter for correcting these errors. Persistent inaccuracies do occur when NWM streamflow predictions are substantially biased ($> 60\%$ mean absolute error between NWM and observed streamflow) and in regions of low relief. Overall, the NWM–HAND method does not accurately capture inundated cells but is quite capable of highlighting regions likely to be at risk in 4th-order streams and higher. While NWM–HAND should be used with caution when identifying flood boundaries or making decisions of whether a cell is dry or wet, its applicability as a high-level guidance tool along larger rivers is noteworthy.

1 Introduction

Floods are one of the deadliest natural hazards in the United States. Since 2010, floods have killed 950 people in the US (NOAA, 2019). While floods used to be associated with coastal zones, 8 of the 10 US states with the largest flooding disasters in the last decade were landlocked (Lightbody and Tompkins, 2018). Such patterns can be expected to continue as climate and land use pressures amplify (Hirabayashi et al., 2013; Yin et al., 2018). While flood damage cannot be eliminated, it can be mitigated. This is particularly true when populations are given advanced warning, have faith in the forecast accuracy, and are provided actionable intelligence (Johnson et al., 2016, 2018). When emergency responders are given maps showing where water is and where it might be in the future, better choices can be made that systematically save time, energy, and resources (National Research Council, 2009). Combined, these challenges suggest a need to better understand and forecast flood events across the entire United States (US).

Since 2015, efforts towards producing real-time and future inundation forecasts for the US have resulted in the compilation of a 10 m resolution Height Above Nearest Drainage (HAND) layer for the continental United States (CONUS: Liu et al., 2018). The methodology for coupling streamflow predictions from the National Water Model (NWM) and HAND was initiated as part of the National Flood Interoperability Experiment (NFIE) (Maidment, 2016). The methodology has since been enhanced and added to the National Water Center (NWC) US flood forecasting framework (NOAA National Water Center, 2018). The cur-

rent 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 these contexts and expectations should be tempered while recognizing the importance of having an operational, continental-scale flood forecasting system.

A number of studies have shown the HAND methodology can produce accurate inundation maps from a known streamflow and carefully created HAND layers (Afshari et al., 2018; Rennó et al., 2008; Rodda, 2005; Zheng et al., 2018b). There has not been, however, an exploration of how errors in the NWM–HAND methodology, and its individual components, might compound and propagate through flood simulation. In this paper we evaluate the integrated skill of NWM–HAND for simulating riverine flooding events by comparing NWM–HAND forecasts against an observational dataset at two scales, the floodplain and the reach-level catchment.

2 Background

The goal of this paper is to study the combined skill of the National Water Model (NWM) and the current CONUS-wide HAND layers to produce inundation maps suitable for emergency response and guidance. We do this by comparing simulated flood extents to an archive of observed events. In this section we introduce the HAND methodology, the existing CONUS-wide datasets, the National Water Model, and the US Flood Inundation Map Repository (USFIMR).

2.1 Height Above Nearest Drainage (HAND) method

As the name implies, a Height Above Nearest Drainage (HAND) map contains the vertical distance between a location and its nearest stream. Producing a HAND map requires a digital elevation model (DEM) and a spatial representation of a region's river network. In the USA, these primarily come from the USGS National Elevation and Hydrography Datasets respectively (NED and NHD).

Each object in the NHD is described as a "reach" and has three realizations including the (1) outflow point, (2) the catchment polygon, and (3) the flow path polyline. Each realization refers to the same feature and is indexed by a unique common identifier (COMID). Figure 1a illustrates a rasterized object in all three realizations.

In the HAND methodology, a catchment mask is created by rasterizing the catchment polygon to the DEM grid and assigning grid cell values equal to the COMID of the reach (e.g., 101) (Fig. 1b). To generate a HAND raster, the flow path is used to determine in-stream and out-of-stream cells. For all out-of-stream cells, the relief between that cell and the nearest in-stream cell is calculated and stored as the HAND cell value (Fig. 1c).

Once computed, HAND raster's can be used to define hydraulic flood plain cross sections. The derived geometries provide cross-sectional areas (A) and hydraulic radius (R) inputs for the Manning's equation (Zheng et al., 2018a, Eq. 1):

$$Q(y) = \frac{1}{N} \times A \times R^{\frac{2}{3}} \times S^{\frac{1}{2}}, \tag{1}$$

where $Q(y)$ is the discharge for a given stage, N is the Manning's roughness coefficient, and S is the reach slope. From this equation, discharge (Q) requirements can be calculated for a range of stage values and stored as a synthetic rating curve (SRC; Zheng et al., 2018a).

Given a flow value at a reach outlet (either observed or modeled) a stage can be generated from the SRC (Fig. 1d). If more than one catchment is being modeled, the synthetic stage values can be stored according to COMID in a table (Fig. 1e). Using this table, a new raster of water depths can be created by reclassifying the catchment mask raster (Fig. 1f). Finally, subtracting the HAND raster from a water depth raster yields an inundation map of above-surface water depths (Fig. 1g). Any value greater than zero indicates inundation and values less than zero indicate dry cells (Fig. 1h).

2.2 HAND products for the continental US

In 2017, HAND raster's and SRCs were generated for CONUS using the 10 m 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., 2018a). In the precomputed HAND rasters, relief was calculated via the TauDEM distance down function (Tesfa et al., 2011). In the SRC creation, reach slope was taken for the NHD attribute table, while a constant channel roughness (N) of 0.05 has been adopted for all of CONUS (Zheng et al., 2018b; Johnson and Coll, 2017). All products are catalogued by HUC6 (the basin level units in the WBD) on the University of Texas (UT) Corral server (https://web.corral.tacc.utexas.edu/nfiedata/, last access: 26 September 2019).

2.3 National Water Model (NWM)

The NWM serves as a cornerstone of the new NOAA Water Initiative to provide integrated predictive capabilities that promote resilience and mitigation of water risks (Graziano et al., 2017). In August 2016, version 1.0 was made operational, expanding the nation's forecasting domain from approximately 9000 gaged locations to 2.7 million reach outlets along the NHDPlusV2 (Gochis et al., 2018). The National Water Model uses the Noah Multi-Parameterization (Noah-MP) land surface model (LSM) to simulate land surface processes across a 1 km grid and a separate 250 m grid to perform diffusive wave surface and saturated subsurface flow routing (Gochis et al., 2018). Once water is in channel, WRF-Hydro implements a standard 1-D Muskingum–Cunge (MC)
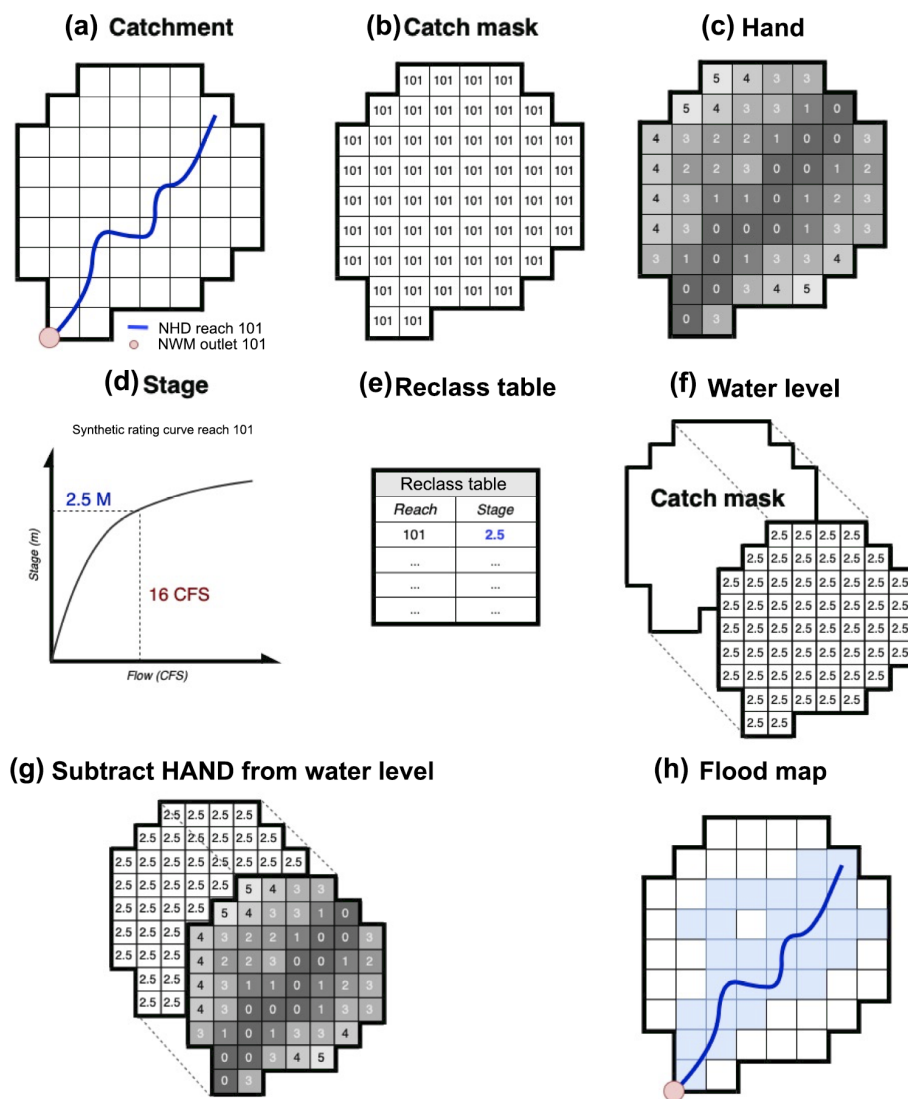
Figure 1. The HAND methodological workflow: **(a)** the contributing catchment to a defined outlet is rasterized to the resolution of the supporting DEM. **(b)** This raster is reclassified to match the ID of the outlet. **(c)** In-stream and out-of-stream cells are identified using the flow line vector. The relief between all out-of-stream cells from their nearest in-stream cell is calculated to define a HAND raster. **(d)** From a given flow rate, a rating curve can be used to convert flow to stage. **(e)** A reclassification table can be built relating reach ID to stage. **(f)** The table can then be used to reclassify the catchment mask, into a water-level raster. **(g)** Subtracting the HAND raster from the water-level raster yields a water-level above-surface raster. All values less than 0 can be set to 0 and the remaining show the estimated flood height at each cell **(h)**.

hydrograph routing method using time-varying parameter estimates.

Complementing 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 forcings 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 a limited number of basins (Gochis et al., 2016). In this study, the v1.2 reanalysis product was used (https://

registry.opendata.aws/nwm-archive/, last access: 26 September 2019).

2.4 US Flood Inundation Map Repository (USFIMR)

Following the release of the NWM, academic partners at the University of Alabama developed the US Flood Inundation Map Repository (USFIMR) to provide inundation maps for past US flood events. These maps were derived using image classification techniques and a number of satellite sensors (e.g., Landsat, Sentinel-1, Sentinel-2) with some ground truthing based on secondary sources (e.g., news reports, so-

cial media). Such maps are useful for model calibration, validation, and flood susceptibility assessment (Cohen et al., 2018; Munasinghe et al., 2018). The USFIMR web portal provides more information on each flood, the specific sensor, and supplementary data including NED elevation and upstream NWIS hyperlinks (http://sdml.ua.edu/usfimr, last access: 26 September 2019). A catalogue of the USFIMR maps used in this study can be found in Table A1. Here it is noted that the flood ID assigned to each flood is consistent with those used in the USFIMR and not necessarily sequential. Using these datasets, we address three main research questions:

1. Is the HAND method viable for continental flood mapping?

2. Are the current NWM forecasts, HAND layers, and SRCs adequate for providing operational flood guidance?

3. How do these three datasets contribute to errors in simulation?

3   Methods

This section will describe the creation of our simulation maps and the metrics used to evaluate accuracy.

3.1   Floodplain maps

For each USFIMR map, the flood polygon was projected from NAD83/Conus Albers (CRS 5070) to a WGS84 coordinate reference system (CRS 4269). For each shapefile, a clipping extent, derived as a concave hull, was created to ensure all pixels being evaluated were within the USFIMR classification bounds. These clipping extents were then used to subset the NHD and extract a list of COMIDs and HUC6 identifiers.

For each HUC6, the HAND, catchment mask, and rating curve products were downloaded from the UT server and processed. 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 (Sect. 2.1).

For each event, a water body 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. These features were then rasterized to the 10 m HAND grid using the fasterize R package (Ross, 2018). All cells that were not within the concave hull or covered by a water body mask, were set to NA prior to comparison. This process was repeated for each USFIMR flood and the process is formalized in the AOI, HydroData, and Flood Mapping R packages (Johnson, 2018, 2019a, b).

Table 1. Flood comparison confusion matrix.

	Modeled wet	Modeled dry
Observed wet	WW (true wet)	WD (false dry)
Observed dry	DW (false wet)	DD (true dry)

3.2   Catchment scale

In this paper we are also interested in isolating the biases stemming from NWM forecasts and the HAND rasters and derived SRCs layers using gaged catchments. To do this, we identified all USGS National Water Information System (NWIS) gages within the concave hull of each USFIMR polygon. For each gage, the collocated COMID was identified using the Network Linked Data Index (HydroData binding). Gages were kept if there was a NWIS and NWM record for the date and hour of interest and the complete catchment was contained within the USFIMR concave hull. In total, 54 unique catchments were identified.

For each catchment two unique flood maps were produced. The first used the NWM forecast and the SRC rating curves, while the second used the NWIS observation. Similar to the floodplain maps, a water body mask was used to remove known cases of standing water.

3.3   Metrics used for comparison

With our simulated and observed flood maps we are interested in quantifying two types of agreement. The first is the agreement between the simulated (NWM–HAND) and observed (USFIMR) maps. The second is the agreement between maps simulated with observed streamflow (NWIS–HAND) and modeled streamflow (NWM–HAND).

3.3.1   Comparing simulations to observed events

To compare simulations to observed extents, a simulated and an observed map are intersected and classified into the four categories seen in Table 1. In these categories, "W" refers to a wet cell and "D" refers to a dry cell. The first letter in each category refers to the state of the cell in the observed map while the second refers to the state of the cell in the simulation.

With this classification, four metrics were established to describe the area ratio, rate of accuracy, rate of overprediction, and rate of underprediction. The last three of these are derivatives of the fit statistic used by Sangwan and Merwade (2015).

The spatially agnostic area ratio (Eq. 2) helps establish if the simulated extent is inundating the correct number of cells regardless of spatial accuracy. Any value less than 1 indicates the NWM–HAND method is not inundating enough cells while a value greater than 1 indicates too many cells were inundated.

$$\text{AreaRatio} = \frac{\text{total wet simulation cells}}{\text{total wet observed cells}} \qquad (2)$$

Overall accuracy is quantified as the number of cells that were correctly identified as wet (WW) divided by the number of all wet cells in the simulated and observed rasters (Eq. 3):

$$\text{accurate}(A) = \frac{\text{WW}}{\text{WW} + \text{WD} + \text{DW}}. \qquad (3)$$

Similarly, areas of underprediction are quantified as

$$\text{under}(U) = \frac{\text{WD}}{\text{WW} + \text{WD} + \text{DW}}. \qquad (4)$$

And areas of overprediction are quantified as

$$\text{over}(O) = \frac{\text{DW}}{\text{WW} + \text{WD} + \text{DW}}. \qquad (5)$$

Together, the "accurate", "over", and "under" prediction metrics form our AOU statistics. The use of a common denominator across these statistics assures the summation of AOU equals 1. The choice of a 2-D 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 capture 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).

3.3.2 Comparing simulation to simulation

In our 54 gaged catchments we have simulations driven by NWM forecasts, simulations driven by NWIS observations, and observed USFIMR events. To compare how the inclusion of known streamflow values change the overall accuracy of the simulated map, we calculate the AOU metrics for each set and then compute the standardized difference in the accuracy (A).

$$\Delta A = \frac{A_{\text{NWIS}} - A_{\text{NWM}}}{A_{\text{NWIS}}} \qquad (6)$$

Positive values indicate the use of NWIS data increased the accuracy of the produced flood map while negative values indicate the use of NWIS decreased overall accuracy. Finally, a normalized mean error (nME) and normalized mean absolute mean error (nMAE) will be used to compare any two measurements (e.g., stage or streamflow). In these, the control is the observed value and simulation is the estimated value.

$$\text{nME} = \frac{\text{control} - \text{simulation}}{\text{control}} \qquad (7)$$

$$\text{nMAE} = \frac{\text{abs}(\text{control} - \text{simulation})}{\text{control}} \qquad (8)$$
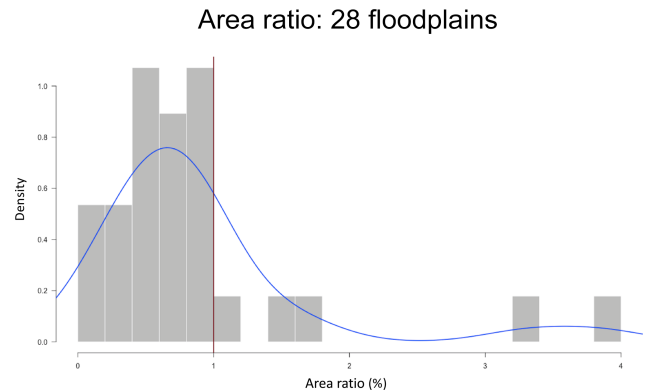
Figure 2. A histogram of area ratios for 28 floodplains shows a tendency for the NWM–HAND to underpredict flood extents.

Table 2. Mean and median AOU statistics for the 28 floodplain studies.

	Median	Mean
Accurate	0.15	0.19
Over	0.25	0.27
Under	0.47	0.53

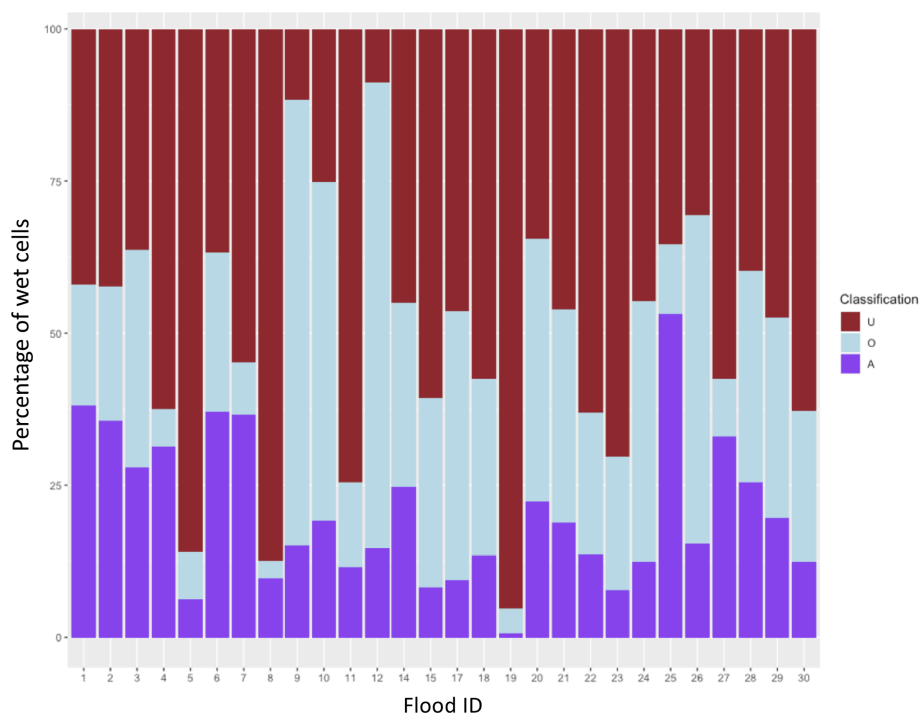## 4 Results

4.1 Floodplain analysis

First, we evaluate the ability of the NWM–HAND method to produce the correct amount of inundated area at the floodplain scale by plotting the area ratios (Eq. 2).

Figure 2 shows that the overall median of our sample was 73 %, indicating a strong bias towards underprediction. Moreover, 82 % of the floodplain maps had an area ratio less than 1. While there are fewer cases of overprediction ($n = 5$), the mean area ratio of 230.4 % indicates that while instances of overprediction are less frequent (~ 18 % of our cases), errors are generally much larger.

From a spatially explicit perspective we evaluated how the NWM–HAND approach captures cell-level inundation using the AOU statistics. Table 2 shows that, of all flooded cells (in either the simulation or observed data), an average of 53 % are underpredicted, 27 % are overpredicted, and 19 % were correctly identified.

Figure 3 shows the AOU metrics for each floodplain as a stacked bar plot, further emphasizing the tendency toward underprediction across the entire set of floodplains.

These statistics show a clear signal that the NWM–HAND method is limited in its ability to capture inundation and is almost twice as likely to underpredict than overpredict the state of a missed cell. Noting the stated intention of NWM–HAND as an operational flood mapping framework for guidance, this tendency raises concerns.

Figure 3. Stacked bar plot of the accuracy, over, and under prediction (AOU) metrics for each floodplain case study. Each bar represents 100 % of the inundated cells in the intersected simulation and observed events.

4.2 Catchment area comparison

In this section we analyze the 54 gaged catchments to better understand the influences driving performance. As a reminder, for each of these catchments we have two simulated extents, one using the NWM forecasts and one using the NWIS observed flows. First, we evaluate the ability of the NWM–HAND method to produce the correct amount of inundated area by plotting the area ratios comparing the total number of wet cells in the simulated and observed extents (Fig. 4).

When compared to Fig. 2, we see the peak in the density curve now sits much closer to 1 with more equal proportions on either side. This indicates that at the catchment level, NWM–HAND is better able to capture the total area of inundation. However, we see that the overpredicted cases are more extreme, causing the right tail to elongate.

4.2.1 Influences of NWM streamflow inputs

The median and mean statistics for the NWM- and NWIS-driven AOU metrics are presented in Table 3.

Table 3 highlights that using known streamflow values generates no change in average accuracy with a slight increase in overprediction and corresponding decrease in underprediction ($\sim$ 3 %). To better understand how well the AOU metrics of the NWIS-driven maps explain the variation in the NWM-driven maps, linear models were fit to describe $(\mathrm{NWIS_A} \sim \mathrm{NWM_A})$, $(\mathrm{NWIS_O} \sim \mathrm{NWM_O})$ and
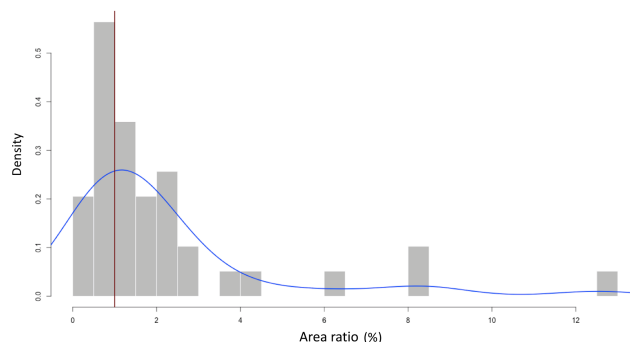

Figure 4. A histogram of area ratios for 54 catchments shows a tendency for the NWM to slightly overpredict flood area and a long right tail of extreme overprediction.

Table 3. Mean and median values for NWM and NWIS AOU metrics and R^2 values for linear fits between NWIS and NWM.

	Median NWM/NWIS	Mean NWM/NWIS	R^2 NWIS/NWM
Accurate	16 %/16 %	25 %/25 %	.93
Over	45 %/47 %	44 %/47 %	.82
Under	26 %/21 %	31 %/28 %	.75
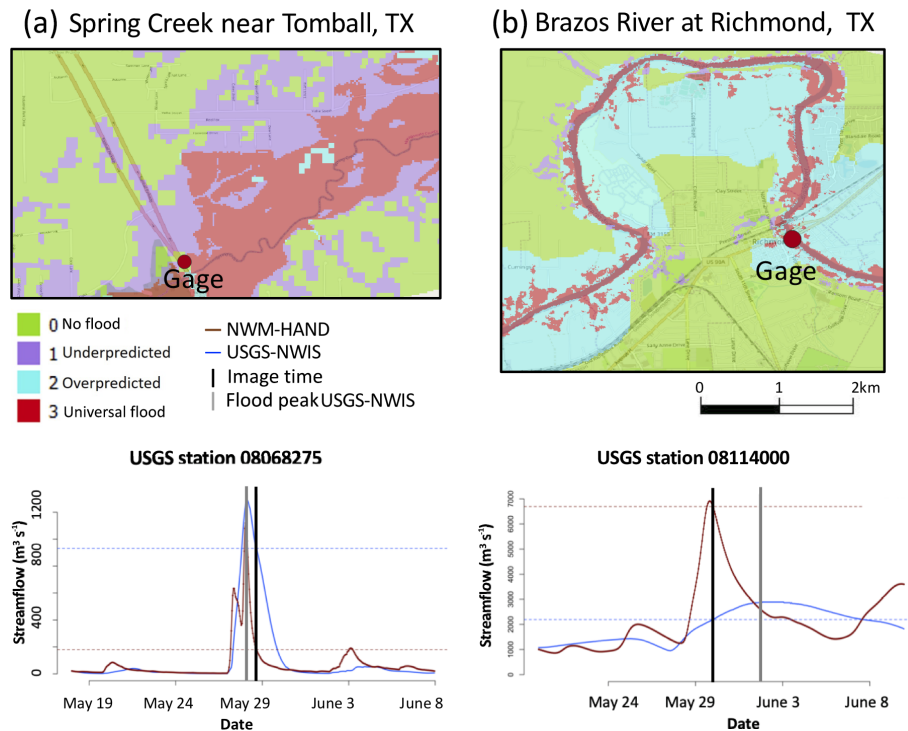
Figure 5. Prediction errors due to NWM flow forecast when nMAE exceeds 60 %: **(a)** NWM underprediction results in underpredicted flood extents. **(b)** NWM overprediction results in overpredicted flood extent.

(NWIS$_U \sim$ NWM$_U$). The R^2 values of these models can be seen in Table 3 as well. From these results we can conclude that approximately 7 % of the variation in accurate prediction can be attributed to inaccuracies in the NWM forecasts. Similarly, 18 % of the variation in overprediction and 25 % of the variation in underprediction can be explained by inaccuracies in the NWM. All said, this suggests that even with perfect streamflow inputs, current implementations of HAND and SRCs are only capturing ~ 25 % of the inundated cells correctly, and a minimal percentage can be explained by variation in NWM forecasts. Moreover, the errors in NWM forecasts have a limited influence on whether cells are accurately captured but a strong influence on whether missed cells are under- or overpredicted.

Table 3 also reveals that, regardless of streamflow input, the tendency at the catchment level is to overpredict (47 %) rather than underpredict (28 %) inundation. This is opposite the pattern seen in the floodplain maps despite the average tendency for the NWM to underpredict streamflow in these events (nME of -20 %).

To confirm our conclusion that NWM forecasts have a minimal impact on inundation accuracy, we need to ensure that the adverse impacts of bad simulations are not masked by many accurate forecasts. To better understand the relative bias introduced by the NWM, we grouped our 54 catchments by their nMAE. For each set of basins grouped by nMAE,

the number of gages and absolute mean change in accuracy (ΔA; Eq. 6) are shown in Table 4.

Looking at the absolute mean change in accuracy between the NWIS and NWM maps, there is evidence that the influence of the NWM on NWM–HAND accuracies is minimal until nMAE exceeds 60 % when there is a clear jump in the absolute mean ΔA from less than a 10 % relative loss to more than 30 %.

When the NWM exceeds nMAE of 60 %, predictable patterns of error occur. Figure 5a (Flood ID 4) shows a case where NWM–HAND underpredicts inundation due to an underpredicted NWM forecast while Fig. 5b (Flood ID 9) shows a case where an overpredicted NWM forecast results in overpredicted flood extents.

In addition to errors in streamflow magnitude, NWM timing errors can introduce issues related to inundation forecasting and its evolution. Even though we did not explicitly study the ability of NWM–HAND to capture the temporal evolution of flood extents in this study, a few instances where the NWM improperly estimated the receding limb of the hydrograph showed the inability of NWM–HAND to capture ponded and receding waters, leading to issues of under- and overprediction, respectively.

At this point we have found that at the floodplain level, NWM–HAND tends to accurately capture ~ 19 % of inundated cells with almost twice the chance of under- than overpredicting a missed cell state. In contrast, at the catchment

Table 4. Absolute mean ΔA gage count and grouped by nMAE between NWIS and NWM streamflow. The bold row highlights a distinct break in the data. Once the normalized mean average error in the NWM forecasts exceeds 60 %, the accuracy in flood maps decreases substantially.

nMAE range	Gages (cumulative %)	Absolute mean ΔA %
0 %–10 %	9 (16.67 %)	0.68
10 %–20 %	11 (37 %)	1.91
20 %–30 %	4 (44 %)	4.18
30 %–40 %	3 (50 %)	6.36
40 %–50 %	3 (55.8 %)	6.81
50 %–60 %	6 (66.7 %)	8.39
60 %–70 %	**8 (81.49 %)**	**31.77**
70 %–80 %	1 (83.3 %)	69.52
80 %–90 %	3 (88.90 %)	30.15
90 %–100 %	2 (92.60 %)	84.36
> 100 %	4 (100 %)	391.42

level NWM–HAND is able to accurately capture $\sim 25\,\%$ of the inundated cells with a more equal tendency to over- and underpredict missed cell states, with a stronger tendency to overpredict.

Dissecting this divergent pattern led us to think about the general state of gaged reaches. Most often gages are put on larger rivers and less so on smaller reaches and tributaries. In fact, of the 54 catchments in our study, 90 % were on a reach with a Strahler order of 4 or higher. In contrast, when all catchments in each of our USFIMR floodplains were aggregated, low-order reaches made up, on average, 81 % of the total stream networks. In many respects, this makes our catchments a biased sample favoring high-order reaches.

Therefore, our hypothesis is that the difference in over- and underprediction patterns seen between the floodplains and catchments can be explained by the difference in the sample and population. That is, low-order reaches tend to underpredict while higher-order reaches are more neutral with a tendency towards overprediction. From here on out, lower-order reaches refer to those with a Strahler order of 3 or less, and higher-order reaches refers to those with a Strahler order or 4 or higher.

4.2.2 Underprediction in lower-order reaches

A visual analysis of each of our floodplains reveals a distinct pattern of underprediction in lower-order reaches; however due the general nature of gaged reaches we only have one gage on a 2nd-order river (Fig. 6, Flood ID 8 USGS station ID 8068325). At this reach, the NWM estimated $50\,\mathrm{m^3\,s^{-1}}$ (SRC stage $= 1.82\,\mathrm{m}$) flowing through the channel. The recorded NWIS flow is $80\,\mathrm{m^3\,s^{-1}}$ (nMAE $= 0.38$, SRC stage $= 2.43\,\mathrm{m}$). A cross section was made across the HAND raster at this reach indicating a stage of 3.3 m is re-
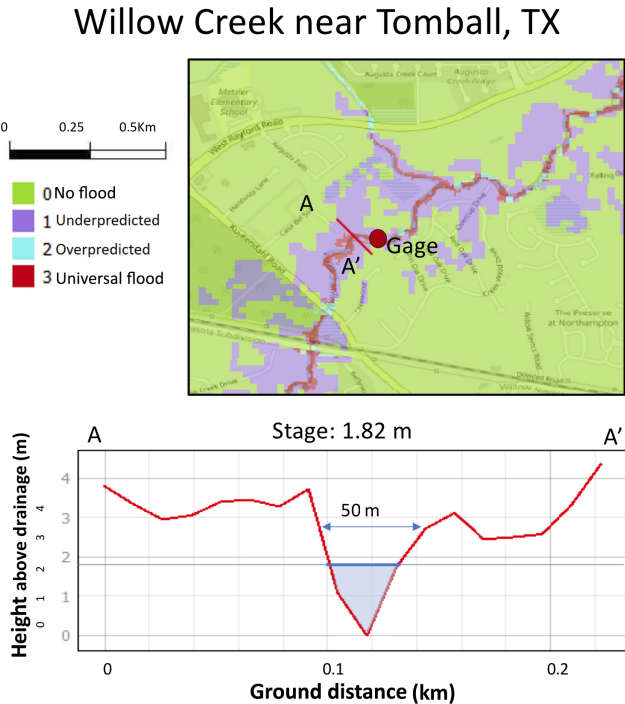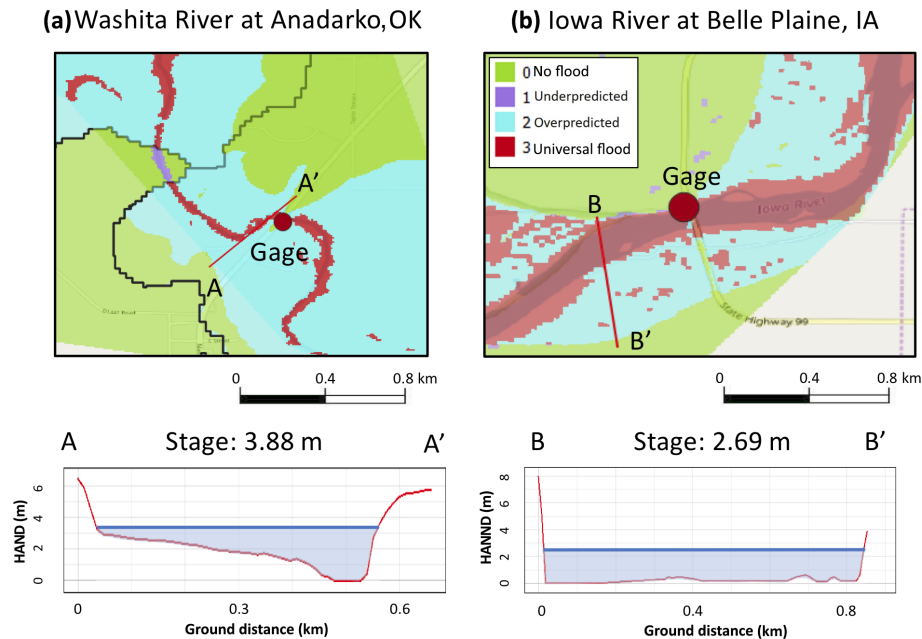

Figure 6. Inaccurate synthetic rating curve relationship causes underprediction in a 2nd-order stream.

quired to inundate the right bank and 3.8 m is needed to inundate the left bank to the levels seen in the USFIMR.

In this case, even if the NWM had accurately predicted streamflow, water would have remained confined to the channel, resulting in underprediction. Recognizing a mismatch between the known Q ($80\,\mathrm{m^3\,s^{-1}}$) and stage (3.8 m) we explored the assumptions driving the SRC for this reach (Eq. 1). 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 $80\,\mathrm{m^3\,s^{-1}}$. 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. Specifically, we found that an N of 0.16 was needed to achieve an observed streamflow–stage relationship at this reach. Based on this representative example we suggest that systematically increasing roughness in lower-order reaches may help mitigate underprediction.

Other research that applies N values by stream order offer similar estimates to what was found in our example ($N = 0.12$ for 2nd-order reaches; Li, 2016). Certainly, more research is needed on the relative role of roughness in the NWM–HAND method and the best way to optimize it across different stream orders and geographies.

Figure 7. Overprediction in catchments with low relief. **(a)** This catchment is influenced by a hypersensitive synthetic rating curve relationship. **(b)** This catchment exemplifies a case where the HAND layers will be unable to accurately capture flooding extents due to essentially 0 relief.

4.2.3 Overprediction in areas of low relief

The other reoccurring error we noticed in our analysis was the extreme right tail seen in the area ratio plot for catchments (Fig. 4). When we isolated the poorly performing catchments it became evident that the extreme overprediction occurred in areas of low relief. Overall, we found two types of low-relief cross sections: those that are hypersensitive to errors in stage and those that are fundamentally limited by the terrain representation in the HAND layer.

Sensitivity to SRCs

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 10 m grid cell resolution. The Washita River at Anadarko, OK (Fig. 7a), is a 6th-order stream with relatively uniform overbank terrain. The NWM estimates $473 \, \mathrm{m^3 \, s^{-1}}$ (SRC stage $= 3.81$ m) while the recorded flow was $495 \, \mathrm{m^3 \, s^{-1}}$ (SRC stage $= 3.88$ m). The nMAE of 4 % suggests the source of overprediction originates in the terrain layer or SRC and not the NWM forecast. For this reach, we optimized the SRC for the given cross-sectional area at an observed stage of 1.2 m and solved for a Q of $495 \, \mathrm{m^3 \, s^{-1}}$. To achieve this, N was reduced to 0.005, effectively increasing the carrying capacity of the river. This solution supports the pattern of decreasing N in higher-order rivers and this em-

phasizes that fact that in areas with low relief, small errors in stage and or discharge can result in extreme overpredictions.

Concerns with raster resolution in areas of low relief

We examined a second type of low-relief profile where optimizing the hydraulic parameters has no effect due to the form of the HAND raster. This is illustrated with a cross section along the Iowa River at Belle Plain, IA (Fig. 7b Flood ID 17). The NWIS-observed streamflow for the reach was $971 \, \mathrm{m^3 \, s^{-1}}$ (stage $= 2.69$ m) while the NWM-predicted streamflow was $958 \, \mathrm{m^3 \, s^{-1}}$ (stage $= 2.66$). The issue with this reach (and others like it) is that the area currently simulated as flooded will flood for any stage exceeding 0.5 m regardless of the volume of water within the channel. It is in these regions that the 10 m grid cell resolution of the underlaying DEM and absolute vertical accuracy of the NED (2.44 m; FEMA, 2007) appear most limiting.

We suggest two possible improvements for low-relief catchments. The first is refactoring of the NHD catchments (or the effective HAND catchments) into more uniform units avoiding large, flat drainages and instances of many small units. Tools for NHD refactoring are already in development (Blodgett, 2018), and the idea of refactoring speaks to issues to be raised in Sect. 4.6. 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:

$$\text{cross sectional area} = \frac{Q}{\text{velocity}}. \tag{9}$$
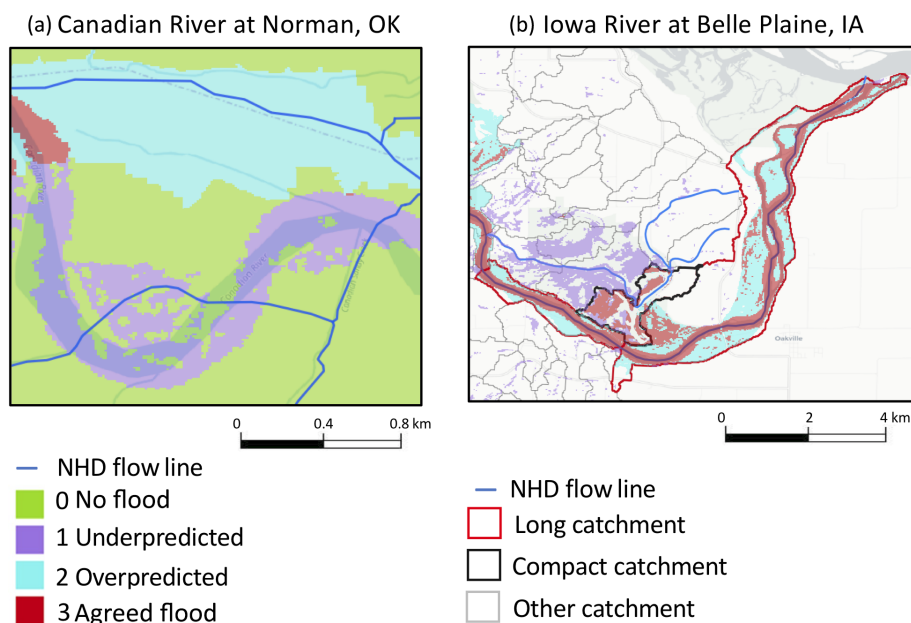
Figure 8. Unaccounted-for challenges when using the NHD as a modeling fabric. **(a)** This catchment shows a case where the NHD and NED are not capturing the true meander of a river channel. **(b)** This set of catchments highlights the influence of long nonuniform catchments and challenges with catchment definition.

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 lookup 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 an understanding of how the NWM is handling hydraulics and thus velocity and a test of how variations in velocity impact volume estimation. Both are interesting pursuits in their own right but out of the scope of this paper.

4.3 Other source of error/influence

While the primary purpose of this study was to explore the impacts of NWM and HAND data layers and assumptions on NWM–HAND simulations, this section looks at two other issues that influence simulation accuracy: namely, the data models underpinning the HAND layers and the limitations in remotely sensed ground truths.

4.3.1 Data models: use, limitations, and adaptions

The digital elevation model (DEM) used to prepare the current HAND layers is a 10 m product that effectively smooths the surface of the earth. In areas of low-relief, higher-resolution elevation data would be able to produce more sen-

sitive rating curves. In other areas, attention to the methods and datasets (streamlines) used to condition the DEM should be considered.

With respect to 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 modeling and routing applications has inherent limitations. In Fig. 8a (Flood ID 12), the NHD does not capture the meandering segment of the Canadian River's main channel. In such a situation, even with accurate NWM forecasts and rating curve relationships, the predicted flood extent will be spatially misaligned. Similar situations can also occur in areas with braided channels or wide estuaries. Standard DEM-based stream delineation tools (like those used in NHD) only allow one channel to be designated for each reach and often result in a poor representation of network density and/or channel width. Prediction errors resulting from network density/missing reach issues can be seen in Flood ID 8, while errors due to channel and or coastal feature representation can be seen in Flood IDs 19, 23, and 24.

On a more theoretical side, the use of a cartographic dataset as a model base layer often involves some form of manipulation. Up until this point, the NHD has been used in its original form without considerations for the implications for modeling. As the NHD becomes a more central component of the NWM framework, the goal of creating more specific and consistent modeling units while preserving the network is of value. To illustrate this, Fig. 8b shows a collection of catchments in Flood ID 17. The two principle catchments experiencing flooding in this image are highlighted in red.
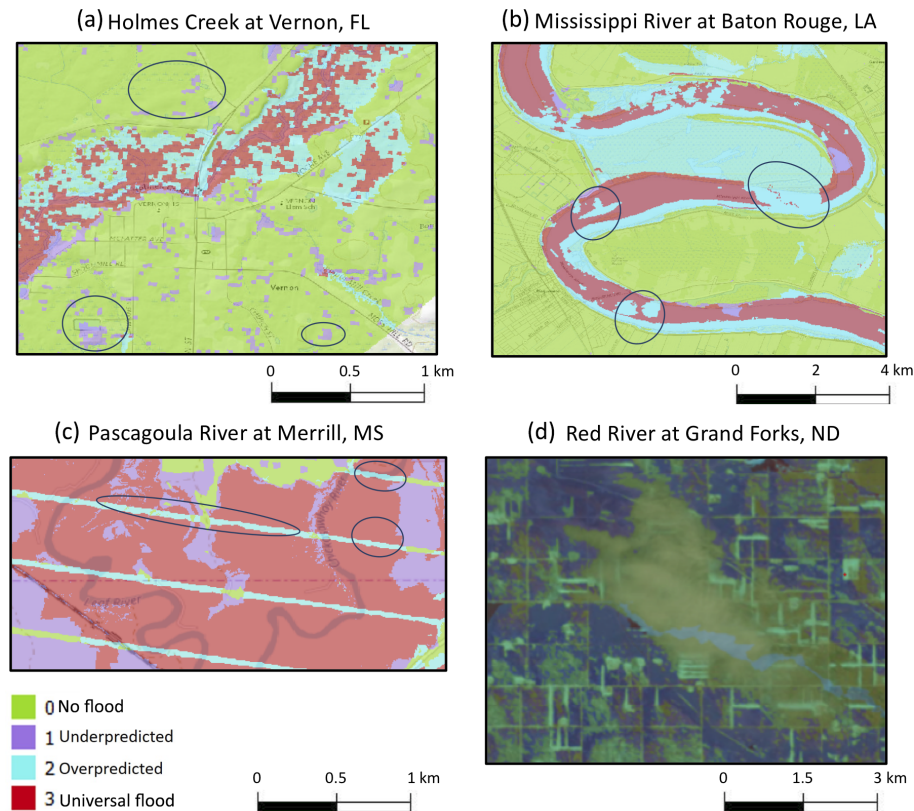
Figure 9. Errors due to classification and image artifacts: **(a)** dark vegetation misclassified as water due to spectral similarity. **(b)** Misclassified channel segments due to spectral similarity. **(c)** Scan line errors in input RS image create gaps in flooded extent. **(d)** Clouds and melted snow create ambiguities in flooded extent recognition.

Both catchments are long and narrow, making them inconsistent with the others in the region. One issue with these long catchments can be seen in the leftmost red catchment. To the north of the catchment (in the "black" catchment) we see a large swath of underpredicted flooding (purple). Looking at the NHD reach draining the black catchment it is easy to see that the area of missed flooding did not stem from the catchment reach but from the long red catchment. This implies that flood water crossed the catchment boundaries in a way that cannot be captured by HAND. Ultimately, integrating cartographic, sensor, and model domain data will be an issue for all involved fields in the near future, and in some respects lay out the challenges for the next 20 years (Clarke et al., 2019).

4.3.2 Remote sensing (RS) challenges

In all analyses we treated the USFIMR maps as ground truth. For the sake of a robust comparison it is worth highlighting some of the issues that arise in the creation of remotely sensed inundation maps and their consequences in these comparisons.

Pixel misclassification

The spectral similarity of dark vegetation and water pixels in optical satellite imagery contributes to uncertainties in both directions. Most often, these errors resulted in isolated patches of dark vegetation being classified as water (Fig. 9a, Flood ID 2). Other cases of pixel misclassification occurred in Flood IDs 1, 3, 16, 17, 18, and 23. Water was also incorrectly classified in a few notable cases. For example, pixels clearly within the main channel of the Mississippi River were classified as unflooded in Fig. 9b (Flood ID 13). In this example, the NWM–HAND methodology correctly identified such locations as being flooded (despite overpredicting the overall flood extent). These error types were also present in flood IDs 19 and 23.

Image artifacts

Image artifacts also impact our comparison. The first is the well-known scan line error present in Landsat 7 images processed after 2003 (Scaramuzza and Barsi, 2005). Despite gap filling, some data gaps still remain. The gaps created by the scan line error (Fig. 9c; Flood ID 4) created situations where both methods missed a flood or NWM–HAND alone pre-

dicts a flood, even though flooding could be detected from surrounding pixels. The overall extent covered by scan line gaps did not have a large impact as this issue affected only a relatively small percentage of the area in one case study.

The second artifact encountered was cloud cover, which had a strong effect in one case study (Fig. 9d; Flood ID 5). The image used to generate the USFIMR map was taken from a Landsat scene with 40 % cloud cover, making it difficult to accurately classify some segments of the image. This comparison also contained large regions of freshly melted snow (visible in the Landsat image as well as from the road network and farm fields in the classified image). Cloud-related classification errors also affected Flood ID 15, but to a lesser degree.

In all cases, these errors only occurred in minor portions of a few floods used in the analysis. We raise them not because they adversely impacted our results but for transparency and to provide readers with an understanding that no form of ground truthing is infallible. The rigorous development, successful use in other studies, and our careful evaluation of each flood extent give us confidence in their overall accuracy. More importantly, such a resource is the only option for validating modeled inundation extents at the scales we are exploring. As the NWM–HAND method continues to improve and be developed and further validated, resources like the USFIMR will only become more necessary.

5   Discussion

In its current form NWM–HAND is capturing 19 %–25 % of inundated cells accurately. At the floodplain resolution, NWM–HAND has a larger tendency to under (53 %) than overpredict (27 %). In contrast, we saw a larger tendency to overpredict (47 %) than underpredict (28 %) inundation at the catchment level. We argue this tendency to underpredict area at the floodplain resolution was an aggregate result of consistently underpredicting many small, lower-order catchments.

While the overall NWM–HAND methods certainly have a long way to go, we contend the results demonstrate the NWM–HAND approach is useful for general guidance and risk identification but may not be suitable for pixel-level analysis, resource allocation, or risk communication. Despite the limited accuracy found in this study, the NWM–HAND is quite an achievement that should not be discounted. The fact that an uncalibrated continental-scale model can rapidly generate flood inundation maps is of great value, and more importantly, the base layers and conceptual framework underpinning the model offer the research community a resource to improve, modify, and manipulate.

Looking towards a future of research and development and the inevitable next version of the base layers, issues of availability arise. Currently, accessing and combining the needed NWM–HAND products is cumbersome and for regions straddling one or more HUC6 can be data and process-ing intensive. While services like the FloodMapping R package can streamline some of the steps, disseminating data by HUC6 is not a convenient choice for users. Instead HAND products could be distributed via a web service like those used to distribute climate data and NWM gridded output – or built into services like the CUAHSI sub-setter (CUASHI, 2019). With workable foundations now in place, engaging with the communities that will use these products for research, for map creation, and ultimately in the field seems like a next step worth considering.

Returning to our research questions, we believe the HAND method is viable for continental flood mapping aimed at guidance, and not cell-by-cell decision-making. With this caveat in mind, current NWM forecasts are accurate enough in about two-thirds of our situations (a condition that will arguably improve in the operational model with data assimilation and stronger calibration) and the HAND terrain layers are adequate in all but areas of low relief. The synthetic rating curves are the source of most error, but can be greatly improved through a closer examination of Manning's N. In all of our tests, the existing cross-sectional geometries could be altered to provide correct Q when varying Manning's N within a range of 0.0001 and 0.2. To us, the potential of the existing data and methods should be weighted more heavily in the mind of readers than the reported low accuracies.

6   Conclusion

This study offers a high-level evaluation of the confidence we can place in the operational NWM–HAND forecasts. In its current state the NWM–HAND methods have limited ability to accurately capture inundation and their skill is more constrained by the terrain and SRC inputs than NWM accuracy.

At floodplain level, NWM–HAND tends to underestimate flooding area (82 % of case studies). Of these case studies, the mean likelihood of underpredicting a missed cell (53 %) is twice that of overprediction (27 %). At the catchment scale, NWM–HAND was better able to capture the total area of inundation but was more likely to overpredict (44 %) than underpredict (31 %) missed cells. We attribute this disparity to systemic underprediction in lower-order reaches. An analysis of NWM–HAND's sensitivity to changes in Manning's N and cross-sectional geometries indicates that SRCs are insensitive to changes in hydraulic radius (ergo wetted perimeters) but are very sensitive to changes in Manning's N.

As a general rule we believe the default Manning's coefficient used by HAND is too small in low-order reaches and arguably too large in higher-order reaches. While we only tested this on the small sample of available catchments, the theory behind the Manning values, and the consistent propensity to underpredict low-order reaches would support such a change. In every cross-sectional geometry we tested, observed streamflow (NWIS)–stage (USFIMR) relationships

could be achieved by varying N except in catchments with zero relief (e.g., Fig. 7b).

We further investigated the level of bias in NWM–HAND maps coming from the NWM forecasts and found the method to be relatively insensitive until the normalized absolute error between NWM and observed flow exceeded 60 %. Above this threshold, predictable patterns of large under- or overprediction occur in the direction of the NWM miss.

Another reoccurring error in our simulations was overprediction in catchments with low relief. Some of these catchments allow incremental changes in stage to have large impacts on inundation extent. We suggest that a refactoring of the NHD to smaller more uniform units and a consideration for how volume rather than stage may be used to fill catchments may be possible solutions. Moreover, the former offers some solutions to issues arising from treating the cartographic NHD as a modeling infrastructure. That said, neither of these is trivial to accomplish or validate.

While the current ability to capture accurate inundation is limited, our analysis revealed no fundamental errors that would render NWM–HAND unusable. Instead, we found that with modifications, accurate inundation is generally achievable at the level needed for guidance and emergency management. Therefore, we end with great optimism that further development, research, and evaluation around the NWM–HAND method can deliver on the promise of a CONUS-scale flood modeling enterprise.

Appendix A

Table A1. Flood events used for comparison.

Flood ID	River name	Flood date	Satellite sensor	State	Flood area (km^2)
1	Choctawhatchee River	4 Jan 2016	Landsat 8 OLI	FL	1493
2	Holmes Creek	4 Jan 2016	Landsat 8 OLI	FL	129
3	Pea River	4 Jan 2016	Landsat 8 OLI	AL	407
4	Pascagoula River	17 Mar 2011	Landsat 7 ETM+	MS	1692
5	Red River	18 Apr 1997	Landsat 5 TM	ND	7777
6	Brazos River	28 May 2016	Landsat 8 OLI	TX	74
7	Spring Creek	28 May 2016	Landsat 8 OLI	TX	160
8	Willow Creek	28 May 2016	Landsat 8 OLI	TX	48
9	Washita River	29 May 2015	Sentinel-1 VH Polarization	OK	623
10	Brazos River	30 May 2016	Sentinel-1 HV Polarization	TX	951
11	San Jacinto River	30 May 2016	Sentinel-1 HV Polarization	TX	212
12	Canadian River	17 Jun 2015	Sentinel-1 VH Polarization	OK	739
14	Trempealeau River	26 Sep 2010	Landsat 5 TM	WI	276
15	Wisconsin River	26 Sep 2010	Landsat 5 TM	WI	8350
17	Iowa River	26 Sep 2016	Landsat 8 OLI	IA	3541
18	Maquoketa River	26 Sep 2016	Landsat 8 OLI	IA	2249
19	Mississippi River	26 Sep 2016	Landsat 8 OLI	IA	1937
20	Wapsipinicon River	26 Sep 2016	Landsat 8 OLI	IA	4148
21	Minnesota River	1 Oct 2010	Landsat 5 TM	MN	7876
22	Redwood River	1 Oct 2010	Landsat 5 TM	MN	1414
23	Mississippi River	3 Oct 2010	Landsat 5 TM	MN	6957
24	Ashley River	13 Oct 2016	Landsat 8 OLI	SC	1804
25	Black River	13 Oct 2016	Landsat 8 OLI	SC	431
26	Cooper River	13 Oct 2016	Landsat 8 OLI	SC	2096
27	Lumber River	13 Oct 2016	Landsat 8 OLI	SC, NC	1116
28	Neuse River	15 Oct 2016	Landsat 8 OLI	NC	1054
29	Tar River	15 Oct 2016	Landsat 8 OLI	NC	1702
30	Rio Grande River	14 Jul 2010	Landsat 7 ETM+	TX	6836

References

Afshari, S., Tavakoly, A. A., Rajib, M. A., Zheng, X., Follum, M. L., Omranian, E., and Fekete, B. M.: Comparison of new generation low-complexity flood inundation mapping tools with a hydrodynamic model, J. Hydrol., 556, 539–556, 2018.

Blodgett, D.: nhdplusTools, Github Repository, GitHub, https://github.com/dblodgett-usgs/nhdplusTools (last access: 26 September 2019), 2018.

Clarke, K. C., Johnson, J. M., and Trainor, T.: Contemporary American cartographic research: a review and prospective, Cartogr. Geogr. Inform. Sci., 46, 196–209, 2019.

Cohen, S., Brakenridge, G. R., Kettner, A., Bates, B., Nelson, J., McDonald, R., Huang, Y., Munasinghe, D., and Zhang, J.: Estimating floodwater depths from flood inundation maps and topography, J. Am. Water Resour. Assoc., 54, 847–858, 2018.

CUASHI: NWM-Subsetting, Github Repository, Github, https://github.com/CUAHSI/nwm_subsetting, last access: 26 September 2019.

FEMA: National Elevation Dataset (NED) and similar USGS Holdings, available at: https://hazards.fema.gov/femaportal/docs/NationalElevationDataset_2007.pdf (last access: 26 September 2019), 2007.

Gochis, D. J., Dugger, A., McCreight, J., Karsten, L. R., Logan, Yu, W., Pan, L., Yates, D., Zhang, Y., Sampson, K., Cosgrove, B., Salas, F., Clark, E., Graziano, T., Maidment, D., Phan, C., Cui, Z., Liu, Y., Feng, X., and Lee, H.: Technical Description of the National Water Model Implementation of WRF-Hydro, CUAHSI Technical Report, Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI), 2016.

Gochis, D. J., Dugger, A., Barlage, M., Fitzgerald, K., Karsten, L., McAllister, M., McCreight, J., Mills, J., Rafieeinasab, A., Read, L., Sampson, K., Yates, D., and Yu, W.: The NCAR WRF-Hydro Modeling System Technical Description, available at: https://ral.ucar.edu/sites/default/files/public/WRF-HydroV5TechnicalDescription.pdf (last access: 26 September 2019), 2018.

Graziano, T., Clark, E., Cosgrove, B., and Gochis D.: Transforming NOAA Water Resources Prediction, in: AMS Meeting, Seattle, WA, 2017.

Hirabayashi, Y., Mahendran, R., Koirala, S., Konoshima, L., Yamazaki, D., Watanabe, S., Kim, H., and Kanae, S.: Global flood risk under climate change, Nature Publishing Group, 3, 816, 2013.

Johnson, J. M.: HydroData, https://doi.org/10.5281/zenodo.1401393, 2018.

Johnson, J. M.: AOI, Github Repository, https://doi.org/10.5281/zenodo.1318687, 2019a.

Johnson, J. M.: Flood Mapping, Github Repository, https://doi.org/10.5281/zenodo.2588009, 2019b.

Johnson, J. M. and Coll, J. M.: National Water Center Innovators Programs Summer Institute Report 2017, Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI), https://doi.org/10.4211/technical.20171009, 2017.

Johnson, J.M., Ruess, P., & Coll, J.: OPERA–Operational Platform for Emergency Response and Awareness: Reimagining Disaster Alerts. National water center innovators program summer institute report 2016, 4, 97, 2016.

Johnson, J. M., Coll, J. M., Ruess, P. J., and Hastings, J. T.: Challenges and Opportunities for Creating Intelligent Hazard Alerts: The "FloodHippo" Prototype, J. Am. Water Resour. Assoc., https://doi.org/10.1111/1752-1688.12645, in press, 2018.

Li, Z.: A Framework of ArcGIS-Based Flood Inundation Modeling and Mapping System, ESRI Proceedings, available at: http://proceedings.esri.com/library/userconf/proc16/papers/265_671.pdf (last access: 26 September 2019), 2016.

Lightbody, L. and Tompkins, F.: Where It Rains, It Floods: Nationwide Disasters Underscore Need for Flood Policy Overhaul, The Pew Charitable Trusts, available at: https://www.pewtrusts.org/en/research-and-analysis/articles/2018/01/25/where-it-rains-it-floods (last access: 26 September 2019), 2018.

Liu, Y. Y., Maidment, D. R., Tarboton, D. G., Zheng, X., and Wang, S.: A CyberGIS Integration and Computation Framework for High-Resolution Continental-Scale Flood Inundation Mapping, J. Am. Water Resour. Assoc., https://doi.org/10.1111/1752-1688.12660, in press, 2018.

Maidment, D. R.: Conceptual Framework for the National Flood Interoperability Experiment, J. Am. Water Resour. Assoc., 53, 245–257, https://doi.org/10.1111/1752-1688.12474, 2016.

Munasinghe, D., Cohen, S., Huang, Y. F., Tsang, Y. P., Zhang, J., and Fang, Z.: Intercomparison of Satellite Remote Sensing-Based Flood Inundation Mapping Techniques, J. Am. Water Resour. Assoc., 54, 834–846, 2018.

National Research Council: Mapping the zone: Improving flood map accuracy, The National Academies Press, Washington, D.C., https://doi.org/10.17226/12573, 2009.

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

NOAA National Weather Service: NWS Preliminary US Flood Fatality Statistics, NOAA, available at: https://www.weather.gov/arx/usflood, last access: 23 October 2019.

Rennó, C. D., Nobre, A. D., Cuartas, L. A., Soares, J. V., Hodnett, M. G., Tomasella, J., and Waterloo, M. J.: HAND, a new terrain descriptor using SRTM-DEM: Mapping terra-firme rainforest environments in Amazonia, Remote Sens. Environ., 112, 3469–3481, 2008.

Rodda, H. J.: The development and application of a flood risk model for the Czech Republic, Nat. Hazards, 36, 207–220, 2005.

Ross, N.: 2018 fasterize: Fast Polygon to Raster Conversion, R package version 1.0.0, available at: https://CRAN.R-project.org/package=fasterize (last access: 26 September 2019), 2005.

Sangwan, N. and Merwade, V.: A faster and economical approach to floodplain mapping using soil information, J. Am. Water Resour. Assoc., 51, 1286–1304, 2015.

Scaramuzza, P. and Barsi, J.: Landsat 7 scan line corrector-off gap-filled product development, in: Proceeding of Pecora, Pecora, 23–27, 2005.

Tesfa, T. K., Tarboton, D. G., Watson, D. W., Schreuders, K. A., Baker, M. E., and Wallace, R. M.: Extraction of hydrological proximity measures from DEMs using parallel processing, Environ. Model. Softw., 26, 1696–1709, 2011.

Yin, J., Gentine, P., Zhou, S., Sullivan, S. C., Wang, R., Zhang, Y., and Guo, S.: Large increase in global storm runoff extremes driven by climate and anthropogenic changes, Nat. Commun., 9, 4389, 2018.

Zheng, X., Tarboton, D. G., Maidment, D. R., Liu, Y. Y., and Passalacqua, P.: River channel geometry and rating curve estimation using height above the nearest drainage, J. Am. Water Resour. Assoc., 54, 785–806, 2018a.

Zheng, X., Maidment, D. R., Tarboton, D. G., Liu, Y. Y., and Passalacqua, P.: GeoFlood: Large-Scale Flood Inundation Mapping Based on High-Resolution Terrain Analysis, Water Resour. Res., 54, 10–13, 2018b.