

(DEM1m) was derived from 2009 aerial LiDAR data acquired by New York City Department of Environmental Protection (RACNE, 2011). This was resampled to create 3m, 10m and 30m resolution DEMs (DEM3m, DEM10m and DEM30m).

DEMs were used to delineate the watershed, calculate flow paths, slopes, drainage areas, and compute gridded values of TI. Based on TI values, the watershed was divided into 10 wetness classes (Fig. 4). Wetness class 1, covering very small fraction of the watershed (0.59%), corresponds to the perennial stream network and is the “wettest” wetness class. We grouped 50% of the watershed with the lowest TI values in the upland as the “driest” wetness class (wetness 10), because saturated areas never exceeded 50% of the watershed based on observations (Harpold et al., 2010) and predictions by other watershed models (SMR (Agnew et al., 2006), SWAT-VSA (Easton et al., 2008) and SWAT-WB (White et al., 2011)). Subsequently, we divided the remaining areas into 8 wetness classes (wetness class 2 – 9) with approximately equal areas (~ 6% each) based on TI values. Applying the same procedure of wetness class division using four DEM resolutions, four SWAT-HS setups have approximately similar areal percentage of each wetness class.

HRUs were created based on 10 wetness classes, 17 soil types, and 11 land use types. A single time series of daily precipitation and temperature data were interpolated from a 4km x 4km gridded PRISM climate dataset (Daly et al., 2008) using the inverse distance weighting method. Solar radiation data were derived as the average of airport stations at Albany and Binghamton supplied by the Northeast Regional Climate Center. Relative humidity and wind speed were generated by the built-in weather generator in SWAT. The procedure outlined above is similar to the SWAT-HS setup used by Hoang et al. (2017).

Four SWAT-HS setups were run on a daily time step from 1998 – 2012. The first 3 years were used as the warming up period and the model was calibrated and validated for the periods 2001-2007 and 2008-2012, respectively. We excluded the year 2011 from the validation period because there were two extreme events (Hurricane Irene and Tropical Storm Lee) in August 2011 that the model could not capture well. The calibration was carried out in 2 stages, i.e. snowmelt calibration and flow calibration, and by applying Monte Carlo sampling method.

Summary of Comments on hess-2018-46-manuscript-version3.pdf

Page: 8

	Number: 1	Author: rver4657	Subject: Inserted Text	Date: 30/09/2018 9:28:54 AM
	Number: 2	Author: rver4657	Subject: Inserted Text	Date: 30/09/2018 9:29:05 AM
	Number: 3	Author: rver4657	Subject: Inserted Text	Date: 30/09/2018 9:29:31 AM

1 For snowmelt calibration, we calibrated 5 snowmelt related parameters in group (i) (Table 1)
2 by generating randomly 10,000 parameter sets, running these sets using SWAT-HS,
3 comparing the streamflow predictions with observations and choosing the best parameter set
4 with the best fit to streamflow observations (highest value of daily Nash Sutcliffe Efficiency
5 (NSE)) to use for the flow calibration stage. For flow calibration, 10,000 parameter sets of 9
6 flow parameters in group (ii) (Table 1) were generated which were then run with SWAT-HS.
7 The simulations in the flow calibration stage were used for uncertainty analysis.

8 We evaluated the effect of DEM resolution on representing topographical characteristics of
9 the watershed by comparing the statistical distributions of elevation, slope angle, upslope
10 contributing area, and TI using DEMs with various spatial resolutions (1m, 3m, 10m and
11 30m). Subsequently, to evaluate the effect of DEM resolution on model uncertainty, we
12 compared the four SWAT-HS setups with different DEM resolutions based on: (i) the
13 uncertainty in streamflow predictions using “good” performance parameter sets, (ii)
14 predictions of saturated areas and their uncertainties, and (iii) uncertainty in parameter
15 estimation. We used the Generalized Likelihood Uncertainty Estimation (GLUE) approach
16 (Beven and Binley, 1992) to estimate the uncertainty in streamflow and saturated area
17 predictions caused by parameter uncertainty. For each model setup, “good” simulations were
18 identified as those with a Nash-Sutcliffe Efficiency (NSE) greater than 0.65 for use in
19 uncertainty estimation of streamflow. Subsequently, from these “good” simulations, we
20 compared predictions of saturated areas with our available field observations of saturated
21 areas to re-select the “good” parameter sets for both simulated streamflow and saturated
22 areas, to estimate the uncertainty in predicted saturated areas. Six observations of saturated
23 areas (28, 29, 30 April 2006, 12 April 2007, 7 June 2007, and 2 August 2007) are available for
24 small area in the headwaters of the Town Brook watershed.

25 2.3.2. Effect of soil and land use complexity

26 We built nine SWAT-HS setups ranging from simple (fewer soil types/land use classes/fewer
27 HRUs) to complex (more soil types/ land use classes, more HRUs) based on three soil maps

 Number: 1 Author: rver4657 Subject: Highlight Date: 30/09/2018 9:35:19 AM

So, in essence there was no real calibration, as you did not specifically optimise using a search algorithm. So while the result is the "best" from the MC runs, there could still be a better solution out there. I think this is worth a mention somewhere in the discussion.

In addition, was each DEM variation run with the same MC sets?

and three land use maps. In all nine setups, the 10m DEM was used based on its performance as the best predictor of saturated areas (see discussion).

Three soil maps were created with increasing levels of complexity (Fig. 2). The simplest map (*TBsoil_1*) had a homogenous soil type, which was created using area-weighted average soil data from the 4 dominant soil types (*Hcc*, *LhB*, *OeB*, *WmB*) in Town Brook. The second soil map *TBsoil_2* has a unique soil type for each wetness class and was created by area-weighted averaging of dominant soil properties in the corresponding wetness class. The most complex soil map *TBsoil_3* consisted of all 17 soil types.

Three land use maps with increasing levels of complexity were created (Fig. 2). The simplest land use map (*TBlanduse_1*), had agriculture as the representative land use for the watershed because it is one of the dominant land uses and potentially has a more significant impact on water quality than other land use types. The more complex land use map (*TBlanduse_2*) classifies Town Brook into 3 diverse land use types: agriculture, forest and urban areas. The most complex one (*TBlanduse_3*) contains all 11 land use types.

HRUs were generated based on a wetness map (10 classes), soil map, land use and slope maps. We assumed that slope does not have an impact on HRU discretization to simplify the set up. We also set a threshold of 1% for soil and 1% for land use to eliminate minor soil types/land uses that cover only less than 1% of the sub-basin area.

The nine model setups are categorized in 3 groups: (i) *simple*: the setups that use either the simplest soil or land use (TB1-TB5), (ii) *intermediate*: the setups that use the average complexity for maps of either soil or land use (TB6 – TB8); and (iii) *complex*: the setup that uses the most complex maps (TB9) (Table 2).

To evaluate the effect of soil and land use data complexity on model uncertainty, we compared the nine SWAT-HS setups using the same methodology used to evaluate the effect of DEM resolution on model uncertainty that is described above.

 Number: 1 Author: rver4657 Subject: Highlight Date: 30/09/2018 9:38:07 AM
See before, so all using the same MC generated parameter sets as before, or were different sets generated for each model?

differences in SWAT-HS performance on streamflow using different DEMs implying the insignificant effect of DEM resolution on streamflow simulation and the uncertainty of streamflow outputs.

① Although the effect of DEMs on streamflow prediction is minor, the setups using coarser resolution DEM10m and DEM30m are slightly better and preferred for application. These two setups give higher NSE value ranges and significantly higher mean NSE values resulted from all random combinations of parameters than the finer resolution setups. These two setups also have more “good” parameter sets indicating higher probability to get “good” representation of the modeled watershed. This implies better streamflow prediction by these two setups even without calibration.

3.1.3. Effect on the prediction of saturated areas

The probabilities of saturation in 10 wetness classes were compared among four DEM resolution setups using only “good” parameter sets for both streamflow and saturated area predictions (Fig. 6). The probability of saturation, which indicates the number of days in the calibration period when the wetness class is saturated, shows no significant difference among the four setups indicating that DEM resolution does not have an impact on the probability of saturation. It is important to note that we tried to keep the areal percentage of each wetness class approximately the same in the four setups using different DEMs. The ‘good’ parameter sets in four setups should give comparable predictions of streamflow, percentage of watershed area that is saturated, and the time that each wetness class was saturated, which results in similar probability of saturation. Wetness classes 7 to 10 are predicted to be mostly dry, implying that almost 70% of the watershed is rarely saturated. Wetness class 1 has a high probability of saturation (80-100%) because its soil water storage capacity is very low, i.e., the wetness class is prone to saturation whenever there is precipitation. The probability of saturation decreases in the more upslope wetness classes: 60-80% in wetness class 2, 30 – 50% in class 3, 5 – 22% in class 4, 1 – 9% in class 5, 0-3% in class 6, 0-1% in class 7, 0-0.3% in class 8, 0-0.08% in class 9, and 0% in class 10. We also observed that the uncertainty of saturation

 Number: 1 Author: rver4657 Subject: Highlight Date: 30/09/2018 9:42:28 AM

Are you going to explain why they are better? It has to do with the smoothing, so worth discussing.

 Number: 2 Author: rver4657 Subject: Highlight Date: 30/09/2018 9:44:06 AM

Not necessarily as it might affect the different flow components

probability of the more upslope wetness classes is lower because they only respond to high rainfall events.

The results of the probability of saturation correspond well with the uncertainty of percentage of saturated areas shown in Figure 7. The four model setups do not have significant differences in the percentage of saturated areas in the watershed. The maximum, minimum, and interquartile range indicated by the top and bottom values of the four box plots are slightly different because of minor differences in division of wetness classes in the watershed. For the majority of the time, no more than approximately 25% of the total watershed area is saturated. The watershed can be saturated up to more than 50% in extreme events that are shown as outliers in the boxplots. The median percentage of saturated areas in the watershed is only around 7-8%.

Although the statistical distributions of saturated areas in four DEM setups are relatively similar, the spatial distributions of saturated areas simulated in a small headwater area (Fig.1) on specific days (28-30 April 2006), when observations are available, appeared to be different as shown in Figure 8. In Figure 8, the saturated areas simulated in four DEM setups correspond to the saturation of wetness classes 1, 2 and 3. Saturated areas cover approximately equal areas of the watershed for the different DEM resolutions, but differ significantly in spatial distribution. The saturated areas resulting from DEM1m and DEM3m are scattered, not well connected, and broadly distributed. For coarser resolution DEM10m and DEM30m, saturated areas connect well with each other and with the areas concentrated near streams. The percentages of simulated saturated areas that intersect with observations increase with coarser resolution DEMs: 34% (DEM1m), 53% (DEM3m), 85% (DEM10m) and 90% (DEM30m). Therefore, based on visual comparison with observations and our calculation, the coarser resolution DEMs give better fits to observed saturated areas than the higher resolution DEMs. Among the four DEMs, DEM10m provides the most realistic representation of saturated areas and reasonable fit to observations.

	Number: 1	Author: rver4657	Subject: Inserted Text	Date: 30/09/2018 10:32:15 AM
---	-----------	------------------	------------------------	------------------------------

	Number: 2	Author: rver4657	Subject: Inserted Text	Date: 30/09/2018 10:32:26 AM
---	-----------	------------------	------------------------	------------------------------

	Number: 3	Author: rver4657	Subject: Highlight	Date: 30/09/2018 10:33:45 AM
---	-----------	------------------	--------------------	------------------------------

Which presumably is a result of the more connected wetness classes as a result of the smoother DEM.

	Number: 4	Author: rver4657	Subject: Highlight	Date: 30/09/2018 10:37:19 AM
---	-----------	------------------	--------------------	------------------------------

This is of course slightly qualitative and assumes that what you call "saturated area" in the "observations" is in fact spatially homogeneous, as represented by the model. Maybe the scattered saturated areas in the finer DEMs actually represent the true nature of the "saturated areas"?

Figure 12 shows the relationships of TI with slope angle, upslope contributing area and elevation using two representative DEM resolutions: 1m and 10m. It is evident that DEM 1m can capture a significantly wider range of slopes than DEM 10m because of its finer resolution. Also, the percentage of grids that have low values of TI is significantly higher in DEM 1m than in DEM 10m (Figure 12 uses red lines for reference), which also can be seen in Figure 3d. Low TI values are usually found in grids with steep slopes or with low upslope contributing areas (according to Equation 1). Because DEM 1m captures steep slopes at a local scale and has a high number of grids with low upslope contributing area (Fig. 3c), the percentage of low TI values in DEM 1m is much higher. If we look at the relationship between TI and elevation, we can see that the distribution of TI values in DEM 1m spread out wider than in DEM10m at all elevations. This explains why the distribution of TI values in DEM1m has a more complex pattern while DEM10m has a more coherent pattern with high TI grids well matched to the stream network (Fig. 13). Realistically, the highest TI value grids should be located in downslope, near-stream, low elevation areas while the lowest TI value grids should be in upslope, high elevation areas. Therefore, in this case study, the coarser DEMs (DEM10m and 30m) give a better and more realistic representation of the landscape than the finer DEMs (DEM1m and 3m). This is possibly the reason why the coarser DEMs setups have higher probabilities for good performance (i.e., a higher number of 'good' parameter sets) and have better performance in all aspects as compared with the finer DEMs.

Our findings are in agreement with Lane et al. (2004) who used a high resolution LiDAR 2m DEM with TOPMODEL, which simulates hydrology based on TI. TOPMODEL predicted the widespread existence of disconnected saturated zones that expanded within an individual storm event but which did not necessarily connect with the drainage network. They found that using the LiDAR 2m DEM, TI has a complex pattern, associated with small areas of both low and high values of the TI, leading to the appearance of disconnected saturated areas. After remapping the topographic data at progressively coarser resolutions by spatial averaging of elevations within each cell, they found that as the topographic resolution is

 Number: 1 Author: rver4657 Subject: Highlight Date: 30/09/2018 11:52:17 AM

So are you suggesting that for these finer DEM values you are actually capturing some fine scale topography, which generate locally higher wetness classes. I guess you are arguing that this level of detail is not needed for the streamflow simulation at the catchment level. So in some way you are arguing there is a relationship between scale of the modelled watershed and required accuracy of the DEM. In other words, if you would study a much smaller or much larger watershed, would your answer on the DEM accuracy be different?

 Number: 2 Author: rver4657 Subject: Highlight Date: 30/09/2018 12:22:43 PM

Only when you observe this at a certain scale! You are really discussing fractals here (in a way :-))

 Number: 3 Author: rver4657 Subject: Highlight Date: 30/09/2018 11:54:06 AM

Not sure if I would call it "better and more realistic". I think it is "more suitable to the scale of the watershed and the simulation"

coarsened, the number and extent of unconnected saturated areas were reduced and the catchments displayed more coherent patterns, with saturated areas more effectively connected to the channel network. Moreover, Quinn et al. (1995) showed how progressively refining model resolution from 50 m to 5 m reduces the kurtosis in the distribution of TI values and increases quite substantially the number of very low index values. Wolock and Price (1994) showed that hydrological predictions are affected by DEM resolutions in TOPMODEL.

Our results show that DEM10m is the best choice among four DEMs tested because of its slightly better performance for streamflow and more importantly, its good fit to observations of saturated areas. Although DEM30m also gives very good results for streamflow and distribution of saturated areas, we did not choose DEM30m because its coarse cell size may overestimate the extent of actual saturated areas. Therefore, DEM10m is the preferred choice to scale-up the application of SWAT-HS to larger watersheds in the New York City water supply system for future applications. Our choice of DEM10m is in agreement with Kuo et al. (1999) who evaluated the effect of DEM grid sizes ranging from 10-400m on runoff and soil moisture for a variable-source area hydrology model and observed that by using the 10x10m grid cells, the overall pattern of simulated wet areas showed a close correspondence with the poorly drained areas defined in the soil survey. Zhang and Montgomery (1994), in a study that evaluated grid size effect using TOPMODEL, also suggested that a 10m grid size presents a rational compromise between increasing resolution and data volume for simulating geomorphic and hydrological processes. In contrast, Thomas et al. (2017) indicated that LiDAR DEM 1-2 m is optimal for modeling hydrologically sensitive areas (runoff generating areas) and is far better than the radar based DEM 5m. However, their case study is a complex agricultural catchment dominated by micro-topographic features, which can only be captured using high resolution DEMs. Our choice of DEM10m is in contrast to Buchanan et al. (2014) who preferred DEM3m rather than DEM10m because of the better fit with the observed patterns of soil moisture collected in five different agricultural field sites.

 Number: 1 Author: rver4657 Subject: Highlight Date: 30/09/2018 11:56:15 AM
My assumption is that these watersheds are at similar scales as your watershed, not looking at micro watersheds...
I think this all needs a qualifier

 Number: 2 Author: rver4657 Subject: Highlight Date: 30/09/2018 11:56:42 AM
At the scale of the studied watershed!

 Number: 3 Author: rver4657 Subject: Highlight Date: 30/09/2018 11:57:28 AM
You might want to test this! It again depends on the scale of the watershed!

The difference in scale of case studies (field scale vs. watershed scale) and characteristics of case studies (agricultural fields vs. a mixture of forest and agriculture) between Buchanan et al. (2014) and our study may have resulted in different conclusions on choice of the appropriate DEM resolution. Therefore, the sensitivity of DEM resolution may depend on the scale and characteristics of the watershed. The dominant hydrological process in the watershed may have a big impact on the sensitivity of DEM on hydrological prediction. In our watershed, lateral flow is a dominant flow component and saturation-excess runoff is a dominant type of surface runoff, thus, topography is the most important factor. Consequently, the DEM that represents a realistic distribution of TI with high TI area compatible with the main stream network gave a better model performance. This also explains why the coarser DEM (10m and 30m) setups have higher probabilities for good performance than the finer DEMs (1m and 3m). In a field-scale watershed, finer DEM resolution is probably better because it can capture a more detailed and realistic representation of TI distribution. In an agricultural area dominated by tile drainage, DEM resolution may not be sensitive.

It should be noted here that all four DEMs in this study are derived from the same source of 2009 aerial LiDAR data with 1 meter resolution. The coarser DEMs (DEM3m, DEM10m and DEM30m) are resampled products from DEM1m. Therefore, the four different DEM resolutions carry similar information, but differ in topographic smoothing. A comparison of various resolution DEMs from different sources may not yield the same results.

4.2. What is the appropriate complexity of the distributed soil and land use inputs?

From our comparison of nine SWAT-HS setups in three groups of complexity (*simple*, *intermediate* and *complex*), we found that with all randomly generated parameter values, the *intermediate* and *complex* groups are better than the *simple* group based on slightly higher mean NSE values and a higher probability of good performance based on randomly generated parameter values. The TB3 setup, which was built from the most complex soil map (17 soil types) and the simplest land use map (1 land use) and the simplest setup TB1 are the

 Number: 1 Author: rver4657 Subject: Highlight Date: 30/09/2018 12:21:28 PM
Yes indeed, I think you could mention this much earlier and that would remove some of the repetition in the above section

and the wetness conditions of areas in the watershed are more important than land use in water quantity modeling. Moreover, SWAT-HS uses TI as the basis for hydrological modeling, thus, the effect of DEM resolution on hydrological predictions is dominant. Therefore, when the appropriate DEM resolution is used, soil and land use information become less sensitive to hydrological predictions. We think that this finding is applicable to watersheds where application of SWAT-HS is suitable, i.e., watersheds dominated by saturation-excess runoff. This finding may be also valid in applications of other topography-based watershed models including: TOPMODEL (Quinn and Beven, 1993; Beven and Kirkby, 1979), SWAT-VSA (Easton et al., 2008), SWAT-WB (White et al., 2011). These results may not be applicable in water quality modeling. Since land use information controls the inputs of nutrients and information of other human activities that affect water quality, the water quality prediction is expected to be very sensitive to the details of land use.

4.3. How does input complexity affect parameter uncertainty and model output uncertainty?

Our results show that regardless of the level of detail of input data, we obtained numerous sets of parameter values that give equally good performance for streamflow and saturated area predictions. Modifying the level of detail in input data changes the number of “good” parameter sets, but the ranges of “good” parameter values and the shape of their distributions remain the same. The number of randomly generated Monte Carlo parameter sets is sufficiently high to give a good coverage of parameter space. Although different inputs result in varied numbers of “good” parameter sets, those numbers in all setups are adequate to represent the distribution of ‘good’ parameter which reflects their sensitivities to hydrological prediction. Therefore, we conclude that for this case study and the particular model SWAT-HS, using higher resolution DEM or adding complex information on soil or land use does not reduce parameter uncertainty or solve the equifinality problem. This statement may not be valid for other areas that are characterized by numerous land uses and complex variations in

 Number: 1 Author: rver4657 Subject: Highlight Date: 30/09/2018 12:26:47 PM

I know you used "a lot" of parameter sets, but could this result be at all related to the fact that you don't strictly "optimise" in the calibration? In other words, there is no search algorithm that specifically searches for the "best solution". We can debate whether a "best solution" exists, but in your case we cannot be sure there is not possibly a single "best solution".

 Number: 2 Author: rver4657 Subject: Highlight Date: 30/09/2018 12:27:13 PM

Do you show this anywhere? Or test this?