NextGen Regionalization

Documentation

2/7/2024

Yuqiong Liu

1.Purpose of regionalization

Given that reliable streamflow observations are only available at a limited set of gaged locations within CONUS and oCONUS domains, NextGen formulations (modules) can only be calibrated for a limited portion of the entire model domain. In order to run NextGen with mosaic formulations and parameters over the CONUS and oCONUS areas, proper formulation and parameters need to be prescribed for each catchment or grid cell in the model domain. This can be achieved through regionalization, where an appropriate donor (i.e., a calibrated basin) is identified for each catchment or grid cell outside of the calibrated basins and the best-performing formulations along with their optimal parameters from the chosen donors are applied to the uncalibrated areas (i.e., the receivers).

2.Background: NWM regionalization

The approach to identifying donor basins for the uncalibrated areas is the most critical part of any parameter regionalization process for large-scale hydrological models. For the National Water Model (NWM), the regionalization approach has evolved from simple techniques based on the Level III and Level IV Ecoregions (v1.0 and v1.1) to more complex techniques based on a combination of physical similarity, hydrologic similarity (i.e., streamflow signatures) and spatial proximity (v3.0), upon which the NextGen regionalization framework is built.

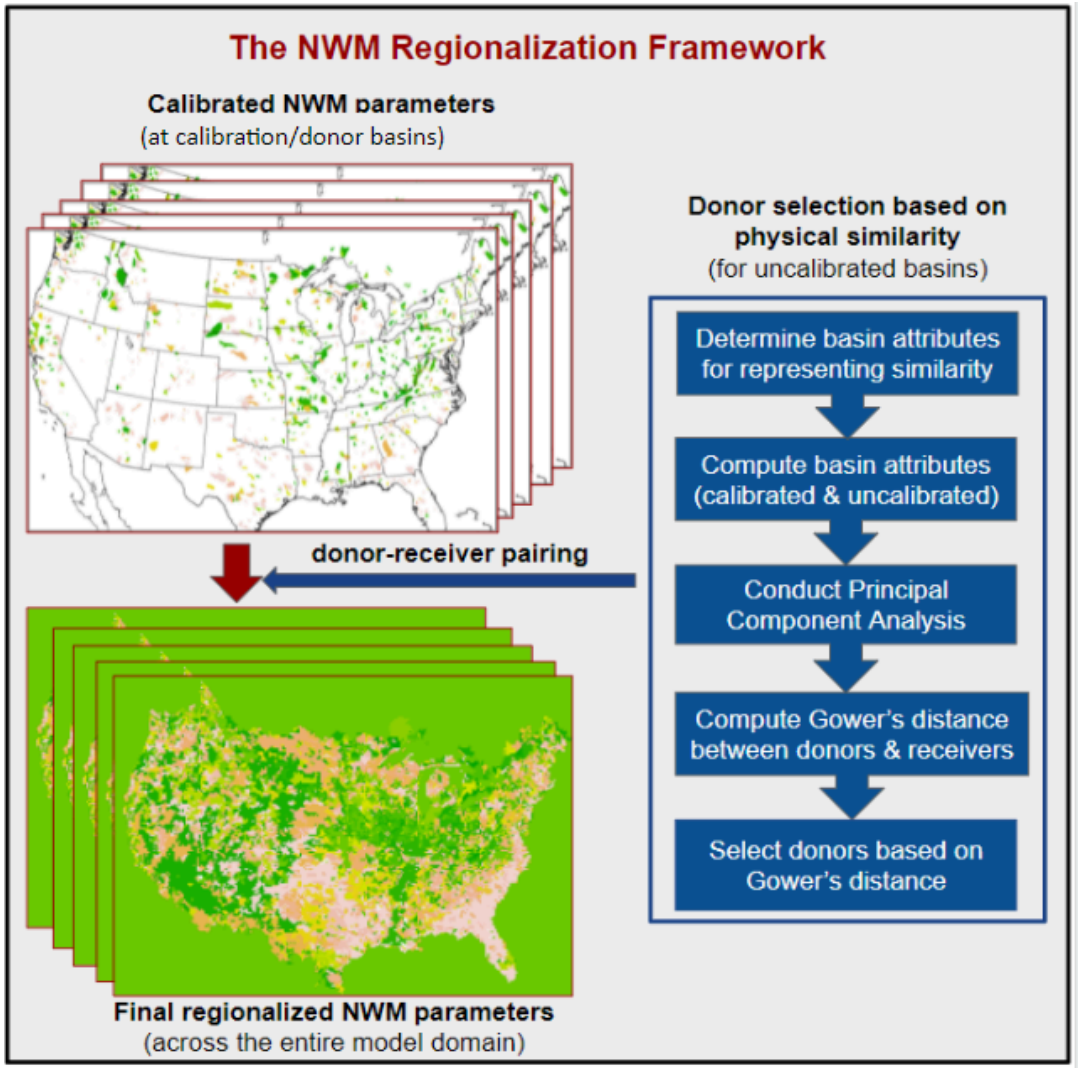

Figure 1 The regionalization framework for the operational NWM v3.0.

3. Regionalization in the context of NextGen

A key difference in regionalization between the NextGen-based NWM v4 and the previous versions of the NWM is that, with NextGen, we'll be regionalizing not only calibrated parameters, but also the calibrated formulations. Figure 2 illustrates regionalization in the context of NextGen with a 5-catchment basin, with the blue catchment calibrated for CFE and the pink catchment calibrated for Topmodel. The three uncalibrated catchments (gray) receive donations of formulation (along with calibrated parameters) from either the blue or the pink catchment, as determined by regionalization.

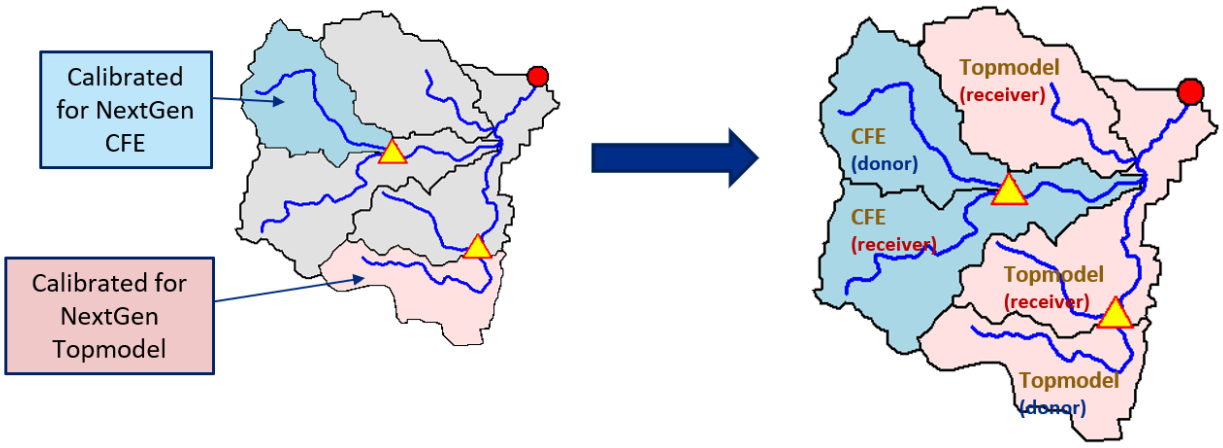

Figure 2. Illustration of NextGen regionalization

4. Regionalization workflow

As illustrated in Figure 3, NextGen regionalization takes as inputs the various static and dynamic datasets made available through the NextGen framework, computes the catchment attributes related to topography, climate, land cover, soil geology, and streamflow signatures, access the physical similarity and spatial proximity between the donor and receiver catchments, and produces outputs in the form of donor-receiver pairing, which can be used to transfer calibrated formulations and parameters from donors to receivers.

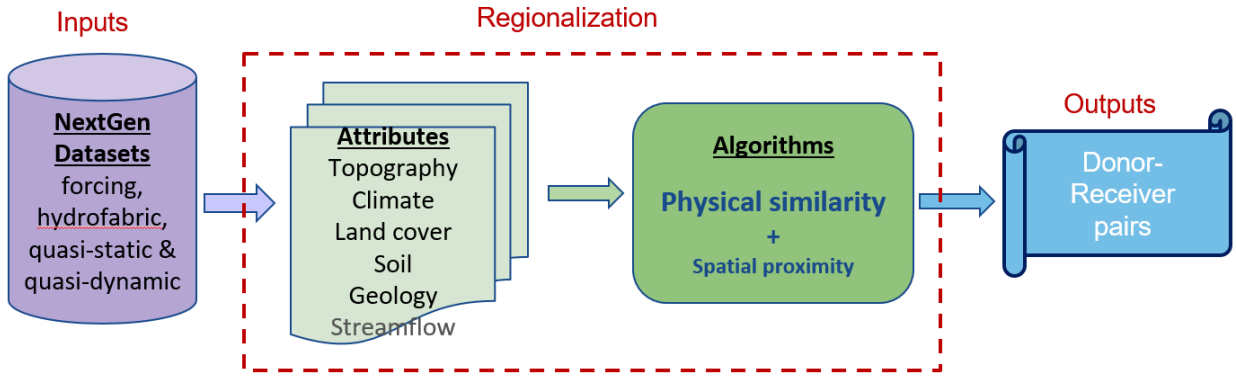

Figure 3. Workflow for NextGen regionalization

5. Regionalization in the NextGen diagram

Figure 4 depicts the envisioned interactions between the regionalization component and the other components of NextGen, which can be summarized as follows: a user uses Runtime

Environment (1) to set up regionalization applications (for selection of domain, attributes, calibrated donor basins, regionalization method and hyperparameters etc); catchment attributes are extracted from the hydrofabric (Component 3) or computed by the "catchment attribute calculator" (based on datasets in 3) as needed; donor-receiver pairing outputs from the regionalization component (11) are passed on to the formulation/parameter component (4) for simulation and evaluation with component (9).

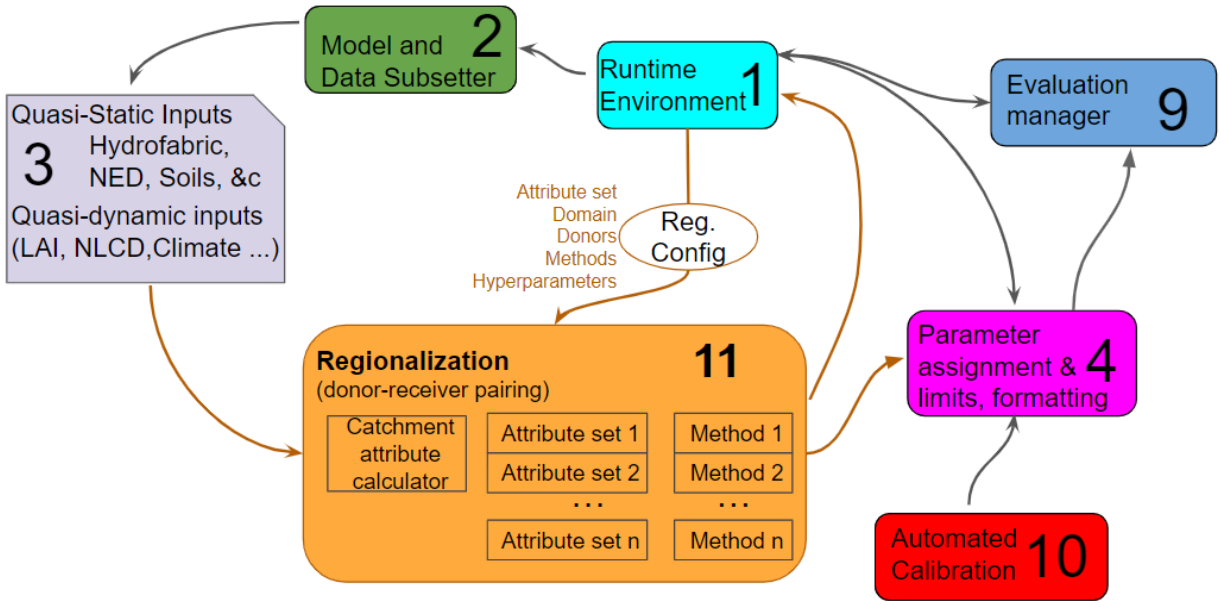

Figure 4. Envisioned interactions between the regionalization component (Box 11) with other components of the NextGen framework.

6. Inputs to NextGen regionalization

There types of inputs are needed for the regionalization algorithm:

1) *Donor and receiver hydrofabric files*. The current regionalization tool is designed for catchment-based NextGen, where the shapefiles (polygons) of donor and receiver catchments are used to calculate the mean-area catchment attributes. The tool allows the donors and receivers to have different versions of hydrofabric. Other information is also needed to determine the upstream catchments of a given gage. To extend to gridded NextGen (e.g., the NWMv3 baseline and NWMv4), donor-receiver mapping can be first produced in the catchment/basin space and then mapped to gridded parameters (as in NWMv3).

2) *Catchment datasets*. Various datasets are needed to calculate catchment attributes to represent donor-receiver similarities for regionalization. Datasets used for the AGU 2023 MVP include the Analysis of Record for Calibration (AORC, Water Year 2008-2019), Subtle Radar Topography Mission (SRTM 30m), National Land Cover Database (NLCD 2019, 30m), the Digital General Soil Map of the United States (STATSGO2), and the Global Hydrogeology Maps dataset (GLHYMPS 2.0). These are used to calculate attributes of climate, topograph, land cover, soil, and geology, respectively. The HYSOGs250m dataset is also used to calculate the hydrologic soil group. Note the streamflow signatures (to be calculated from USGS streamflow observations) have not been used in NextGen regionalization but can be easily incorporated based on scripts used in NWMv3 where streamflow signatures were actually used in the regionalization process.

3) *Calibrated donor formulations and parameters*. Prior to regionalization, the donor basins should be properly calibrated for chosen formulations. The calibrated formulations (along with their optimal parameters) will then be transferred to the uncalibrated areas based on the donor-receiver pairings to be produced by the algorithm. When multiple formulations are calibrated at donor basins, the optimal formulation (determined from validation statistics) can be chosen for each donor and all of its receivers to construct an optimal mosaic implementation of NextGen over the model domain.

7. Calculation of basin attributes

Currently the catchment attributes are developed based mainly on two conceptual frameworks: Hydrologic Landscape Regions (HLR, Wolock et al. 2004) and Catchment Attributes and Meteorology for Large-Sample studies (CAMELS, Anddor et al. 2017). Other attributes, e.g., related to stream networks or anthropogenic influences, can be added as needed in the future. Note NWM v2.1 adopted a modified version of the HLR framework.

Table 1 lists all the attributes calculated based on the HLR and CAMELS frameworks. These attributes are grouped into 5 broad categories: climate, topography, land cover, soil and geology. Note streamflow signatures of the CAMELS framework, along with a few other attributes are not included yet but should be investigated in future work. In addition, the land cover attributes are not included in the original HLR framework (Wolock et al., 2004) but are added here, while the bedrock permeability attribute of the original HLR framework is omitted here due to lack of reliable data sources.

The original datasets for calculating each attribute are also listed in Table 1. These are either raster or vector datasets. Given the shapefiles defined for the donors and receivers in the hydrofabric files, the zonal package in the R-based Hydrofabric tools is used to calculate the values of each attribute for all donors and receivers, using the “mean” function for all attributes

except for soil_conductivity, where the geometric mean function (gm_mean) is used. If the data source for calculating an attribute has spatial gaps, the data coverage for the attribute is also calculated to help determine whether the calculated attribute is valid. Currently, the minimum coverage (configurable) is set to 50%. In other words, the data coverage for a given catchment needs to be at least 50% for the attribute to be considered valid for use in regionalization.

As an illustration, Figure 5 presents the catchment attributes calculated for HUC-01 receivers for a few selected attributes: aridity, snow fraction, sand fraction, and slope.

Table 1 Catchment attributes for HLR and CAMELS used for regionalization in AGU 2022 MVP

Category	Attribute	Description	Unit	Framework	Data Source
Climate	p_mean	mean daily precipitation	mm/day	CAMELS	AORC
	pet_mean	mean daily PET based on Hargreaves method	mm/day	CAMELS	
	aridity	PET/P (i.e., p_mean/pet_mean)	-	HLR & CAMELS	
	snow_frac	fraction of precip. falling as snow (i.e., on days colder than 0 degC)	-	CAMELS	
	high_prec_freq	frequency of high precipitation days (>5 times mean daily precip.)	days/yr	CAMELS	
	high_prec_dur	average duration (consecutive days) of high precipitation events	days	CAMELS	
	low_prec_freq	frequency of low precipitation days (< 1 mm/day)	days/yr	CAMELS	
	low_prec_dur	average duration (consecutive days) of low precipitation events	days	CAMELS	
Topography	elev_mean	catchment mean elevation	m	CAMELS	SRTM digital elevation
	slope_mean	catchment mean slope	m/km	CAMELS	
	prcFlatTotal	total fraction of flat land (slope < 0.01)	-	HLR	
	PrcFlatLowland	fraction of flat land in upland areas (below middle elevation)	-	HLR	
	prcFlatUpland	fraction of flat land in upland areas (above middle elevation)	-	HLR	
	relief	Difference between highest and lowest elevations	m	HLR	
	cidx	circularity index	-	HLR	Hydofabric
Landcover	forest_frac	fraction of foreast	-	HLR & CAMELS	NLCD 2019
	cropland_frac	fraction of cropland	-	HLR	NLCD 2019
	urban_frac	fraction of urban area	-	HLR	NLCD 2019
	gvf_max	max monthly mean green vegetation fraction (based on 12 monthly means)	-	CAMELS	MODIS
	gvf_diff	difference between max & min monthly mean GVF	-	CAMELS	
	lai_max	max monthly mean leaf area index (based on 12 monthly means)	-	CAMELS	
	lai_diff	difference between max & min monthly mean of lead area index	-	CAMELS	
Soil	sand_frac	fraction of sand in the soil	-	HLR & CAMELS	STATSGO
	clay_frac	fraction of clay in the soil	-	HLR & CAMELS	
	soil_porosity	volumetric porosity	-	CAMELS	
	soil_conductivity	satuated soil conductivity	cm/h	CAMELS	
	soil_depth	depth to bebrock	m	HLR & CAMELS	Pelletier
Geology	geol_porosity	subsurface porosity	-	CAMELS	GLHYMPS
	geol_permeability	subsurface permeability (log10)	m^2	CAMELS	

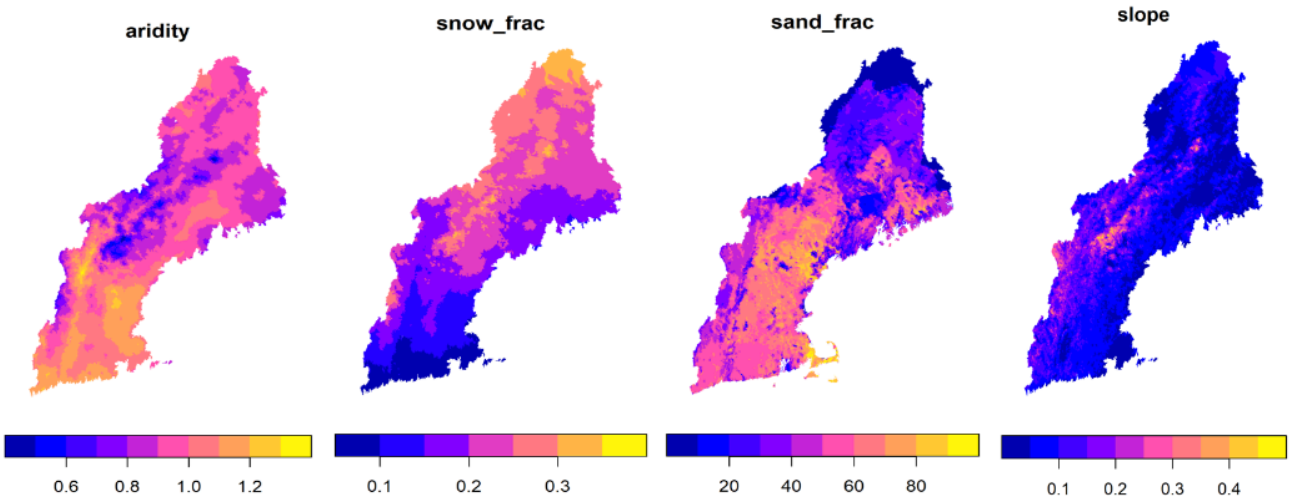

Figure 5 Spatial distribution of aridity, snow fraction, sand fraction, and slope over HUC-01

8. NextGen regionalization algorithms

So far seven different regionalization algorithms have been developed for NextGen (Fig. 6). These include 1) spatial proximity, 2) unsupervised random forest (URF), 3) Gower's distance (Gower, 1971), 4) k-means clustering, 5) k-medoids clustering, 6) balanced iterative reducing and clustering with hierarchy (BIRCH), and 7) hierarchical density-based spatial clustering of applications with noise (HDBSCAN). The first algorithm is based purely on spatial proximity (where each receiver is paired with the spatially closest donor) and can be used as a benchmark for assessing the other more sophisticated algorithms that are primarily based on physical proximity (but supplemented with spatial proximity when needed). Note the Gower's approach is also used in the operational NWM (v3.0). Each of these algorithms comes with its own unique strengths and weaknesses and so it is critical to choose an appropriate algorithm for a given application. Ultimately though, assessing the performance of these regionalization algorithms would require running simulations with regionalized formulations and parameters to evaluate the skill in predictions of streamflow and other relevant hydrologic variables.

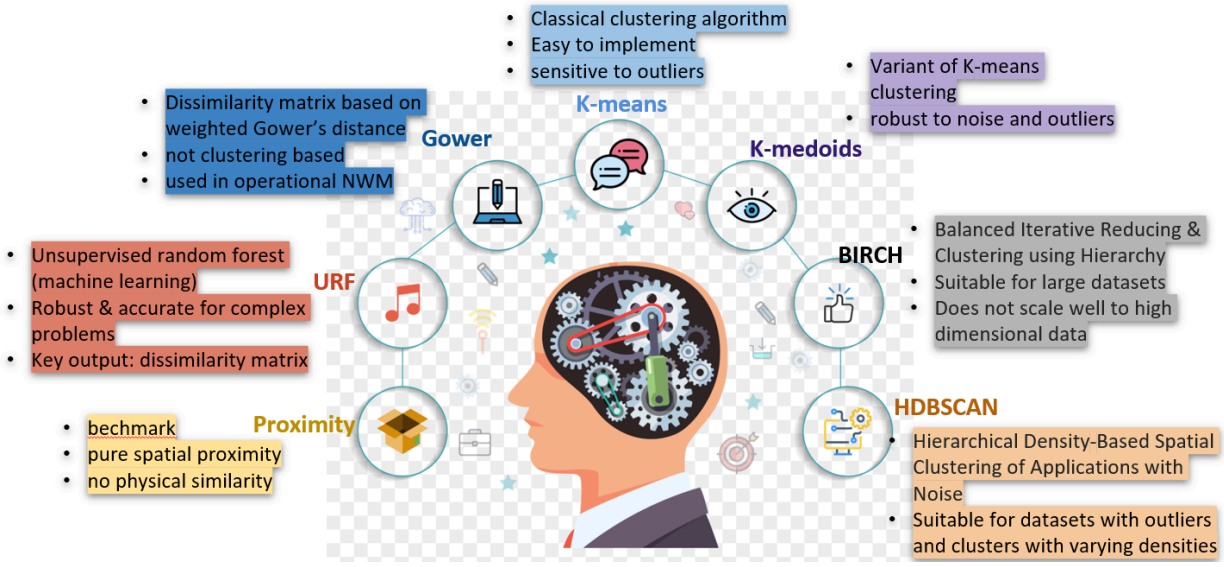

Figure 6. The seven algorithms accurately available for NextGen regionalization

Below are some important notes about the implementation of NextGen regionalization algorithms:

1) The six physical proximity based algorithms can be grouped into two categories: 1) dissimilarity-base approaches (i.e., URF and Gower) where a physical dissimilarity matrix is first computed for all the receivers and potential donors, and then the dissimilarity matrix is used to assign a donor to each receiver based on the smallest dissimilarity principle; 2) clustering-based approaches (i.e., k-means, k-medoids, BIRTH, HDBSCAN) where clustering is first used to group all receivers and donors into separate clusters and then donors are assigned based on the shared cluster membership and spatial proximity.

2) Because many of the catchment attributes are correlated with each other and hence collectively contain redundant information, it is important to remove that redundant information before the attributes can be used to compute physical similarity between donors and receivers. This is achieved through principal component analysis (PCA), which reduces the original attributes to a few principal components that are largely independent of one another. The scores of these principal components are then fed into the regionalization algorithms for donor-receiver pairing. This preprocessing step is critical for all the physical similarity-based methods except for the URF, which can effectively learn the relationships between donors and receivers even in the presence of redundant information.

3) The development of the clustering-based methods (k-means, k-medoids, BIRCH, HDBSCAN) leverages the python package scikit-learn, with online documentation about each of the clustering method and the comparisons between different clustering methods available at https://scikit-learn.org/stable/modules/clustering.html. Information on URF-based clustering can be found at http://gradientdescending.com/unsupervised-random-forest-example/. The calculation of Gower's distance is detailed in Section 9 below.

9. Compute Gower's distance between donor and receiver basins

The Gower's distance metric (Gower, 1971) can be used to calculate the distance or dissimilarity between two objects characterized by multiple variables (numerical or categorical), where weights can be applied in the calculation to reflect the relative importance of these variables. For two objects represented by the same N variables, the Gower's distance can be calculated as follows:

$$S = \frac{\sum_{k=1}^{N} W_k S_k}{\sum_{k=1}^{N} W_k} \quad (1)$$

Where W_k is the weight on variable k and S_k is the distance for variable k, respectively. For numerical variables, S_k is the absolute difference between the two objects in the values of variable k, normalized by the range of the values of variable k over all observations. For categorical variables, S_k is assigned to 1 if the two objects are equal on variable k and 0 if they are not.

The variables for calculating the Gower's distance are the scores of the principal components from the PCA analysis. The weights W_k are calculated based on the percentages of the total variance explained by individual principal components.

10. Identify donors based on dissimilarity matrix (URF and Gower)

The following steps are followed to identify a donor for each receiver catchment when the dissimilarity-based approach (URF or Gower) is used:

1) Compute the centroid-based spatial distances between each receiver and all available donors within the model domain.

2) In the first round of processing, for a chosen framework (HLR or CAMELS), perform a PCA using all attributes of all donors and receivers to obtain the principal components as well as the percent variance explained by each component.
3) Compute the dissimilarity matrix between all receivers and donors using URF or Gower's distance. For the latter, the weighted Gower's distance is calculated using the scores of the principal components and the weights computed from the explained variances.
4) For each receiver, start with a small search radius (e.g., 200 km) and iteratively increase the search radius until at least one qualified donor is found. Here a donor is considered qualified if the Gower's distance between the donor and the receiver is no greater than a predefined threshold.
5) If multiple qualified donors are found, calculate the differences in a few primary screening attributes between the donors and the receiver and keep only donors with differences smaller than predefined values, while making sure at least one donor is retained. Currently, the screening attributes include elev_mean, snow_frac, cropland_frac, urban_frac, and forest_frac.
6) If multiple donors are left, further narrow down by keeping only donors that are within the same hydrologic soil group as the receiver, while making sure at least one donor is retained.
7) If still multiple donors are left, choose the donor that is spatially closest to the receiver.
8) For catchments with no donors found, perform a second round of processing by repeating steps 2)-7) using a set of (user configurable) base attributes, currently including elev_mean, aridity, forest_frac, sand_frac, geo_porosity.
9) For the remaining catchments still with no donors found, perform a final round of processing by identifying a donor within the predefined maximum spatial distance using steps 5)-7).

Similar iterative search procedures are also built into the clustering-based methods in order to identify donors with the optimal combination of physical similarity and spatial proximity.

11. Evaluation of regionalization algorithms

A few steps can be followed to test and evaluate the outputs of the regionalization algorithm.

1) First, given the donor-receiver pairing outputs from the regionalization algorithm, the calibrated formulations (along with optimal parameters) for each donor are transferred to all of its receiver catchments.
2) Second, a realization file is created to provide information on the chosen formulation and its optimal parameter values for all receiver and donor catchments in the model domain, along

with other information necessary for a NextGen retrospective simulation. Alternatively, one can realize the transfer of formulations and parameters by modifying the BMI files.

3) The NextGen framework is then run with the realization file (and BMI files) to produce a retrospective simulation of streamflow.
4) Statistics for various metrics are then calculated to evaluate the streamflow simulation.
5) Finally, multiple such retrospective simulations can be produced for various calibration and regionalization scenarios. An optimal heterogeneous implementation can then be established for the model domain via a mix-and-match type of approach, where the best-performing formulations are chosen for each local region (e.g., a HUC-8 region).

12. Code and other documents

Scripts related to the regionalization tool and its testing/evaluation can be found in OWP GitHub (NextGen_Regionalization). The most recent developments can be found in the NWMv4 folder, for which the sub-folders are described below:

- **algorithm**: python scripts related to the regionalization algorithms
- **attr_calc**: R scripts for calculating the various attributes from hydrofabric data and other catchment datasets
- **ams2024**: scripts related to implementation of regionalization for three different regions (HUC-01, HUC-12, and HUC-17) using scripts in the algorithm and attr_calc folder
- **config**: yaml-based configuration for setting up regionalization
- **data**: datasets to be used for regionalization (e.g., list of donor basins)
- **func**: functions copied from NWMv3 regionalization that can be used for computing streamflow signatures

The input and output datasets for regionalization testing and evaluation can be found at the following directory (accessible from NWC virtual machines):

`/home/yuqiong.liu/NextGen_Regionalization`

In addition, the following presentations related to regionalization might be useful:

AMS 2024 meeting

AGU 2022 meeting

NWMv3 regionalization training for RFCs

NWMv3 regionalization prep for operational implementation

References

Addor, N., A.J. Newman, N. Mizukami, and M.P. Clark, 2017. The CAMELS data set: catchment attributes and meteorology for large-sample studies. Hydrol. and Earth Syst. Sci., 21, 5293-5313, doi:10.5194/hess-21-5293-2017

Gupta, H.V., H. Kling, K.K. Yilmaz, G.F. Martinez, 2009. Decomposition of the mean squared error and NSE performance criteria: implications for improving hydrological modelling. J. Hydrol., 377, 80-91, doi: 10.1016/j.jhydrol.2009.08.003

Wolock, D. M., Winter, T. C. & McMahon, M. (2004) Delineation and evaluation of hydrologic-landscape regions in the United States using geographic information system tools and multivariate statistical analyses. Environ. Manage. 34, S71–S88.

Gower, J. C. (1971). A general coefficient of similarity and some of Its properties. Biometrics, 27(4), 857-871.

J.E. Nash and J.V. Sutcliffe, 1970. River flow forecasting through. Part I. A conceptual models discussion of principles. Journal of Hydrology, 10, 282-290