READ ME FILE for TNC Database of Estimated Natural Monthly Flows for CA Streams

Daren Carlisle
US Geological Survey, 22 September 2016

Callifornia_COMIDs.csv, California_COMIDS.xlsx **(Files)**

Contains list of all NHDPlus V2 unique segment identifiers (COMID) within the geographic area of California. File is used in the application of models to all CA stream segments.

COMID_Extractor.R **(File)**

Contains R script that uses a user-created input file that has a list of desired COMIDs, extracts from the database all monthly statistics predictions for those COMIDs, organizes the files, then outputs results.

PerformancePLots.R **(File)**

Contains R scripts used to create performance plots located within the "Performance" Folders for each region (see below).

AllGages_Observed_Monthlies **(Folder)**

Contains 4 files, one each for monthly maximum, minimum, mean, and median values. These are observed values of flow, in cubic feet per second, obtained directly from NWIS for all sites in California. Each file has the following fields: ID (USGS gage station identifier with leading "T" added), Year, Month, count_nu (number of days in month for which a daily mean value is available), max/min/mean/p50 (indicates value of monthly statistic—note that p50 is median), NHDV2_COMID (unique NHDPlus V2 identifier for stream segment), and comidyear (combined COMID and year fields for convenience in data processing). Files within this folder are used in model development.

COAST, MNTS, XERIC **(Folders)**

These folders have identical structure and file contents, and contain all model predictions. Each folder contains:

A. "MonthlyScripts2rev .R" (R script used to develop, evaluate, and apply models of monthly statistics for region. See Supplemental MetaData Folder for detailed explanation of R scripts),
B. "Mean", "Min", "Max", and "Median" folders. Each of these folders has identical structure and contents. Each folder contains:

    i. "Perform_xx.csv", where _xx indicates month. The file has the following fields: num (bootstrap number), RSR (root mean square error normalized by the standard deviation of observed data, Moriasi et al. 2007), RMSE (root mean square error),

rsquared (squared correlation of observed and predicted values), pbias (percent bias, Moriasi et al. 2007), NSE (Nash-Sutcliffe Efficiency, Moriasi et al. 2007), MnOE (mean ratio of observed to predicted values), MedOE (median ratio of observed to predicted values), SDOE (mean standard deviation of observed to predicted values), RtMSEt (root mean square error, VanSickle et al. 2006, where RMSE is computed using all observations), RtMSEs (root mean square error, Van Sickle et al. 2006, where the RMSE is computed separately for each site, then averaged).

ii. "x.x.xx_xx_Pred.csv", where x.x.xx is the code for the Level 3 (North American) Ecoregion, and _xx indicates month. The file has the following fields: COMID (NHDPlusV2 segment identifier), AREA (total upstream drainage area in square kilometers), Year, P10 (the 10th percentile of the 1000 predictions of flow [unadjusted for drainage area] from the individual regression trees in the random forest), P50 (the 50th percentile of the 1000 predictions…ibid), P90 (the 90th percentile of the 1000 predictions…ibid), MEAN (the mean of the 1000 predictions…ibid), P10_Q (the 10th percentile of flow, in cubic feet per second, of the 1000 predictions…ibid. This is the lower confidence bound of predicted flow), P50_Q (the 50th percentile of flow, in cubic feet per second, of the 1000 predictions…ibid), P90_Q (the 90th percentile of flow, in cubic feet per second, of the 1000 predictions…ibid. This is the upper confidence bound of predicted flow), Estimated.Q (predicted flow, in cubic feet per second), Month.

C. "Performance" Folder. Each of these folders contains plots of 4 measures (RSR, MnOE, NSE, SDOE) of model performance, by month, for each monthly statistic. Files of plots are named "xxPerform.pdf" where xx indicates statistic (Mean, Median, Max, or Min). Plots are box plots using default settings in R, with outliers removed.

***NHDV2* (Folder)**

Contains data files, organized by Level 3 Ecoregion, for stream segments that are used in applying models. Files in this folder are used in model development and application. Within each folder (numbered using Level 3 Ecoregion codes), the following files exist:

i. "segnatpred.csv" contains static natural predictor data for each stream segment. See Supplemental MetaData Folder for field names and variable descriptions,

ii. "tavxx.csv", where xx indicates month. Each file has the following fields: Year, t0 (mean air temperature [PRISM] for month indicated in file-name suffix), t1 (mean air temperature 1-month prior to the month indicated in file-name suffix), t2 (mean air temperature 2-months prior to the month indicated in the file-name suffix), t3-t12 (same as just described), COMID (NHDPlus V2 segment identifier).

iii. "pptxx.csv", where xx indicates month. Same file structure and variable names as for "tav," but these files are for monthly precipitation.

iv. "runxx.csv", where xx indicates month. Same file structure and variable names as for "tav," but these files are for modeled monthly runoff.

***ReferenceSiteFiles* (Folder)**

Contains files used for efficiency in model development. The following files exist:

i. "CA_USGS_Gage_Reference_Screen_Aug2016.csv" File is identical to one presented to TNC under separate cover. Field names include:

ii. "staticpreds.csv" File has identical structure and field names as "segnatpred.csv" file within the NHDV2 Folder. This file contains the data from that file, but just for reference sites in California.

iii. "xxXX.csv", where xx indicates month and XX indicates runoff ("run"), precipitation ("ppt"), or air temperature ("tav"). These files have the same structure and variables as those in the NHDV2 Folder. These files contain the data from those files, but just for reference sites in California.

Graphical P*erformance Summaries* (Folder)

Contains folders for each region, all with identical internal structure. Within each folder, there are files "xx_Performance.pdf", where xx indicates monthly statistic (Mean, Median, Max, Min). For each statistic, four performance metrics were selected from the "Perform_xx.csv" files, then plotted vs month of year.

***OldFiles* (Folder)**

Contains old files from previous iterations. Used solely in place of a trash bin.

Citations

Moriasi, D. N., J. G. Arnold, M. W. Van Liew, R. L. Bingner, R. D. Harmel and T. L. Veith (2007). "Model evaluation guidelines for systematic quantification of accuracy in watershed simulations." Transactions of the ASABE 50(3): 885-900.

Van Sickle, J., D. D. Huff and C. P. Hawkins (2006). "Selecting discriminant function models for predicting the expected richness of aquatic macroinvertebrates." Freshwater Biology 51(2): 359-372