Metadata for Supporting data for Savoy et al. (2019): Metabolic regimes in flowing waters: An approach for classifying river productivity regimes

Table 1. Description of the dataset

Title of dataset	*Average annual productivity regimes for 47 rivers*
URL of dataset	*The current dataset is posted to CUAHSI hydroshare (https://www.hydroshare.org/resource/eba152073b4046178d1a2ffe9a897ebe/)*
Abstract	*These files comprise the derived dataset of gross primary productivity (GPP) regimes across 47 streams and rivers in the United States from Savoy et al. (2019). A table of basic site information and the cluster (regime) membership of each site from the clustering analysis is included. Additionally, representative time series of GPP were calculated for each site based on original and gap-filled data. Finally, the normalized representative time series actually used to generate the clusters are presented for each site.* *Both two and four clusters are presented in Savoy et al. (2019). Under the four clustering solution, individual sites were classified as "spring peak" rivers, "summer peak" rivers, "aseasonal" rivers, or "summer decline" rivers according to the temporal pattern in GPP as described in Savoy et al. 2019, Limnology and Oceanography (https://doi.org/10.1002/lno.11154).*
Keywords	*River metabolism, metabolic regimes, gross primary production, classification, StreamPULSE*
Lead author for the dataset	*Philip Savoy*
Title and position of lead author	*Postdoctoral researcher*
Organization and address of lead author	*Duke University* *Department of Biology* *130 Science Dr., Durham, N.C., 27708*
Email address of lead author	*prs15@duke.edu*
Additional authors or contributors to the dataset	*Alison P. Appling, James B. Heffernan, Edward G. Stets, Jordan S. Read, Judson W. Harvey, Emily S. Bernhardt*
Organization associated with the data	*Duke University*
Funding	*The work described here was supported by working group funds through the USGS Powell Center Synthesis and by a NSF Macrosystems Program Grant (EF 1442439)*
License	*CCBY*

Geographic location – verbal description	*This data set consists of average annual gross primary production regimes from 47 streams and rivers located within the continental United States.*
Geographic coverage bounding coordinates	*Northern: 45.403452* *Southern: 28.332787* *Western: -122.754819* *Eastern: -75.673271*
Time frame - Begin date	*1 January 2013*
Time frame - End date	*31 December 2016*
General study design	*The study was a synthesis of estimates of river gross primary productivity (GPP) on a subset of data from Appling et al. (2018), (https://doi.org/10.1038/sdata.2018.292).* *The primary objectives were to examine whether there were characteristic GPP regimes and whether these regimes reflected differences in site characteristics or environmental drivers related to GPP.*
Methods description	*We used a subset of 47 streams and rivers to perform a classification of stream GPP regimes. Dynamic time warping (DTW) was used to define the similarity between time series because of its widespread application in time series analysis. The DTW dissimilarity matrix was then used to perform a hierarchical agglomerative clustering. We defined between two and ten clusters because no a priori optimal number of clusters exists. A suite of indices that effectively describes a combination of within-cluster cohesion and between-cluster separation was used to assess the clusters and determine a final set of clusters.*
Laboratory, field, or other analytical methods	*No new data were collected as part of this study, so we do not detail additional lab, field, or other analytical methods here. River metabolism estimates (rates of gross primary production, GPP and ecosystem respiration, ER) for these 47 stream and river sites were originally presented in Appling et al. (2018), (https://doi.org/10.1038/sdata.2018.292).*
Taxonomic species or groups	*Not applicable*
Quality control	*The subset of sites in this analysis were filtered based on a combination of data quality and coverage to select a subset of 47 rivers. Only sites that had GPP estimates for at least 50% of each year from 2013-2016 were selected. Additionally, sites were selected based on their proximity to dams.*
Additional information	

Table 2. Site and basic information

Dataset filename: *site_basic.csv*
Dataset description: *A table containing basic site information and clustering assignments for each site*

Column name	Description	Units
Site_ID	The National Water Information System (NWIS) unique identifier for each site. Sites represent USGS gauged sites, and each Site_ID corresponds to a location in the dataset available from Appling et al. (2018), (https://doi.org/10.5066/F70864KX).	none
Site_name	The full site name for each location	none
Lat	Latitude (NAD83)	decimal degrees
Lon	Longitude (NAD83)	decimal degrees
WS_area	Watershed area	km^2
Width	Channel width (m) derived from regional hydraulic geometry coefficients (Gomez-Velez et al. 2015)	m
two_clus	The cluster each site was assigned to for a two cluster solution (summer peak, spring peak)	none
four_clus	The cluster each site was assigned to for a four cluster solution (summer peak, spring peak, summer decline, aseasonal)	none

Dataset filename: *avg_gpp.csv*

This file contains representative time series of GPP (gO_2 m^{-2} d^{-1}) for each of the 47 sites used. This data is derived from a subset of daily estimates of stream metabolism described in Appling et al. (2018), (https://doi.org/10.1038/sdata.2018.292). The original data are freely available to download (Appling et al. 2018, https://doi.org/10.5066/F70864KX) and full descriptions of the original datasets can be found within these sources. The original set of 356 sites was filtered based on a combination of data quality and coverage to select a subset of 47 rivers that all had data for the time period of 2013-2016. This file consists of the mean time series of GPP for each site that was calculated by taking the mean GPP for each day of the year across all four years of data. The first column (***DOY***) is the day of year and each subsequent column corresponds to a specific ***Site_ID***.

Dataset filename: *avg_gpp_filled*

This file contains representative time series of GPP (gO_2 m^{-2} d^{-1}) for each of the 47 sites used; however, the time series of GPP has been gap-filled. To create these series the original daily GPP estimates were gap-filled using a generalized additive model with both seasonal and trend components. These gap-filled series were then used to calculate the mean time series of GPP for each site that was calculated by taking the mean GPP for each day of the year across all four years of data. The first column (***DOY***) is the day of year and each subsequent column corresponds to a specific ***Site_ID***.

Dataset filename: *normalized.csv*

To calculate similarity with DTW it is necessary to z-normalize each time series. The representative time series of gap-filled GPP as described above were thus z-normalized. The first column (***DOY***) is the day of year and each subsequent column corresponds to a specific ***Site_ID***.

Table 3. Data provenance

If you used data derived from other sources, provide the information here so future users know where the data came from.

Dataset title	**Dataset DOI or URL**	**Creator (name & email)**	**Contact (name & email)**
The metabolic regimes of 356 rivers in the United States	https://doi.org/10.1038/sdata.2018.292	*Alison P. Appling* (aappling@usgs.gov)	*Alison P. Appling* (aappling@usgs.gov)
Metabolism estimates of 356 rivers in the United States (2007-2017)	https://doi.org/10.5066/F70864KX	*Alison P. Appling* (aappling@usgs.gov)	*Alison P. Appling* (aappling@usgs.gov)