**SWOT Calibration and Validation Hydrology Toolboxes**

**Camryn Kluetmeier, UNC**

Toolboxes for the hydrology corrections of cal/val field data for comparison with SWOT satellite data.

To run the toolboxes, the following inputs from the PO.DAAC repository are required:

PT inputs:

* PT data directory containing all raw PT csv files. First 11 rows of the csv are metadata, header on line 12 contains the following data:

|  |  |
| --- | --- |
| **Variable** | **Description** |
| Date | yyyy/mm/dd format |
| Time | UTC |
| Level | meters |
| Temperature | °C |
| Offset | meters |

* PT key csv file that links PTs to SWORD reaches/nodes, install/uninstall GNSS drifts, and contains spatial information about the PT

|  |  |
| --- | --- |
| **Variable** | **Description** |
| PT\_Serial | 7-digit PT serial number |
| Label | Field label (e.g. PT1, BL1) |
| Node\_ID | SWORD node ID |
| Reach\_ID | SWORD reach ID |
| US\_Reach\_ID | Upstream SWORD reach ID |
| DS\_Reach\_ID | Downstream SWORD reach ID |
| Lat\_WGS84 |  |
| Long\_WGS84 |  |
| Easting\_UTM |  |
| Northing\_UTM |  |
| Install\_method | E.g. rebar, cinderblock |
| Date\_install | m/dd/yyyy format |
| Time\_Install\_UTC | UTC 24 hour format |
| Date\_Uninstall |  |
| Time\_Uninstall |  |
| GNSS\_install\_date |  |
| GNSS\_install\_start |  |
| GNSS\_install\_end |  |
| GNSS\_offset\_m | meters |
| Receiver | GNSS receiver used (e.g. Rec3) |
| Original\_log\_file |  |
| Final\_log\_file |  |
| GNSS\_uninstall\_start |  |
| GNSS\_uninstall\_end |  |
| Receiver |  |
| Original\_log\_file |  |
| Final\_log\_file |  |

GNSS inputs:

* GNSS drift data directory with all raw GNSS netCDF files containing:

|  |  |
| --- | --- |
| **Variable** | **Description** |
| Latitude |  |
| Longitude |  |
| WSE | Water surface elevation (meters) wrt geoid |
| Time\_tai | Seconds since January 1 2000 at midnight without leap seconds |
| Motioncode\_flag | 0, 1, or 2. Only 2 indicates quality data |
| Surfacetype\_flag | 10, 11, or 12. Only 12 indicates quality data |
| ellipsoid\_semi\_major\_axis |  |
| ellipsoid\_flattening |  |
| postion\_3drss\_formal\_error |  |
| infoEventDescription |  |
| infoEventStartTime |  |
| infoEventEndTime |  |

SWORD inputs:

* SWORD data in the netCDF format brought in at the continental scale with centerline, node, and reach data. See SWORD Production Description PDF for detailed dataset documentation.

Empty directories are created to store munged data outputs for PTs, GNSS drifts, and integrated SWORD products in addition to directories created for quality control flagged PT and GNSS drift output data.

**Creating data frames from GNSS drifts**

First, check for un-munged GNSS files in the GNSS drift data directory that have not already been processed. If there are, then run the create\_GNSS\_dataframe function.

The function reads in the netCDF GNSS drifts sourced from PO.DAAC and creates a data frame in an xml format for cleaning and quality flagging.

Event code times are converted to POSIXct UTC. A two-minute buffer is added to the start and end time of the event codes, before powerline and bridge events are filtered out.

All of the GNSS times in the TAI format are converted to POSIXct UTC. Only data with the surface flag of 12 and motion code of 2 are kept. Data with an uncertainty greater than 5cm are filtered out.

If the resulting data frame has no lines, the toolbox will print that the filename bonked.

Finally, the munged data frame is written to a csv in the GNSS output directory with the following fields:

|  |  |
| --- | --- |
| **Variable** | **Description** |
| GNSS\_Lat | WGS84 |
| GNSS\_Lon | WGS84 |
| GNSS\_wse | In meters |
| GNSS\_time\_tai | Seconds since January 1 2000 at midnight without leap seconds |
| GNSS\_uncertainty | In meters, only values < 0.05 m |
| GNSS\_surf\_flag | Should be 12, indicating quality data over water |
| GNSS\_motion\_flag | Should be 2, indicating moving data collection |
| GNSS\_time\_UTC | yyyy-mm-dd hh:mm:ss format |
| GNSS\_ellipsoid | Corrected by JPL |
| Drift\_ID | Following calval naming conventions |

**Correct PTs to GNSS**

Error thresholds for data:

**Distance threshold:** 150 meters, PT must be within a 150 m radius of a GNSS drift

**Time threshold:** 900 seconds (15 minutes), centered around PT collect, so 15 minutes before and after

**GNSS standard deviation:** 0.15 meters, how much variance is allowed in GNSS data when it is within the distance threshold

**Offset standard deviation:** 0.10 meters, how much variance is allowed in PT data from put in to take out

**Change threshold over 15 minutes**: 0.05 meters, does it change more than 5cm in 15 minutes? If so, that is a discontinuity in offset

These will be specific to each river and are subject to change/should be up for debate!

First, check for un-munged PT data in the PT data directory that have not already been processed. If there are, then run the correct\_pt\_to\_gnss function.

The function reads in the raw PT files and the PT key file and IDs PTs based on their serial numbers. There are 11 lines of headers to skip when creating PT data frames in csv format. *For the mock Willamette data there is a missing header so have to skip the 12th line and create a header.*

The last step in data tidying is to merge PT files with the PT key file with a left join based on PT serial number. The PT date and time are also mutated to the standard POSIXct UTC time.

Next up, come the actual corrections to the PT data to get corrected water surface elevation. First, check to make sure the key file is properly joined to the PT data and there is data in the cleaned PT files, if not return NA.

Read in the munged GNSS drift data and mutate the datetime into R’s POSIXct format, as this is lost when the file was written to a csv.

Do an inner join of the PT and GNSS data by datetime, applying the site-specific time threshold. For example, with a threshold of 15 minutes, all GNSS data at 1 Hz within 30 minutes of every PT observation will be pulled.

Use a haversine to determine the distance between all PT data and moving GNSS data. Filter the GNSS drifts to only keep data that are under the distance threshold. For example, with a threshold of 150 m, only GNSS data within a 150 m radius of a PT will be pulled.

Ensure there is at least one GNSS drift to make a valid offset correction by checking that a linked drift exists and contains more than zero rows. Otherwise, return NA.

Multiple GNSS points go to one PT observation, so group the PT by timestep and create the correction of GNSS water surface elevation minus the PT level (both in meters). Every time a drift goes by, take the mean PT offset at that location. Differences in these offsets is the uncertainty of this mapping, so take the standard deviation of the PT offset as the uncertainty metric.

To create the PT corrections, the mean PT offset and the standard deviation of the PT offset are joined to the PT data frame with a left join. Corrected PT water surface elevation is added as the PT level plus the PT correction. Corrected PT water surface elevation standard deviation is added as the standard deviation of the PT offset plus 0.001 m (this is added for PT measurement error).

There is a lot of exception handling that has to take place for a valid offset correction to take place. Errors are thrown in all the following cases and the PT file will be moved into the flagged PT output directory:

* The PT isn’t listed in the key file
* There are no GNSS data within the space-time thresholds
* There are not enough data to create offset (change your thresholds or double check the data), the current threshold is under 5 rows of data
* GNSS range is too large (check thresholds or data), as defined by the GNSS standard deviation threshold
* Correction too noisy (check thresholds or data), as defined by the offset standard deviation threshold derived from the GNSS and PT data
* Offsets are too different over time (check thresholds or data), as defined by the change over 15 minutes threshold

Finally, join the corrected PT and standard deviation measurements with the munged PT data using the nearest time between the GNSS time and PT time in UTC. Remove PT data that is before/after the install/uninstall time. *This process is slow and should be parallelized!*

The last step is to print out the filename and ‘this file has passed all checks’ and write a csv of the final PT data saved to the QA QC PT output directory with the following fields:

|  |  |
| --- | --- |
| **Variable** | **Description** |
| pt\_time\_UTC | yyyy-mm-dd hh:mm:ss format |
| pt\_lat | WGS84 |
| pt\_lon | WGS84 |
| pt\_install\_UTC | yyyy-mm-dd hh:mm:ss format |
| pt\_uninstall\_UTC | yyyy-mm-dd hh:mm:ss format |
| install\_method | E.g. rebar, cinderblock |
| pt\_serial | Seven digit ID number |
| pt\_level | In meters |
| temperature | In °C |
| driftID |  |
| datetime | yyyy-mm-dd hh:mm:ss format |
| pt\_correction | GNSS water surface elevation minus PT level, in meters |
| pt\_wse\_sd | Uncertainty in meters, equal to all variance in all offsets used to create PT water surface elevation plus PT measurement error |
| pt\_wse | PT correction plus PT level, in meters |

**Calculate slopes and heights from drifts within nodes and reaches**

Read in the river-specific SWORD reach data and pull all river reach and node IDs. Set the UTM zone and define a buffer of 50 m to extend the reach.

Run the calculate\_slope\_wse\_fromdrift function. First, set up a function to project latitude and longitude into UTMs.

Bring in SWORD data in the netCDF format and index the reach and node IDs for the specific river.

For node variables, create a data frame with the following fields:

|  |  |
| --- | --- |
| **Variable** | **Description** |
| lon | Node longitude in UTM |
| lat | Node latitude in UTM |
| node\_id | Unique node ID |
| node\_wmax | Max node width, in m |
| node\_length | Node length, in m |
| reach\_id | Reach ID that node falls in |

For reach variables, create a data frame with the following fields:

|  |  |
| --- | --- |
| **Variable** | **Description** |
| lon | Central reach longitude UTM |
| lat | Central reach latitude UTM |
| reach\_id | Unique reach ID |
| reach\_xmax | Max longitude UTM |
| reach\_xmin | Min longitude UTM |
| reach\_ymax | Max latitude UTM |
| reach\_ymin | Min latitude UTM |
| reach\_length | Length, in m |

For both of these processes overwriting the original indexes limits the amount of RAM needed.

Bring in the SWORD centerlines in netCDF format for the specific river and create a data frame with the following variables:

|  |  |
| --- | --- |
| **Variable** | **Description** |
| reach\_id | Reach ID for centerline |
| lon | Centerline longitude UTM |
| lat | Centerline latitude UTM |
| cl\_id | Centerline ID |
| cl\_node\_id | Centerline node ID |

**Calculate node water surface elevation**

Read in the GNSS drift csv file and convert the latitude and longitude to UTMs and the time to POSIXct UTC.

Join the centerline data frame to the node data frame using a left join and compute the angle that the centerline intersects each node at. Using that node angle, make a polygon geometry of four points centered around the node that are spaced out the maximum width and length of the node*.* Convert the points into a box for each node that is a spatial object. *Eventually this process could be replaced with aerial imagery to delineate the water extent.*

Make the GNSS drift data a spatial object and find the intersection between the node boxes and drift points. Group these points by node ID and then take the mean of all GNSS water surface elevation points within the boxes as the mean node water surface elevation (in meters). Compute the mean GNSS time for all drift data within the box and save the reach ID, drift ID, and node water surface elevation precision as 0.05 m.

**Calculate reach water surface elevation and slope from drifts**

Read in the GNSS drift csv file and convert the latitude and longitude to UTMs and the time to POSIXct UTC.

Create an index of relevant drift data that that is within each reach, then for each reach calculate the mean GNSS water surface elevation (in meters), save the reach water surface elevation precision as 0.05 m, water surface elevation start time, and water surface elevation end time.

Search for centerline points within the buffer distance of the top and bottom of each reach to define the slope and pull the length of each reach.

Match GNSS drift data to the centerline points at the top and bottom of the reaches through a geospatial search.

To calculate the slope, take the mean of the water surface elevations at the top of the reach over the buffer area minus the mean of the water surface elevations at the bottom of the reach over the buffer area divided by the length of the centerline. Calculate the standard deviation of the slope as the uncertainty metric and save all outputs to a data frame.

If the whole reach wasn’t floated, don’t compute the slope. If the slope doesn’t exist for a reach get rid of the water surface elevation, because water surface elevation should be averaged over an entire reach’s worth of data.

From these calculations, there are three csv files saved into the node and reach output directories, respectively:

Node water surface outputs saved as ‘rivername’ \_drift\_node\_wses:

|  |  |
| --- | --- |
| **Variable** | **Description** |
| node\_id | Unique SWORD node ID |
| node\_wse | Node water surface elevation, in meters |
| node\_wse\_precision\_m | In meters, instrument error, 0.05 m |
| time | yyyy-mm-dd hh:mm:ss format |
| reach\_id | Unique SWORD reach ID |
| drift\_id | Following calval naming conventions |

Node geometry is saved as a csv with node id and geometry.

Reach water surface elevation and slope as ‘rivername’ \_drift\_reach\_wse\_slope:

|  |  |
| --- | --- |
| **Variable** | **Description** |
| reach\_id | Unique SWORD reach ID |
| wse\_bar | Water surface elevation, in meters |
| wse\_precision | In meters, instrument error, 0.05 m |
| wse\_start | yyyy-mm-dd hh:mm:ss format |
| wse\_end | yyyy-mm-dd hh:mm:ss format |
| slope |  |
| slope\_precision | Standard deviation |
| drift­\_id | Following calval naming conventions |

**Calculate slopes and heights from PTs within nodes and reaches**

Bring in the munged PTs, the key file, and SWORD reaches and nodes for the river of interest. Next, run the calculate\_slope\_wse\_fromPT function.

Make a data frame with the key file and munged PTs by a left join.

To calculate node water surface elevation, group the PT data frame by node ID and find the mean water surface elevation (in meters) for each node. Set the mean PT water surface precision to 0.001 m, the error of the PT instruments.

To calculate reach water surface elevation, group the PT data frame by reach ID and find the mean water surface elevation (in meters) for each reach. Set the mean PT water surface precision to 0.001 m, the error of the PT instruments.

To calculate slope, group PT data by upstream and downstream reach IDs and average water surface elevation over these variables. Uncertainty is the standard deviation of PT water surface elevations at the upstream and downstream reaches. Slope is the mean upstream water surface elevation minus the mean downstream water elevation divided by reach length, segmented by time.

From these calculations, there are three csv files saved into the node and reach output directories, respectively:

PT node water surface elevation:

|  |  |
| --- | --- |
| **Variable** | **Description** |
| node\_id | SWORD node ID |
| pt\_time\_UCT | yyyy-mm-dd hh:mm:ss format |
| mean\_node\_wse\_m | Mean node water surface elevation in meters |
| mean\_pt\_wse\_precision\_m | In meters, instrument error, 0.001 m |

PT reach water surface elevation:

|  |  |
| --- | --- |
| **Variable** | **Description** |
| reach\_id | SWORD reach ID |
| pt\_time\_UCT | yyyy-mm-dd hh:mm:ss format |
| mean\_reach\_wse\_m | Mean reach water surface elevation, meters |
| mean\_pt\_wse\_precision\_m | In meters, instrument error, 0.001 m |

PT reach slope:

|  |  |
| --- | --- |
| **Variable** | **Description** |
| reach\_id | SWORD reach ID |
| pt\_time\_UCT | yyyy-mm-dd hh:mm:ss format |
| slope | Slope |
| slope\_precision | Standard deviation of PT measurements |

**Define what drift goes with what SWOT overpass**

Right now, this is set up with a fake SWOT pass ID and time.

Define thresholds for linking GNSS drifts to SWOT overpasses:

**Time threshold:** Two hours

**Water surface elevation threshold:** Within 5cm

**Distance threshold:** Within 200m

These will be specific to each river and are subject to change/should be up for debate!

Next, run the select\_appropriate\_drift function.

First, read in the node water surface elevation csv, convert time to POSTIXct, and calculate the time to SWOT overpass.

Make an index of what drift node observations to keep and remove based on the time threshold.

For node water surface elevations from drifts within the time threshold, add the time of the SWOT overpass and the SWOT pass ID to a direct match data frame and save to a csv of SWOT drift pairs.

For drifts outside of the time threshold get the drift data at 1 Hz.

Pull the PT levels at the time of the SWOT overpass from the munged PT csv.

Next, compare the drift node levels with the PT levels with a difference join based on latitude and longitude: *(This process is slow.)*

Read in the PT key file. Join the drift and PT data using a nearest neighbor haversine geospatial left join. Compute the difference in water surface elevation between the GNSS and PT data. Then join the key file to the data frame with a left join. Next, filter the data by the distance threshold.

This data frame has now found all drifts within the distance threshold of the PTs.

Next, filter by water level within the water surface elevation threshold to find ‘good’ matches of PT and drift data.

Group by node and drift ID before computing summary statistics to be output in the final csv data frame of SWOT drift pairs:

|  |  |
| --- | --- |
| **Variable** | **Description** |
| node\_id | SWORD node ID |
| drift\_id | Following calval naming conventions |
| drift\_pt\_dist\_km\_bar | Distance from drift to PT in km |
| wse\_difference\_m\_bar | Water surface elevation difference between drift and PT, in meters |
| wse\_difference\_m\_sd | Standard deviation of water surface elevation difference, in meters |
| swot\_passid | Fake swot pass (for now!) |
| swot\_time\_UTC | yyyy-mm-dd hh:mm:ss format |