**Summary Documentation for Streamflow Signatures (in helperFunctions.R)**

**1. Summary of Functions**

Each function, its purpose, assumptions, and key data processing decisions are summarized below:

**process\_gages\_rawData()**

* **Purpose**: Processes raw gage data (USGS/Canadian) to calculate streamflow metrics for individual gages and save results to a file. Data are available up to the present day, but no climate data are directly available.
* **Key Decisions**:
  + A folder of USGS and HYDAT (Canadian) metadata must be stored locally in order to process watersheds.
  + A minimum number of valid years (min\_num\_years) is required for a gage to be included.
  + Data is filtered by a minimum flow threshold (min\_Q\_value\_and\_days). A watershed with fewer than the minimum days of > min\_Q\_value is not processed, but if the threshold is surpassed then non-zero flow is still passed to subsequent functions for processing.
  + Missing or invalid years are excluded from analysis.

**process\_caravan\_gages()**

* **Purpose**: Processes Caravan NetCDF timeseries data for watersheds, calculates streamflow metrics, and saves results. Coincident climate data are available, but the record of streamflow is truncated to as early as the early 2000s (as late as 2018 for HYSETS).
* **Requirements and Decisions**:
  + Minimum years of valid data (min\_num\_years) are required for processing.
  + Filters years based on minimum flow thresholds and valid days.
  + Handles redundancy between datasets (e.g., CAMELS and HYSETS).

**generate\_streamflow\_dt() and generate\_streamflow\_dt\_caravan()**

* **Purpose**: Converts raw gage data (USGS/Canadian) or Caravan NetCDF data into a standardized streamflow data.table.
* **Requirements and Decisions**:
  + Streamflow is provided in appropriate units (or converted, e.g., m³/s to mm/day).
  + Filters data by specified date ranges and required years of data.
  + Handles missing or invalid values (e.g., flagged data).
  + Adds derived columns (e.g., year, month, day of year).

**calculate\_flow\_vols\_by\_year()**

* **Purpose**: Calculates annual and seasonal flow volumes, as well as flow percentiles (e.g., Q1, Q10, Q50, Q90).
* **Requirements and Decisions**:
  + Requires year, Q, month, and doy columns in input data.
  + Seasonal periods are defined as:
    - Winter: December-February
    - Spring: March-May
    - Summer: June-August
    - Fall: September-November
  + Excludes years with insufficient data (<250 days).

**analyze\_baseflow\_indices()**

* **Purpose**: Calculates baseflow indices (BFI) using Eckhardt and Lyne-Hollick digital filters.
* **Requirements and Decisions**:
  + Streamflow (Q) must be valid for at least 250 days/year to calculate annual BFIs.
  + Default parameters for Eckhardt filter: BFImax = 0.8, a = 0.98.
  + Excludes years with >20% missing data.

**analyze\_recession\_parameters()**

* **Purpose**: Analyzes recession events to calculate parameters like log(a), b, and concavity.
* **Requirements and Decisions**:
  + Recession events are defined as periods of at least 5 consecutive days with monotonic decreases in flow (Q) and its derivative (|dQ/dt|).
  + Watersheds with fewer than 25 event observations are excluded.
  + Splits recession events temporally (first vs. second half) for concavity analysis.
  + Fits sinusoidal models to analyze seasonal variation in log(a).

**analyze\_flashiness\_trends()**

* **Purpose**: Calculates the Richards-Baker (RB) flashiness index by year and analyzes trends.
* **Requirements and Decisions**:
  + Requires at least 30 valid days/year to calculate flashiness.
  + Missing values are interpolated if <20% of the data is missing.

**calculate\_pulse\_metrics()**

* **Purpose**: Calculates metrics for high/low flow pulses (e.g., frequency, duration) and flow reversals.
* **Requirements and Decisions**:
  + High/low flow pulses are defined using the 90th and 10th percentiles of flow (Q).
  + Years with <250 days of valid data are excluded.
  + Calculates separate metrics for annual and period-of-record thresholds.
  + Flow reversals are counted based on changes in flow direction exceeding 2% of current flow.

**analyze\_flow\_timing\_trends()**

* **Purpose**: Analyzes the timing of flow events (e.g., day of maximum flow, cumulative flow percentiles).
* **Requirements and Decisions**:
  + Requires at least 300 valid days/year to calculate timing metrics.
  + Calculates Julian days for cumulative flow percentiles (e.g., D10, D50).

**analyze\_Q\_PPT\_relationships()**

* **Purpose**: Analyzes runoff ratios (streamflow divided by precipitation) and their trends.
* **Requirements and Decisions**:
  + Precipitation (PPT) data must be available alongside streamflow (Q).
  + Calculates separate runoff ratios for annual and seasonal periods (winter, spring, summer, fall).
  + Handles cases with near-zero precipitation by assigning NA to ratios.

**2. Glossary of Streamflow Signatures**

**Flow Volume Metrics**

* **Qann**: Annual mean streamflow.
* **Qwin**: Winter mean streamflow.
* **Qspr**: Spring mean streamflow.
* **Qsum**: Summer mean streamflow.
* **Qfal**: Fall mean streamflow.
* **Qxx**: Flow percentiles (e.g., Q10, Q50, Q90).

**Baseflow Metrics**

* **BFI\_Eckhardt**: Baseflow index calculated using the Eckhardt filter.
* **BFI\_LyneHollick**: Baseflow index calculated using the Lyne-Hollick filter.

**Recession Metrics**

* **log\_a\_pointcloud**: Median log(a) from point cloud analysis of all recession events.
* **log\_a\_events**: Median log(a) from individual recession events.
* **b\_pointcloud**: Median recession exponent b from point cloud analysis.
* **b\_events**: Median b from individual events.
* **concavity**: Difference in b between the first and second halves of recession events.
* **log\_a\_seasonality\_amplitude\_all**: Seasonal amplitude of log(a) for all years.
* **log\_a\_seasonality\_minimum\_all**: Julian date of minimum log(a) for all years.

**Pulse Metrics**

* **n\_high\_pulses\_year**: Number of high-flow pulses per year.
* **n\_low\_pulses\_year**: Number of low-flow pulses per year.
* **dur\_high\_pulses\_year**: Mean duration of high-flow pulses per year.
* **dur\_low\_pulses\_year**: Mean duration of low-flow pulses per year.
* **TQmean**: Percentage of days with flow above the annual mean.
* **Flow\_Reversals**: Number of flow reversals (annual or seasonal).

**Flashiness Metrics**

* **R-B Index**: Richards-Baker flashiness index.

**Flow Timing Metrics**

* **Dxx\_day**: Julian day when cumulative flow reaches a given percentile (e.g., D10, D50, D90).
* **D25\_to\_D75**: Days between 25% and 75% cumulative flow.
* **Dmax**: Julian day of maximum flow.

**3. Glossary of Statistical Metrics**

* **slp**: Theil-Sen slope, a robust trend estimator over time.
* **rho**: Spearman's rank correlation coefficient, a non-parametric measure of monotonic trends.
* **pval**: P-value associated with rho, indicating statistical significance.
* **mean**: Arithmetic mean of the signature across all years.
* **median**: Median value of the signature across all years.