Input data scale impacts on modeling output results: A review

Gurdeep Singh¹, Eeshan Kumar²
¹Research Engineer, The Climate Corporation, St. Louis, Missouri, United States
²Water Resources Engineer, Wood plc., Wichita, Kansas, United States

ABSTRACT. The measurement scale is one of the most important aspects of remote sensing and hydrological modeling studies. Changes pertinent to scale or resolution of input data are reflected in modeling results. This study discusses different types of scales, resampling and rescaling techniques and several potential scale issues as addressed by past studies. Specifically, scale impacts of digital elevation model (DEM) and land use and land cover (LULC) data impacts on hydrological modeling results are discussed. Higher resolution data can accurately describe watershed characteristics but can be computationally intensive. Therefore, it becomes important to select a data resolution, which aides in model running efficiency by not compromising with model accuracy. The discussions in this paper could provide references for development, application, and improvement of watershed hydrological models, and increase accuracy and efficiency of hydrological model simulations and can aid in making informed decisions by watershed managers.

Keywords. Scale, watershed, digital elevation model, hydrological modeling, soil and water assessment tool (SWAT)
1. Introduction

The use of watershed modeling has increased in the recent years with the dramatic increase in computing powers (Borah and Bera, 2004). Sustainable use of water resources is important to meet various water demands of the world, and hydrological models can prove to be great tools for planning purposes (Abbaspour et al., 2015). Watershed models are mainly used for best management practices analysis, prediction of non-point source nutrient loads, analyze total maximum daily loads (TMDL) and management of water resources (Singh and Saraswat, 2016; Pai et al., 2011). The rise in complexity among watershed models has contributed to their usefulness in handling this task (Fernandez et al., 2000). However, as the complexity of model increases, it becomes pertinent that input data selected should accurately describe the characteristics of a watershed, but does not compromise with its running efficiency (Kumar, 2015).

Our ability to model hydrologic processes with greater accuracy and at a finer spatial and temporal resolution will continue to improve with increased use of remotely sensed data (e.g., satellite observations), increased computational capacity, improvements in GIS and database management systems. However, computational capacity, data availability and model complexity will not increase at the same rate. Considering this, there could be two potential errors in modeling: (1) developing an overly complex model that cannot be calibrated properly and verified using available data, or (2) developing a model that fails to make proper use of available, high-quality data (Mirchi et al., 2010).

This paper focuses on one of the important aspects of remote sensing and hydrological modeling-scale or resolution of input data. Here four major sections: (1) scale in hydrological modeling, (2) DEM resolution effects on hydrological modeling, (3) LULC resolution effects on hydrological modeling; and (4) other input data resolution effects on hydrological modeling are discussed which provide a detailed account of input data effects on watershed modeling. These sections are followed by the summary and conclusion section with a provision of future directions for modelers.

2. Scale in hydrological modeling

In context to analysis, modeling, and perspective scales are broadly categorized in six following categories (Cao et al., 1997; Bloschl et al., 1995; Lam et al., 1992; Bierkens et al., 2000, and Wu and Li, 2009):

1) Observation Scale: Observation scale or measurement scale depends on the methods or characteristics of the instrument. This scale refers to the description of the resolution, time interval, spectral range or solid angle in the field of remote sensing.

2) Modeling Scale: Models are built so that reliable outputs could be generated at the scale. The influence of measurement and operational scale is seen on the modeling scale. Observations that are sampled at measurement scale are used as inputs for a model. The measurement scale should coincide
with the modeling scale, and if there is any difference between them, then proper scaling conversions are required. The modeling scale should also coincide with the operational scale.

3) Operational Scale: Any process is operated at the scale. If the operational scale is smaller than the scale used for modeling then the variability, lower than the modeling scale that was used is lost and the phenomenon that is expected could not be seen.

4) Geographic Scale: Also called for coverage, the geographic scale refers to the spatial extent of research. A large geographic scale study may involve a larger spatial area, whereas a smaller scale corresponds to a smaller area.

5) Policy Scale: It is the scale at which decisions are made, or policy could be implemented. A policy scale can be used to judge crop yield of a specific area to see if it is reduced or increased on a yearly basis. For good inferences, policy scales should be larger than the operational scales.

6) Cartographic Scales: It is the ratio of the distance on the ground and the map. This is used for the representation of spatial results. A scale, 1:10,000 corresponds a larger area in comparison to scale 1:5000. The scale of 1:10,000 may not provide enough details as this would cover a larger area on the ground.

For modeling, observation and modeling scale should be smaller than operation scale. Also, observation and modeling scale should be consistent with each other. The changes of resolution may also alter the modeling results. For example, while dealing with the land use/land cover rasters, if the resolution is decreased then it might give rise to a possibility that one pixel of the raster could contain more than one cover type. This could disturb the heterogeneity which is considered as the fundamental characteristic of any landscape (Wu et al., 2000).

If any higher resolution data are processed to decrease its resolution, then, although the storage space required by the same could be reduced, and it could respond quickly to process and display, but it would have lower feature spatial accuracy, and one could lose on the prediction of good results in modeling (Burrough and Pfeffer 2003). Thus, converting high-resolution data to lower resolution results in a reduction of file size, decrease of processing time faster displays.

Xiangyi et al. (2016) also analyzed how the resolution of the DEM and LULC input layers impacted a hydrologic model output. They concluded that too high and too low resolution of data impacted simulation results. Improvement of DEM resolution did not improve the simulation accuracy. An increase in the resolution of land-use data also resulted in a larger number of HRUs in the SWAT model that impacted model efficiency.

3. DEM resolution impacts on hydrological modeling

One of the important factors in determining watershed response to runoff and water quality parameters is watershed topography. All hydrological models require a digital elevation model (DEM) as one of the spatial inputs. For example, the SWAT model uses a DEM to delineate a watershed and further divide the watershed into HRUs (hydrologic response units) based on the land use and soils data.
Input DEM data resolution affects the watershed that is being delineated, a number of subwatersheds and HRUs were created (Chaubey et al., 2005). The effects of DEM resolutions on the SWAT model outputs in the Moores Creek Watershed, Arkansas was analyzed by Chaubey et al. (2005). They analyzed a total of seven different DEM resolutions, namely, 30m, 100m, 150m, 200m, 300m, 500m, and 1000m. The modeling period was 1997 to 2000, and the objective function optimized to be a relative error (RE). Initially, the SWAT model was developed and calibrated on an annual scale for flow, nitrate-nitrogen, and total phosphorus using the 30*30m DEM. The calibrated model outputs were considered as the baseline for comparing with the outputs generated by models using different DEM resolutions. All DEMs (except 30*30m) were resampled from 30*30m DEM using the bilinear interpolation method. Differences in DEM resolutions affected the watershed delineation, stream network, and subbasin classification. Specifically, DEM resolutions resulted in the slightly different watershed area and some sub-basins and HRUs. The watershed and stream network representation became increasingly less accurate with a decrease in DEM resolution. The range of RE for stream flow was 4–25% and for NO3-N was 10–51%, respectively. Coarser DEMs resulted in decreased flow and nitrate-nitrogen. However, total phosphorus did not always decrease. As per the results stated in this study, it can be said that models should be developed with finer resolution DEMs as opposed to coarser DEMs especially when the water quality outputs need to be analyzed.

Panda and Proctor (2008) analyzed the water quality and quantity of West Fork Little River (WFLR) watershed in Georgia using one arc second (30 m), 1/3 arc second (10m), and 1/9 arc second (3m) DEM with SWAT. They obtained higher organic N and P output with finer spatial resolution DEMs. They concluded that low-resolution DEM showed flatter spatial area as an average, ignoring the actual elevation in places and ultra-high-resolution DEM provides actual elevation and thus resulting in an increased runoff in the study area, which was sloppy as being the foothill of the Appalachian mountain ranges (Panda and Proctor, 2008). They also concluded that low-resolution DEMs might provide similar results compared with high-resolution DEM in flatter coastal areas, but would provide different model outputs in mountainous undulated watersheds.

Cho and Lee (2001) studied the effect on the runoff of the SWAT model output in the Broadhead watershed; New Jersey is using two different DEMs of resolution of 30 m and 90 m. They found that the use of the 90-m resolution DEM resulted in the underestimation of the total runoffs whereas with the use of 30 m resolution DEM higher simulated runoff values were predicted. They stated that this increase could be due to the finer resolution.

Cotter et al. (2003) analyzed the effect of the different resolutions (30m, 100m, 150m, 150m, 200m, 300m, 500m, and 1000m) on the SWAT model outputs for flow, sediments, NO3-N and the TP in the Moores Creek watershed, Arkansas. They reported that the model output was most sensitive to the DEM resolution, flow predictions were not significantly affected by the resolution of the land use data and the resolution of the soil data did not have any significant effect on NO3-N and TP predictions.
It could be said that the 30m resolution DEM data produce adequate simulation results as per the past studies (Chaubey et al., 20005 and Cotter et al. 2003) and a DEM of 30 m resolution could be used, but a LULC layer resolution of 28.5 m could be used. If the resolution of LULC is changed to 5 m then it could disturb the heterogeneity of the landscape as discussed before (Wu et al., 2000). As per the metadata information about the DEM data available from http://www.geostor.arkansas.gov/G6/Home.html the resolution of the data is 5m. The resolution of DEM could be resampled to 30m, but LULC should not be done.

Input data resolution affects the parameterization of the model output. As a result, Cotter et al. (2003) have analyzed the effect of spatial resolution variability of input data on model output uncertainty. The study area was the Moores Creek watershed (1890 ha) – a subbasin of the Illinois River watershed – in Washington, Arkansas. Pasture (55%) and mixed forests (39%) were the major land uses in the watershed. Various spatial resolutions of DEM, land use, and soil were considered: 30 x 30 m, 100 x 100 m, 150 x 150 m, 200 x 200 m, 300 x 300 m, 500 x 500 m, and 1,000 x 1,000 m. The impact of varying spatial resolutions was analyzed on the uncertainty of SWAT predicted flow, sediment, nitrate-nitrogen, and total phosphorus. The model was calibrated from 1997 to 1998 on an annual and monthly scale, and RNS2 (Nash-Sutcliffe Efficiency) was the objective function selected. The calibrated annual flow was overpredicted by 11% in 1997 and underpredicted by 13% in 1998. The monthly flow calibration yielded RNS2 value of 0.76 and was comparable to other reported values for SWAT calibration. The calibration yielded RNS2 values of 0.48, 0.44, and 0.66 for sediment, NO3-N, and TP, respectively. Courser DEM resulted in decreased watershed area representation and slope, but increased slope length. Flow predictions were not significantly affected by land-use data resolutions. Courser land-use affects the distribution of pasture, forest, and urban areas, resulting in significant uncertainty in predicting sediment, nitrate-nitrogen, and total phosphorus. The soils data resolution does not affect flow and nitrate-nitrogen. However, sediment was overpredicted, and total phosphorus was underpredicted.

Decreases in land use and soil resolution resulted in smaller predicted flow and water quality parameter uncertainty than did decreases in DEM resolution. Overall, it was recommended that the minimum DEM data resolution should range from 30 to 300m whereas minimum land use and soil data should range from 300 to 500m for achieving less than 10% model output error for flow, sediment, nitrate-nitrogen, and total phosphorus.

The DEM, land use and soil data resolutions should be kept under the limits mentioned above for achieving less than 10% model output error for variables of interest. Alternatively, finer resolution data should be preferred as compared to coarser datasets. Improvement of DEM resolution does not always result in the improvement of simulation accuracy after the resolution reaching certain accuracy, while the decrease of slope degree caused by the decrease of DEM resolution would decrease the flow production and delay the summit of peak flow (Xiangyi et al., 2016).
4. Land Use and Land Cover (LULC) resolution impacts on hydrological modeling

Land use and land cover changes in any watershed over the years. A single land-use layer or multiple land-use layers could be used to set up a hydrological model as per the availability of data (Kumar, 2015). Similarly, there can be one or multiple land-use sources. For instance, National Land Cover Database (NLCD) in combination with Cropland Data Layer (CDL) was used by Singh et al. (2017) whereas NLCD in combination with Center for Advanced Spatial Technologies (CAST) was used by Singh (2012) and Singh (2015). Several studies have analyzed the impact of using different resolutions of LULC layers. Sometimes map comparisons necessitate reprojecting and rescaling raster based categorical maps obtained from multiple sources.

Christman and Rogan (2010) have analyzed various methods of raster data transformation for land change analyses. Each transformation resulted in differences in the relative area of classes. Even between the two maps which underwent the same transformation operations in opposite order (2 and 3), 9.8% of the pixels were different. The relative class area was best preserved by a nearest-neighbor resampling method. However, the contiguity of thematic classes and the overall fragmentation of the landscape was lowest with the vector-based method.

The relative area and position of various land use classes should be cross-checked with the actual land use map after reprojecting/resampling. Frequency-based aggregation method was suggested by the authors for transforming raster datasets. Land cover maps are often used for mapping change detections in land uses. However, the land-use images might not be available at the same resolutions. Hubert et al. (2012) analyzed and explained what approach should be used for existing upscale classifications in the land use maps. The study area comprised of three regions: the communes of Geneve (urban dominated), Thonex (urban and agriculture dominated), and Satigny (agriculture dominated) in Switzerland. Four methods were used to test the degradation of a higher resolution map (0.25m) to a lower resolution (5m): direct method, nearest-neighbor, majority method, and statistical method. Upscaled maps were compared to the original by visual examination and computing kappa coefficients.

The majority and statistical methods were unable to preserve thin linear features such as roads. As far as statistical analysis is concerned, all methods except direct method performed well. For the upscaling based on the high thematic resolution, between 8% (Satigny), 19% (Genève Cité) and 22% (Thônex) of the information was lost. The nearest-neighbor approach yields the smallest loss of information with Kappa values up to 92%.

The nearest-neighbor approach should be preferred for resampling raster datasets. However, a preliminary analysis can be done with all the methods mentioned above to see if any other method yields more accurate results.

As Land Use Land Cover (LULC) maps are categorical geospatial data layers linking pixel based spectral reflectance and corresponding ground truth information, errors are associated with each LULC
category. Pai and Saraswat (2013) evaluated the effect of published LULC categorical errors on SWAT model output uncertainty. In other words, how much uncertainty can be expected in the SWAT model output from LULC categorical errors?

A new method was also used in this study to distribute land use while maintaining the spatial autocorrelation structure. The SWAT model was run ten times from 2000 to 2006 with errors ranging from ±8% FRST (forest), ±28% BERM (bermuda grass), ±19% FESC (fescue), ±22% URLD (urban-low intensity), and ±9% URHD (urban-high intensity). Variation in water yield was observed as a result of LULC categorical errors at the subwatershed scale.

Three factors that affected the uncertainty due to LULC categorical errors were LULC type, baseline LULC composition, and the percentage misclassification error. Water yield deviations had a range of 0% to 1.6%, 0% to 0.2%, 0.7% to 3.2%, 0.2% to 8.0%, and 0.3% to 2.8% for the URLD, URHD, FRST, BERM, and FESC land uses, respectively. Deviations in monthly output had a range of 0.0% to 4.8%, 0.0% to 3.4%, 0.1% to 6.8%, 0.0% to 19.9%, and 0.0% to 5.4% for the URLD, URHD, FRST, BERM, and FESC land uses, respectively. Overall, LULC categorical errors resulted in variations in water yield ranging from 0% to 8% annually and 0% to 19.9% monthly.

5. Other input data resolution impacts on hydrological modeling

The spatially distributed precipitation data could increase the accuracy of surface runoff simulation because the distributed precipitation data could embody the locally intensified rainfall events that could significantly influence the surface runoff. The results could provide references for the development, application, and improvement of the watershed hydrological model and increase the accuracy of the hydrological model simulation. Effects of resolution of spatial land use data and soil data on the simulation results of the model mainly result from their impact on the generation of HRUs (Xiangyi et al., 2016). Soil data are an essential input to SWAT model. It is available in two forms of spatial resolution, i.e., State Soil Geographic (STATSGO) and Soil Survey Geographic (SSURGO) and the latter being higher spatial resolution and more detailed. Panda and Proctor (2008) completed SWAT models to analyze the water quality and quantity of WFLR, a mountainous watershed in north Georgia, using both STATSGO and SSURGO data in combination with NLCD 30 m LULC and 10 m DEM. They obtained higher sediment, organic N, organic P, and NH4 output and lower NO3, NO2, and Chl-a output with finer spatial resolution SSURGO data combination (Panda & Proctor, 2008). It was clear from their study that detailed soil data would provide different water quality and quantity result in comparison to low-resolution STATSGO soil data.

6. Summary & Conclusion

The simulation accuracy of a watershed hydrological model depends largely on the description of the watershed characteristics of the input data. High-resolution input data can accurately describe the watershed characteristics, and at the same time increase the difficulty of data collection and processing.
As a result, it is important to study the response of the model simulation results to spatial data resolution, which would help to improve model running efficiency without reducing model simulation accuracy. As per discussions in this study, the choice of input data resolution for a watershed model depends on the output of interest. Since the use of models in making watershed response predictions can be expected to increase in future, every effort must be made to collect input data at a finer resolution to minimize uncertainties in the model predictions. Future climate scenarios can also be analyzed using robust models which can help in risk assessment of floods and droughts. Hence, the results of such models can be used by watershed managers to make more informed decisions and aid in better watershed management.

7. References
Cao, C.Y., and N. Lam. (1997). Understanding the scale and resolution effects in remote sensing and GIS. In Scale in remote sensing and GIS, Lewis Publishers: Boca Raton, FL, USA. 57-72


