Download USGS NWIS observations using HydroDesktop, and format the data for RAPID

By Cédric H. David (cedric.david@jpl.nasa.gov),
11 Feb 2013, updated 21 Jan 2015

Goal

Download streamflow measurements from USGS NWIS gages (http://waterdata.usgs.gov/usa/nwis/) and format them for use in RAPID. RAPID observations inputs consist in two files: the first is a list of river IDs where gages are located and the second contains the corresponding measurements. Here, we focus only on those gages that have complete record for a given period of time (no data gap) and that are located on river reaches for which the NHDPlus (http://www.horizon-systems.com/nhdplus/data.php) knows the direction of flow. Along the way, shapefiles to be used in ArcGIS are also created.

Requirements

ArcGIS (http://www.esri.com/software/arcgis).
HydroDesktop (http://his.cuahsi.org/hydrodesktop.html).
Microsoft Excel (http://office.microsoft.com/en-us/excel/).
A Fortran compiler (here we use gfortran http://gcc.gnu.org/fortran/).
A text editor (here we use vim http://www.vim.org/).

Preliminary notes

The shapefile used here for USGS gages is downloaded from the NHDPlus dataset (http://www.horizon-systems.com/nhdplus/data.php) and named StreamGageEvent.shp. One could be tempted to use another file (USGS_Streamgages-NHD_Locations.shp) that is also available online. Unfortunately, the latter was not "snapped" to NHDPlus river reaches and is therefore much more challenging to use when associating an NHDPlus river reach to a USGS gage as needed for RAPID. Using StreamGageEvent.shp allows to successfully Select by location without threshold for all stations. However, the Spatial join seems to need a threshold for intersection.

Previous versions of this tutorial

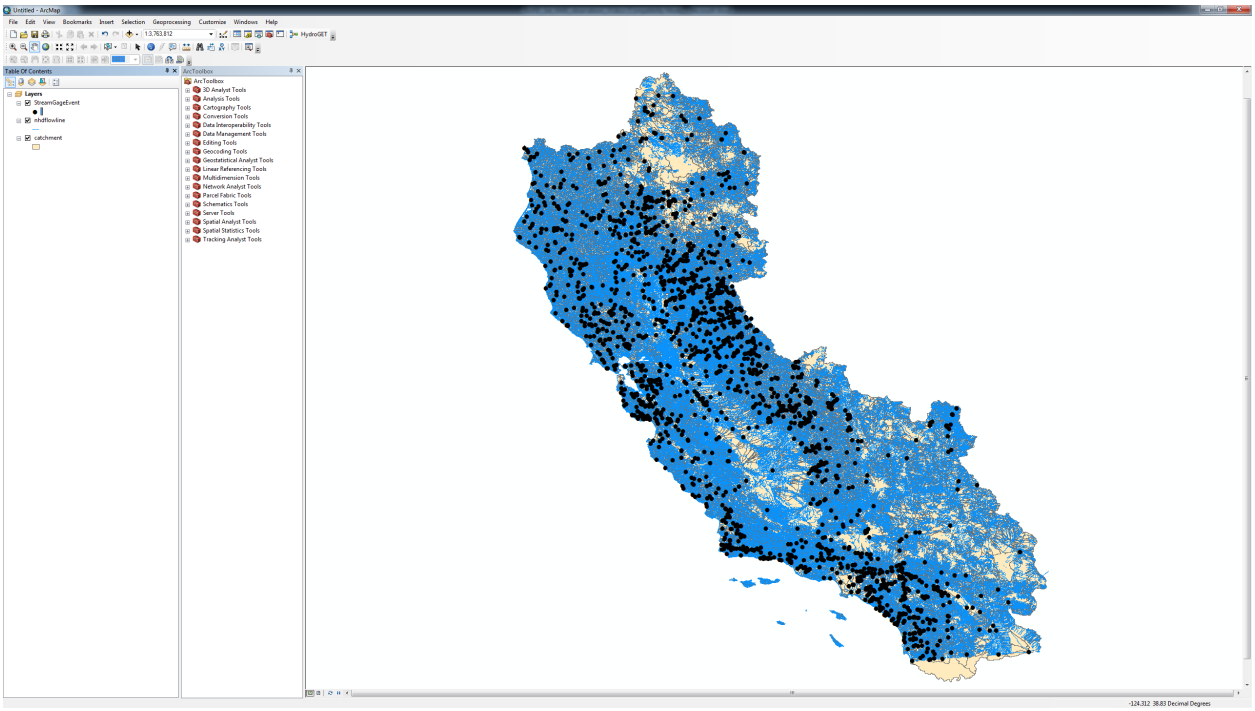
As RAPID increasingly addresses domains of size necessitating the download of observation time series made of millions of data points, it has become common that a few time series fail to properly download (likely due to variations in internet communication) when attempting to obtain time series for several hundred of gages. The original version of this tutorial used a tool developed for ArcGIS called HydroGET (http://his.cuahsi.org/hydroget.html) which is well suited for downloading time series for on the order of 100 gages. Later versions of this tutorial used HydroGET along with a Microsoft Excel file called HydroExcel (http://his.cuahsi.org/hydroexcel.html). HydroExcel was helpful to perform a pre-selection of which gage may have available data prior to running HydroGET hence limiting the number of stations

to be downloaded. This HydroExcel/HydroGET combination worked well for time series from on the order of 500 stations.

This tutorial uses the more-recently developed HydroDesktop (http://his.cuahsi.org/hydrodesktop.html) which offers similar downloading capabilities to that of HydroGET. However, HydroDesktop has an option for re-downloading selected time series if the original download failed. This option is currently lacking from HydroGET, and turns out to be crucial when attempting to download time series from on the order of 1000 stations. However, the method presented here remains applicable to smaller domains.

Preparing GIS data for NHDPlus Region 18 (California)

Download catchment.shp, nhdflowline.shp and StreamGageEvent.shp from NHDPlus (http://www.horizon-systems.com/NHDPlus/NHDPlusV1_data.php). Open these in ArcGIS. There should be 136,883 catchments, 176,142 river reaches and 2,311 gages.

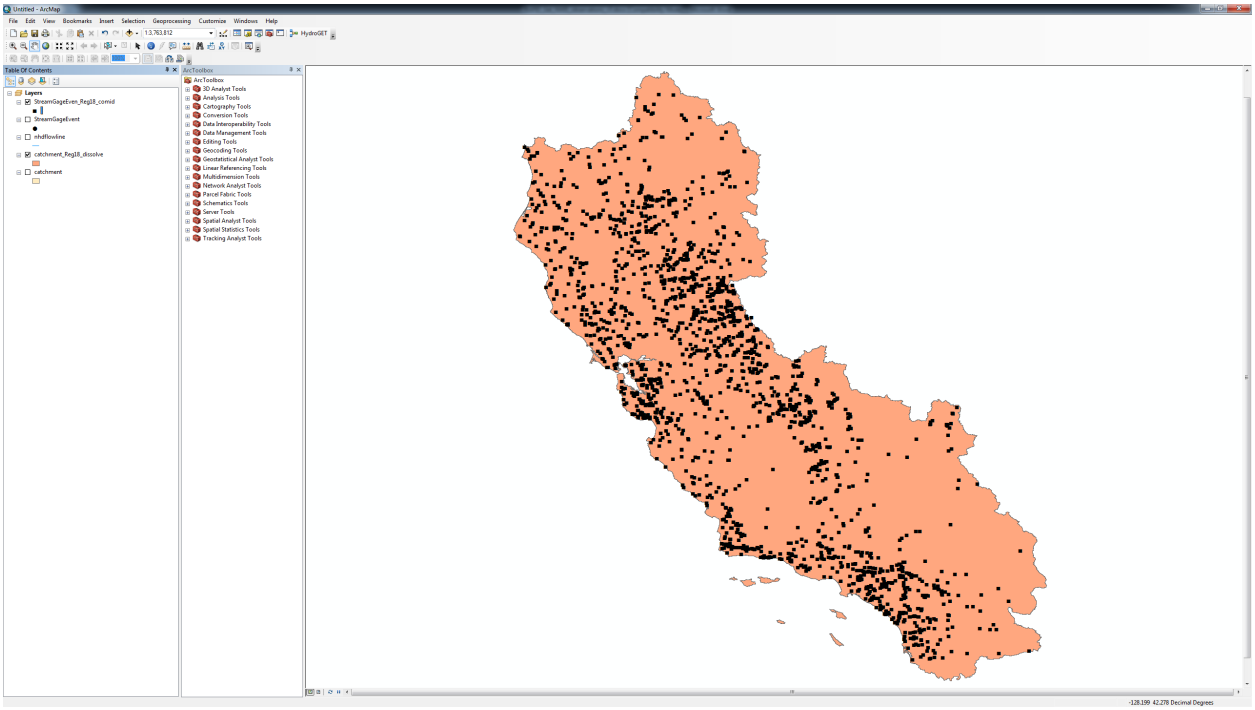


Rename these files as catchment_Reg18.shp, nhdflowline_Reg18.shp and StreamGageEvent_Reg18.shp

RAPID needs to know what river reach each gage is located on. Each NHDPlus river reach has a unique identifier called COMID that is used in RAPID as the unique river ID. Therefore, we need to know what COMID corresponds to each USGS gage. A COMID field is available in StreamGageEvent_Reg18.shp, but for some reason it is not populated. The first step is therefore to fix that. One could be tempted to join both features using REACHCODE, but it is not appropriate here because several COMIDs can correspond to the same REACHCODE. Instead, we'll use a Spatial Join. Select an intersection search radius of 1 decimeter, keep all fields of StreamGageEvent_Reg18.shp, add COMID and FLOWDIR from

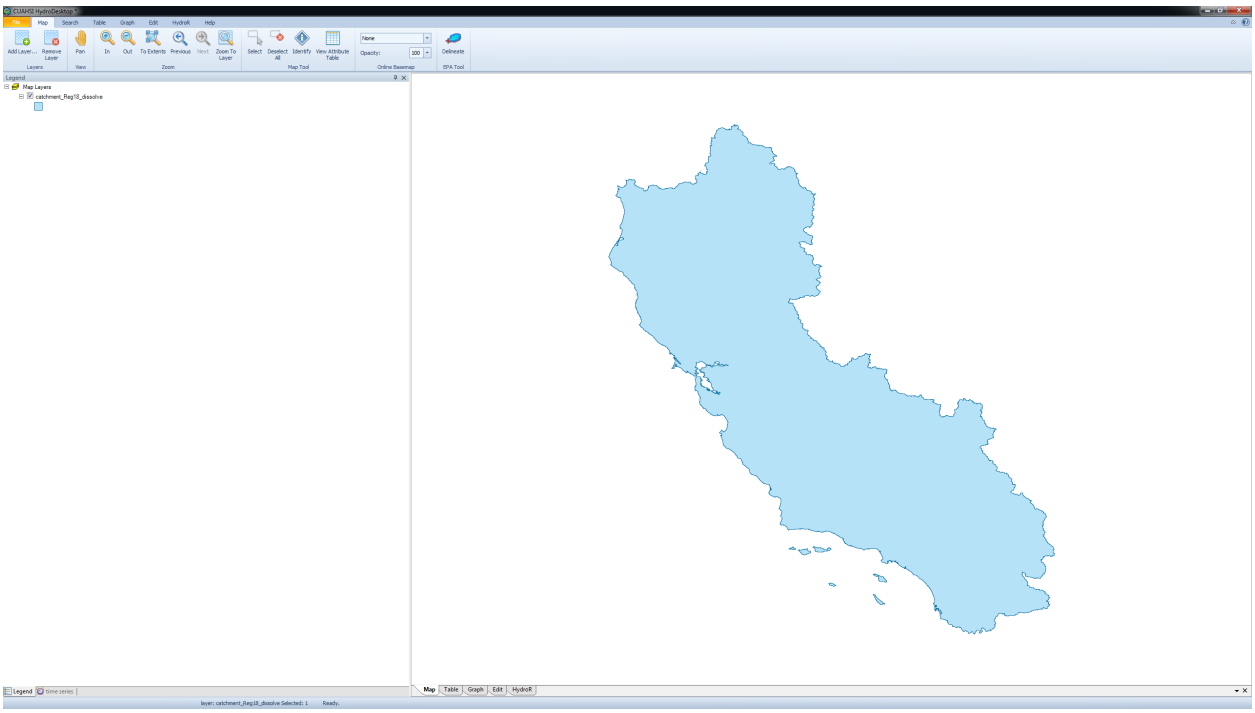
NHDFlowline_Reg18.shp. Save as StreamGageEvent_Reg18_comid.shp. All 2,311 stations should have a COMID associated to them.

Use the Dissolve tool to create a unique feature covering the entire domain catchment_Reg18_dissolve.shp based on the file catchment_Reg18.shp .

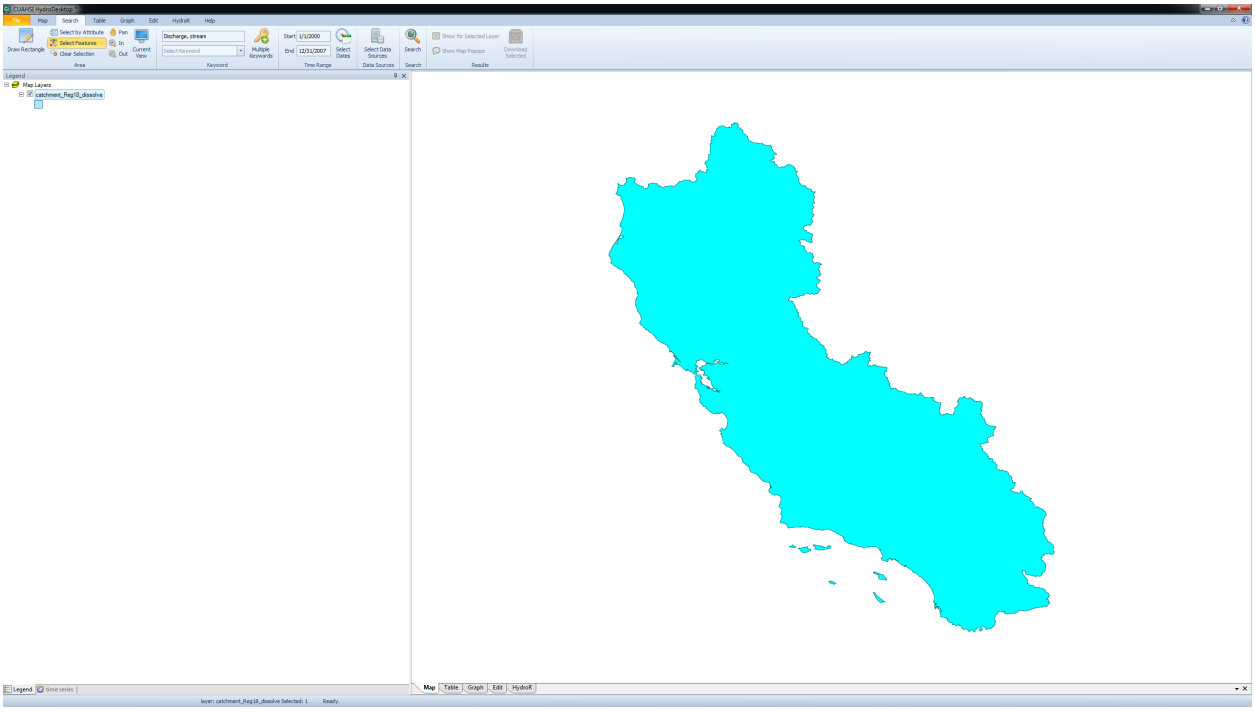


Downloading USGS NWIS data for NHDPlus Region 18 (California) between 2000/01/01 and 2007/12/31 using HydroDesktop

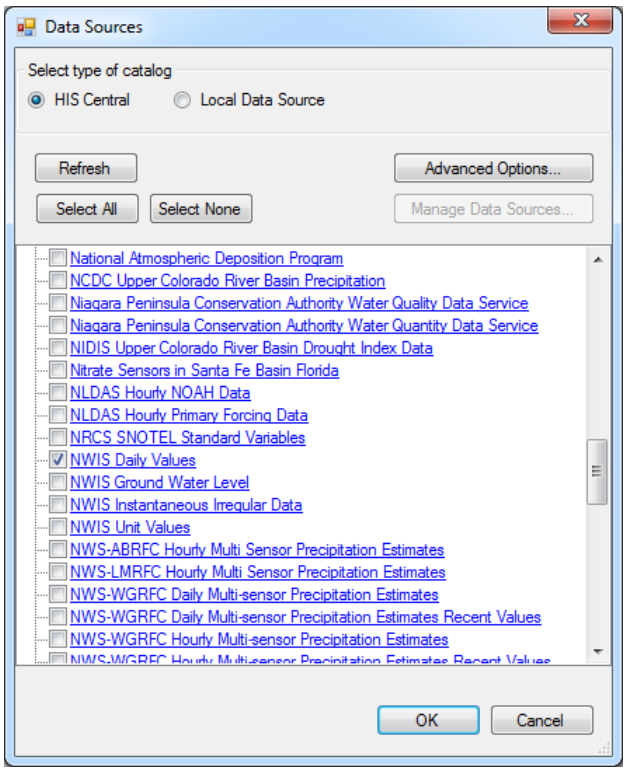
Open HydroDesktop and use the Add layer button from the Map tab to add a polygon shapefile for NHDPlus Region 18. Here we use the shapefile called catchment_Reg18_dissolve.shp prepared above.



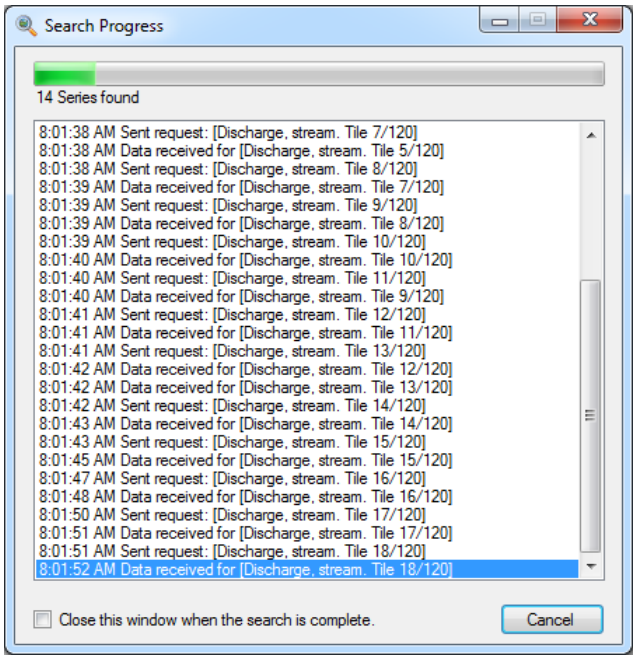
Now move from the Map tab to the Search tab. Click on Select Features and on California from the map to have HydroDesktop search only data for this given location.



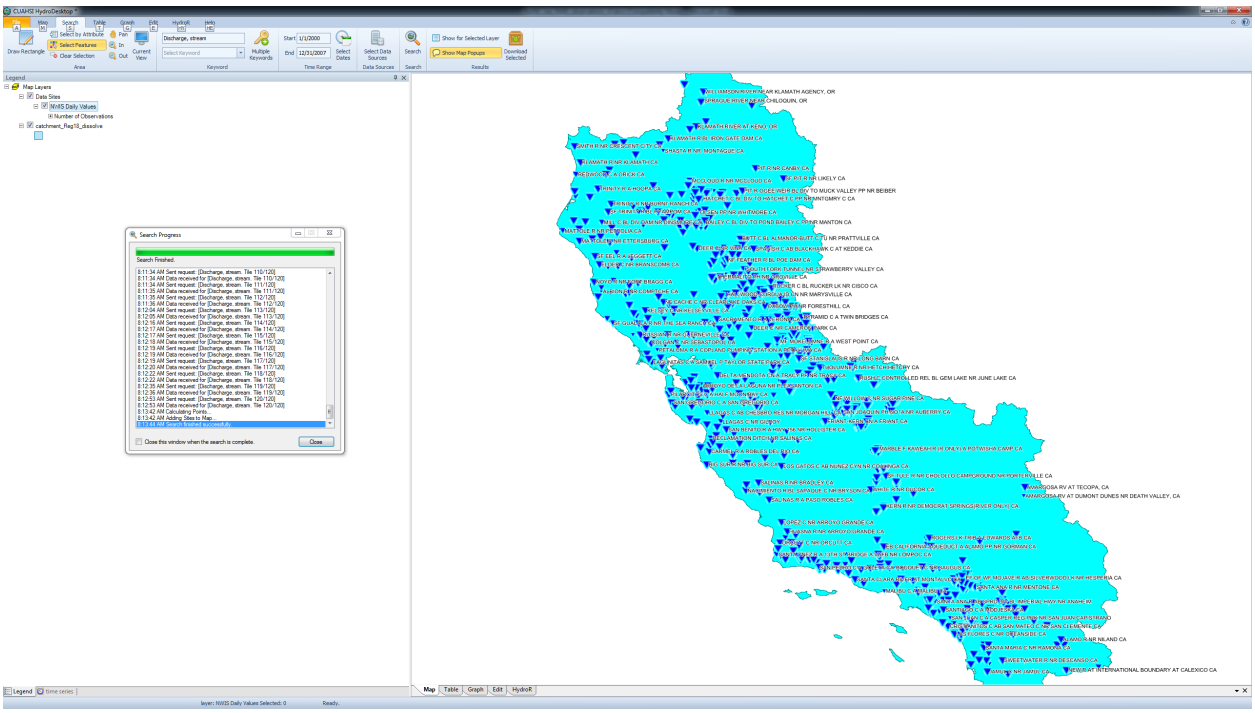
Now click on Select Data Sources and select only NWIS Daily Values as from HIS Central as the source.



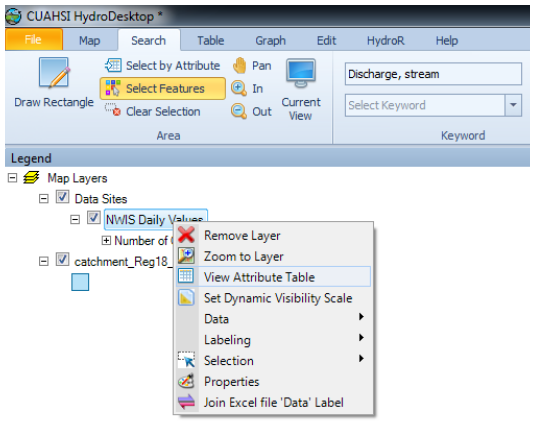
Type Discharge, stream as the Keyword. Select 1/1/2000 for Start and 12/31/2007 for End of the Time Range. Click on Search.



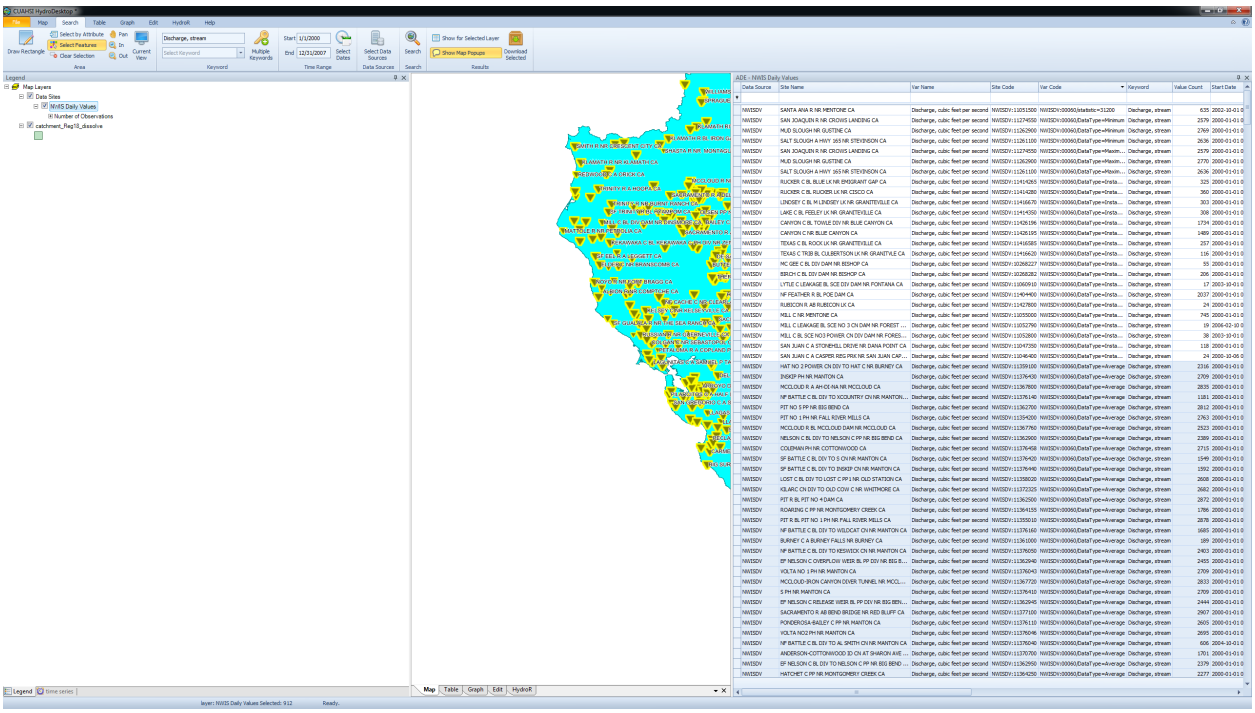
Once the search is over, the following map appears with the gages that were found:



Open the attribute table through right clicking on the NWIS Daily Values in Map Layers.

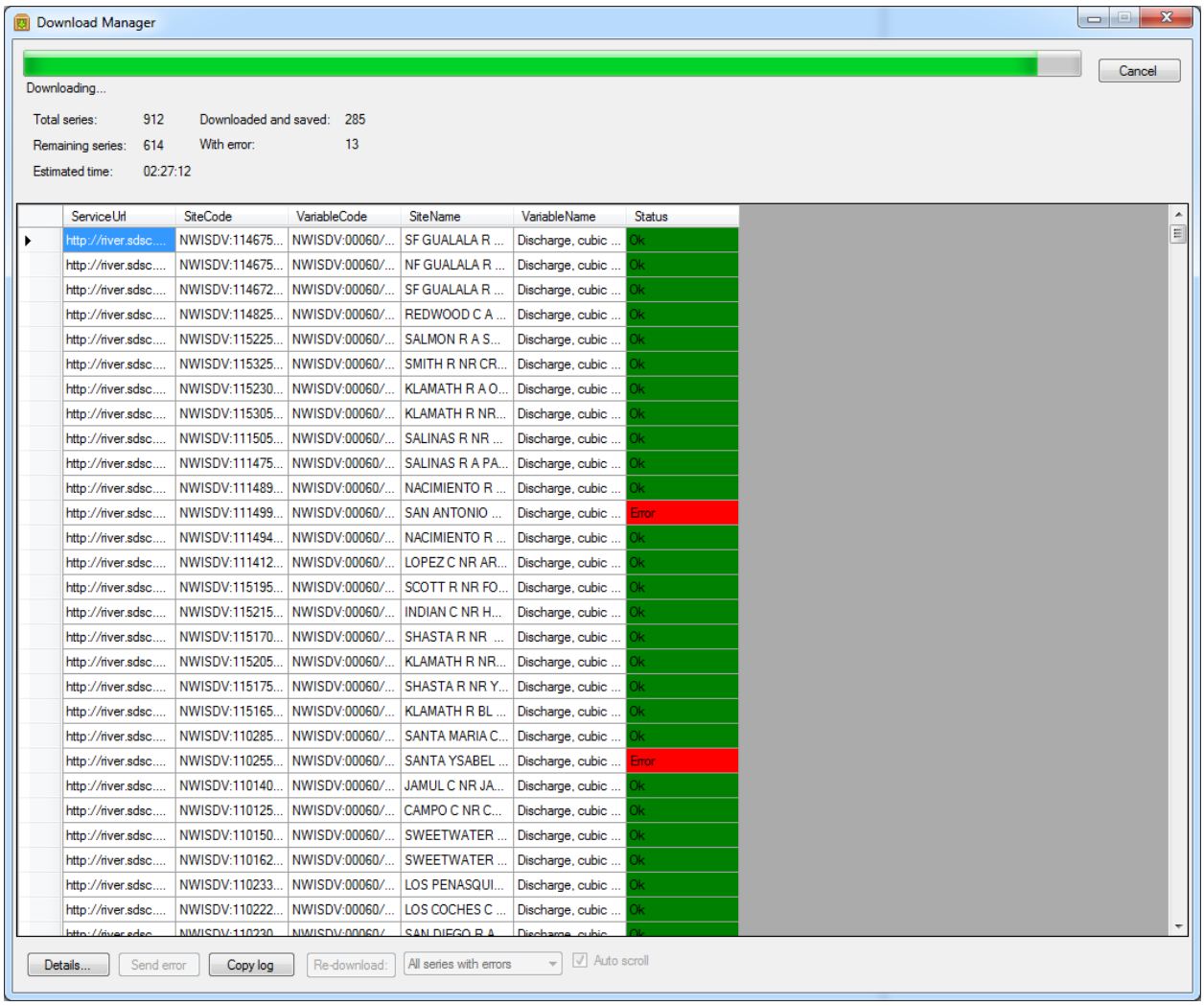


There should be 937 gages available.

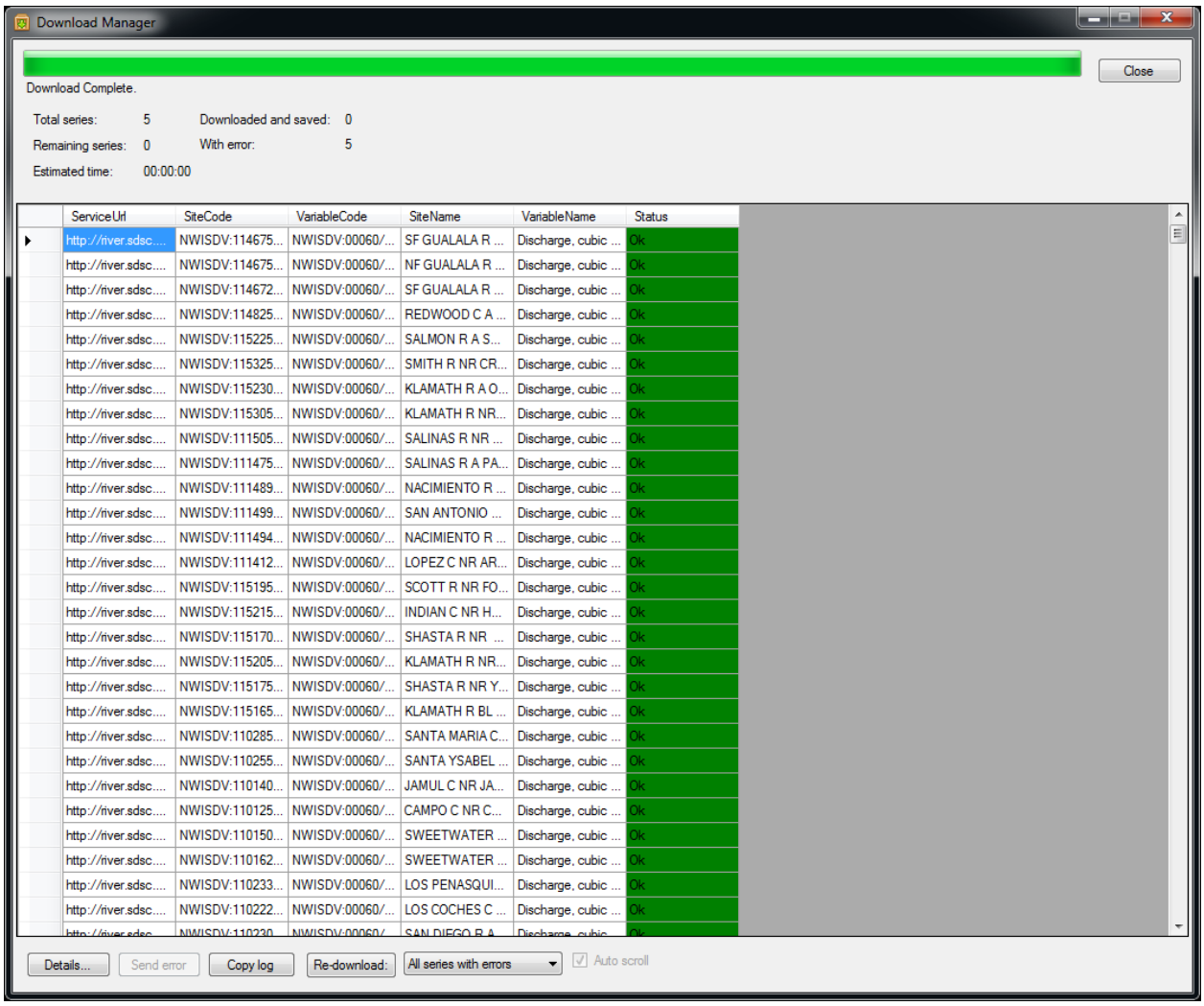


Only 912 of these have NWISDV:00060/DataType=Average as the Var Code. Select these and click on Download Selected.

The downloading part will take a while (likely 2 or 3 hours depending on internet connection speed).



As expected the time series for a few stations failed to download properly. However, clicking on Re-download and specifying All series with errors allows to rerun the download part only for those places for which there was an error. This single feature is crucial when needing to obtain millions of data points as done here.

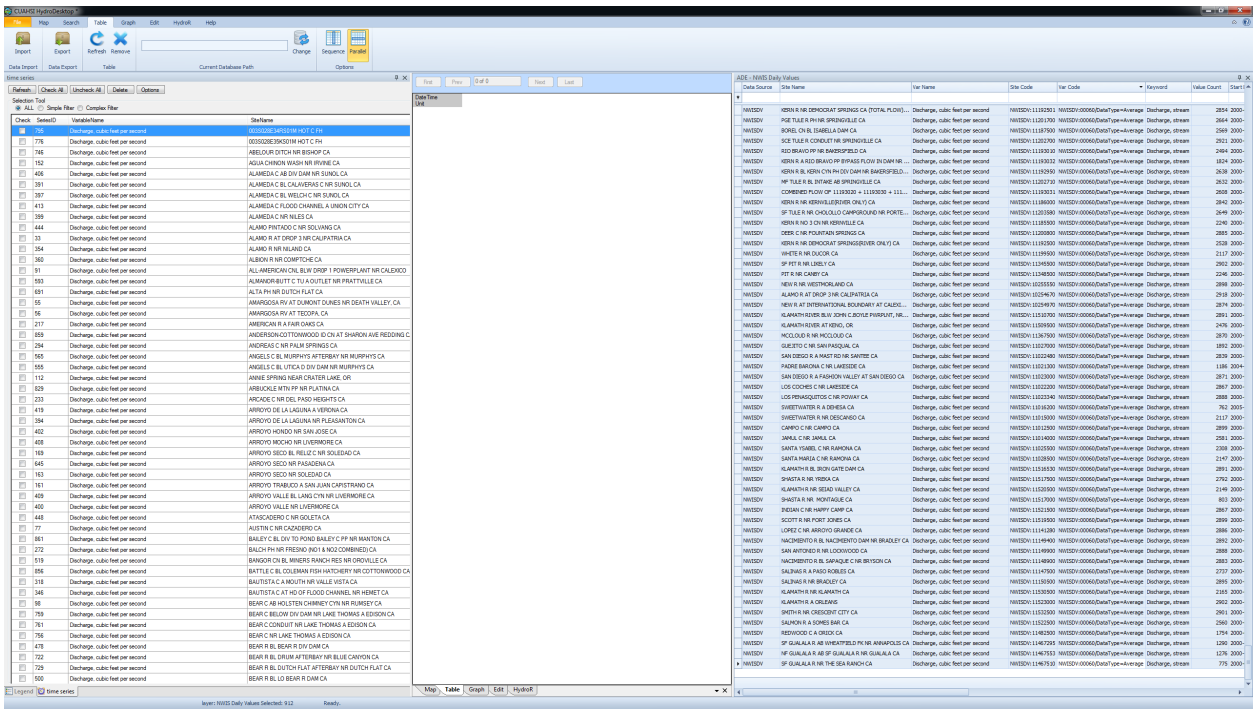


Out of the 912 series that have NWISDV:00060/DataType=Average as the Var Code, and that HydroDesktop selected as potentially having data for 1/1/2000 - 12/31/2007, 5 series consistently have errors in the download process.

The 5 USGS stations corresponding to these series are: 11154700, 11460000, 11120520, 11136500 and 11277100. Checking the data available at http://waterdata.usgs.gov/usa/nwis/ informs that:

- 11154700 actually has data from 10/1/2007 to current. Since this wouldn't be a full record for 1/1/2000 - 12/31/2007 as attempted here, this is not an issue.
- 11460000, 11120520 and 11136500 have their period of record outside of 1/1/2000 - 12/31/2007 so it is normal they would not have any data for our period of interest.
- 11277100 seems to be a station with instantaneous data which could allow to have daily averages for our period of interest, but daily data is not currently directly accessible.

Once the download is complete, you should be in the Table tab.



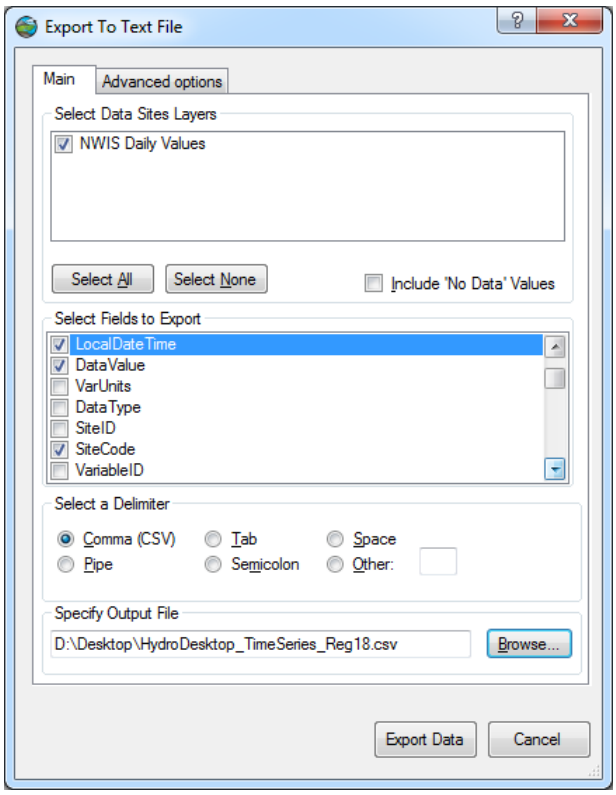
Click on Export.

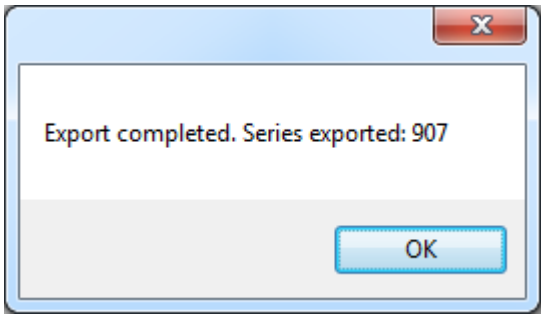
For Select Data Sites Layer, pick NWIS Daily Values.

For Selct Fields to Export, pick LocalDateTime, DataValue and SiteCode.

For Select a Delimiter, pick Comma (CSV).

For Specify Output file, choose HydroDesktop_TimeSeries_Reg18.csv.





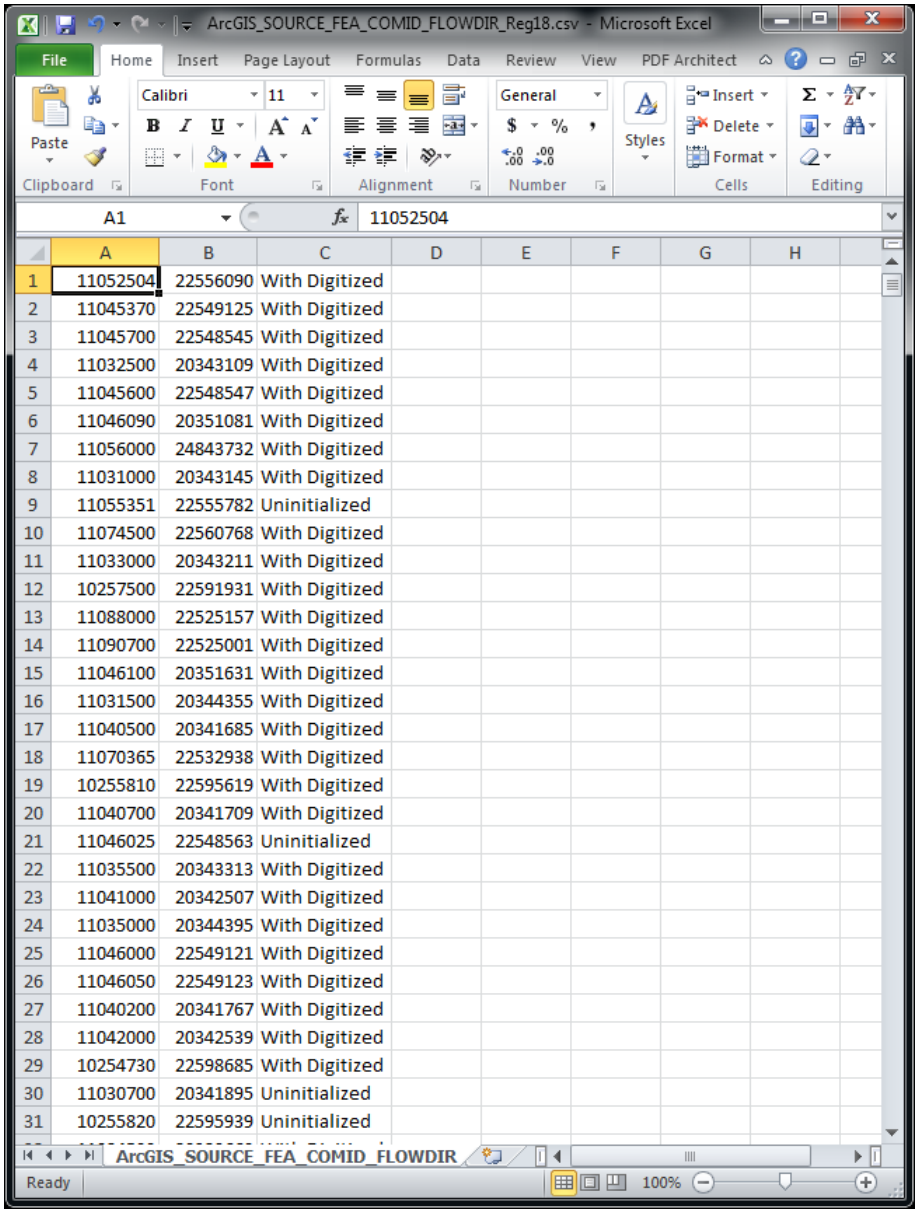
These 907 series exported correspond to the 912 that have NWISDV:00060/DataType=Average minus the 5 series that consistently have errors in the download process. There should be 2,230,656 lines in HydroDesktop_TimeSeries_Reg18.csv. The first line is the header line, followed by 2,230,655 data points.

Prepare .csv files using Excel

RAPID can only run on those reaches which have known flow direction, and therefore we need to select only the stations located on these reaches. The RAPID data model allows for only one gauge per river reach, and therefore one gauge has to be picked in case of duplicates. These two tasks will later be done in a Fortran program for which some input files prepared here will be needed.

From ArcGIS files

Open StreamGageEvent_Reg18_comid.dbf using MS Excel, remove all columns except for SOURCE_FEA, COMID_1 and FLOWDIR. Save the file as ArcGIS_SOURCE_FEA_COMID_FLOWDIR_Reg18.csv. Make sure to keep the field called SOURCE_FEA as text and not as number otherwise Excel will remove the leading zero in the USGS code (when applicable) which is unwanted.



	A	B	C	D	E	F	G	H
1	11052504	22556090	With Digitized					
2	11045370	22549125	With Digitized					
3	11045700	22548545	With Digitized					
4	11032500	20343109	With Digitized					
5	11045600	22548547	With Digitized					
6	11046090	20351081	With Digitized					
7	11056000	24843732	With Digitized					
8	11031000	20343145	With Digitized					
9	11055351	22555782	Uninitialized					
10	11074500	22560768	With Digitized					
11	11033000	20343211	With Digitized					
12	10257500	22591931	With Digitized					
13	11088000	22525157	With Digitized					
14	11090700	22525001	With Digitized					
15	11046100	20351631	With Digitized					
16	11031500	20344355	With Digitized					
17	11040500	20341685	With Digitized					
18	11070365	22532938	With Digitized					
19	10255810	22595619	With Digitized					
20	11040700	20341709	With Digitized					
21	11046025	22548563	Uninitialized					
22	11035500	20343313	With Digitized					
23	11041000	20342507	With Digitized					
24	11035000	20344395	With Digitized					
25	11046000	22549121	With Digitized					
26	11046050	22549123	With Digitized					
27	11040200	20341767	With Digitized					
28	11042000	20342539	With Digitized					
29	10254730	22598685	With Digitized					
30	11030700	20341895	Uninitialized					
31	10255820	22595939	Uninitialized					

From HydroDesktop files

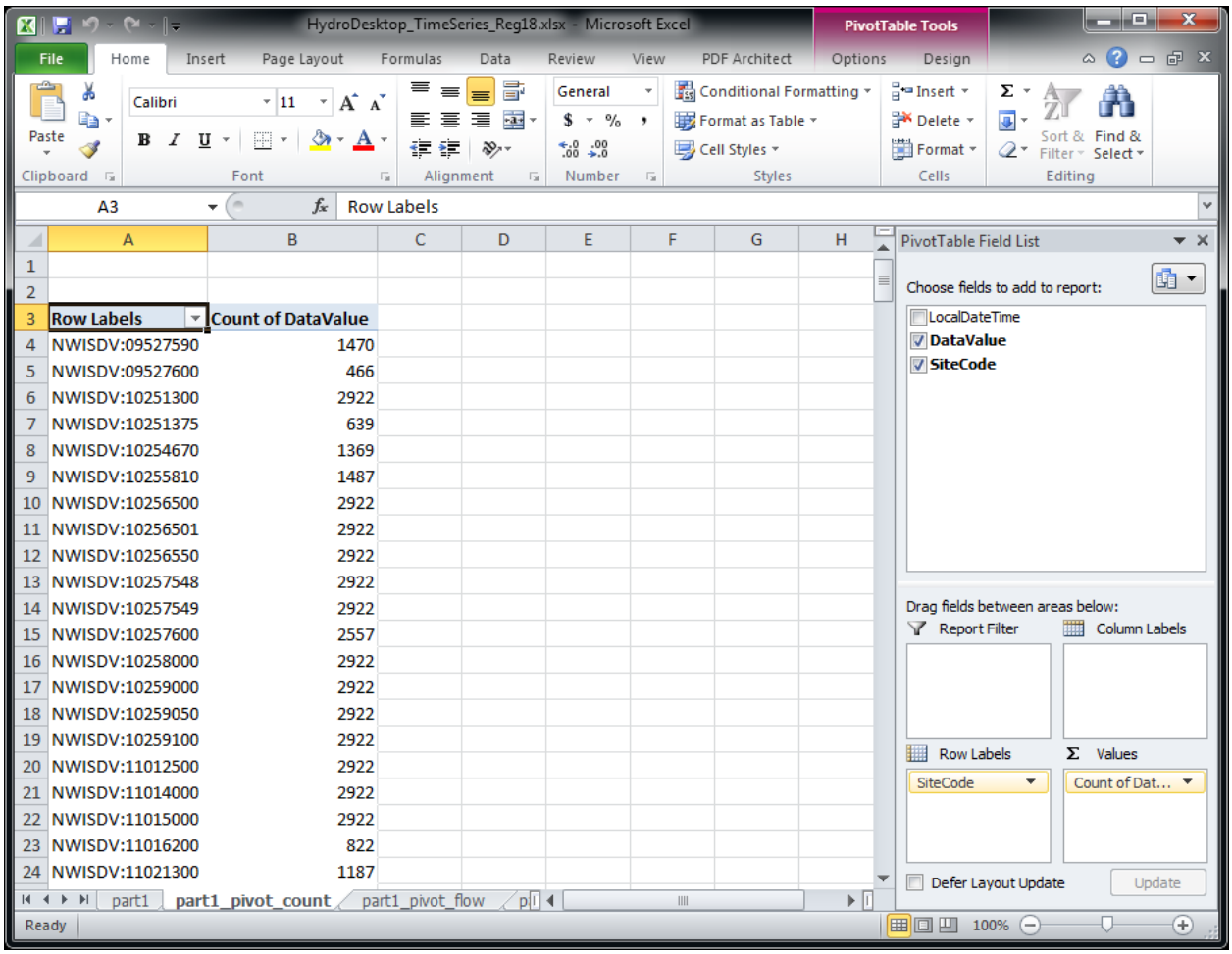
Open HydroDesktop_TimeSeries_Reg18.csv in MS Excel. This file has 2,230,656 lines which turns out to be larger than the limitation of MS Excel (1,048,576 lines).

One can get around this limitation by cutting HydroDesktop_TimeSeries_Reg18.csv in three files using vim (http://www.vim.org/). Each of these files should contain the same header, and less than a total of 1,048,576 lines.

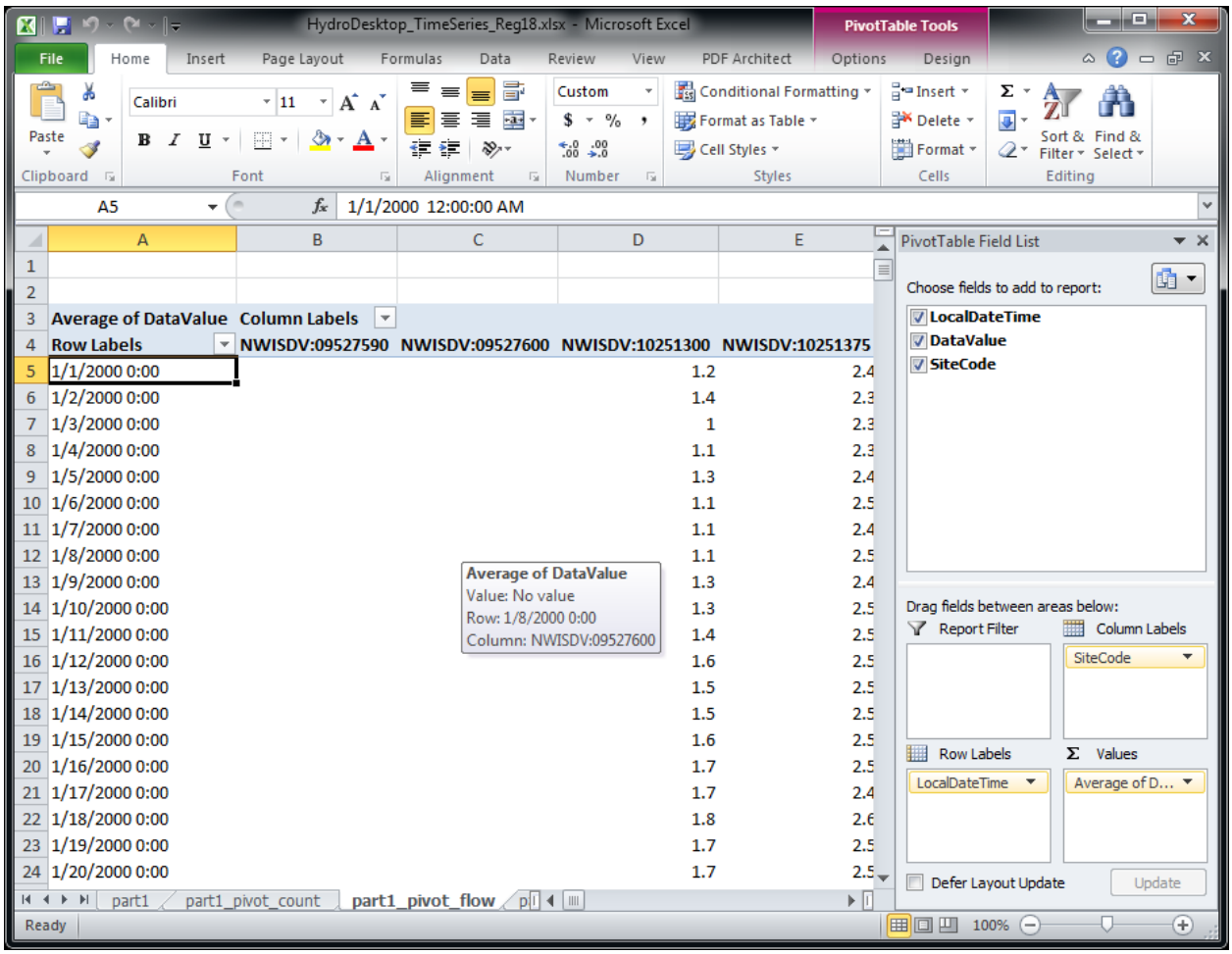
- HydroDesktop_TimeSeries_Reg18_part1.csv
  - First line (line 1): LocalDateTime,DataValue,SiteCode
  - Second line (line 2): 11/18/2000 12:00:00 AM,3.5,NWISDV:11467295
  - Last line (line 801,795): 12/31/2007 12:00:00 AM,0,NWISDV:11045700
- HydroDesktop_TimeSeries_Reg18_part2.csv
  - First line (line 1): LocalDateTime,DataValue,SiteCode
  - Second line (line 801,796): 1/1/2000 12:00:00 AM,0,NWISDV:11070150
  - Last line (line 1,601,487): 12/31/2007 12:00:00 AM,0.88,NWISDV:11062000
- HydroDesktop_TimeSeries_Reg18_part3.csv
  - First line (line 1): LocalDateTime,DataValue,SiteCode
  - Second line (line 1,601,488): 10/1/2007 12:00:00 AM,0.52,NWISDV:11058600
  - Last line (line 2,230,656): 12/31/2007 12:00:00 AM,3.1,NWISDV:11401165

Once these three smaller files are created, open them separately in MS Excel and perform two pivot tables for each. The first pivot table uses SiteCode as Row Labels and count of DataValue as Values. The second pivot table uses SiteCode as Column Labels, LocalDateTime as Row Labels and Average of DataValue as Values.

First pivot table:



The second pivot table:



Part1, Pivot table 1 has 801,794 data points corresponding to 326 USGS stations.
Part1, Pivot table 2 has data points corresponding to 2,922 days (i.e. the total number of days between 1/1/2000 and 12/31/2007).

Part2, Pivot table 1 has 799,692 data points corresponding to 325 USGS stations.
Part2, Pivot table 2 has data points corresponding to 2,922 days (i.e. the total number of days between 1/1/2000 and 12/31/2007).

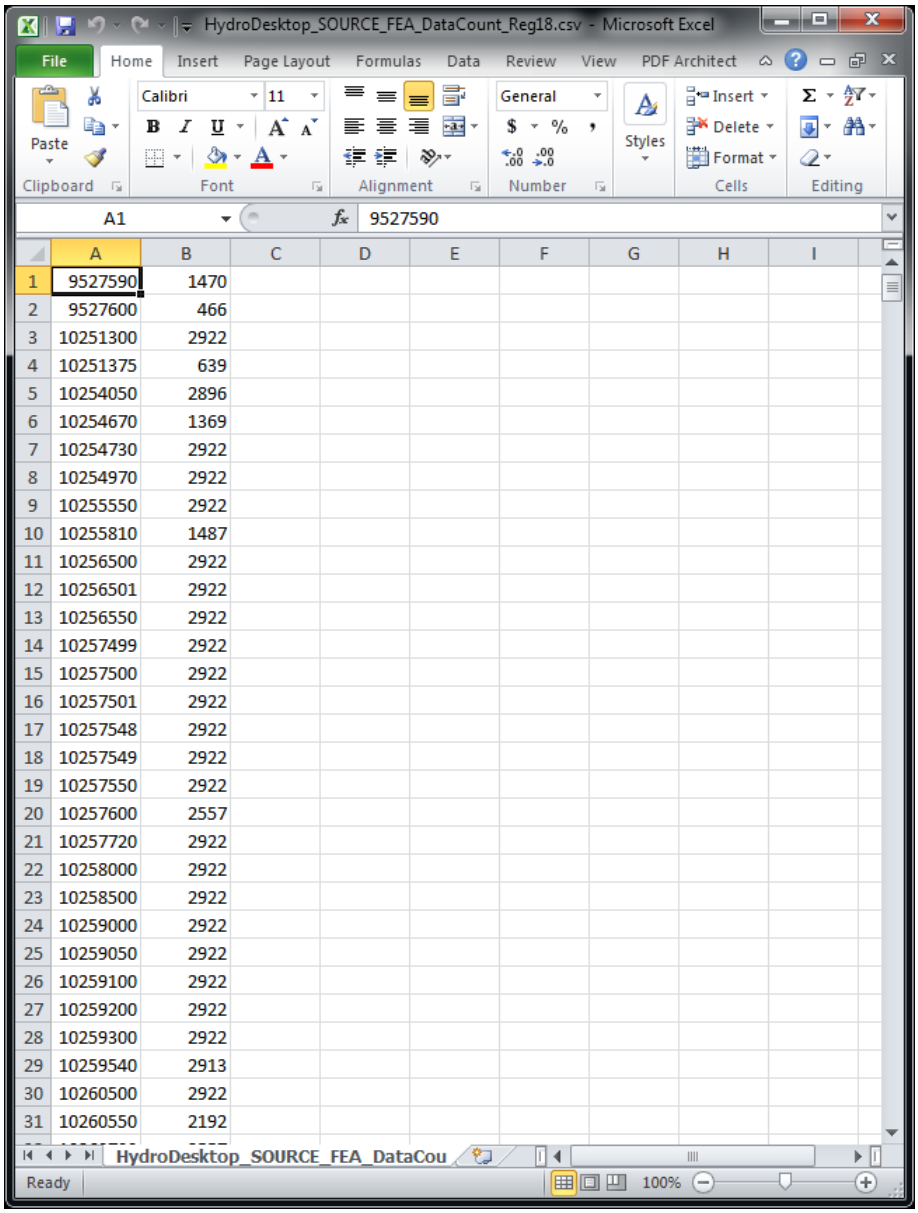
Part3, Pivot table 1 has 629,169 data points corresponding to 256 USGS stations.
Part3, Pivot table 2 has data points corresponding to 2,922 days (i.e. the total number of days between 1/1/2000 and 12/31/2007).

One can check that the sum of the number of stations available in the first pivot table for each one of the three parts is 326+325+256=907, i.e. the total number of stations in the original file.

One can also check that the sum of the number of data points in the first pivot table for each one of the three parts is 801,794+799,692+629,169=2,230,655, i.e the total number of data points in the original file.

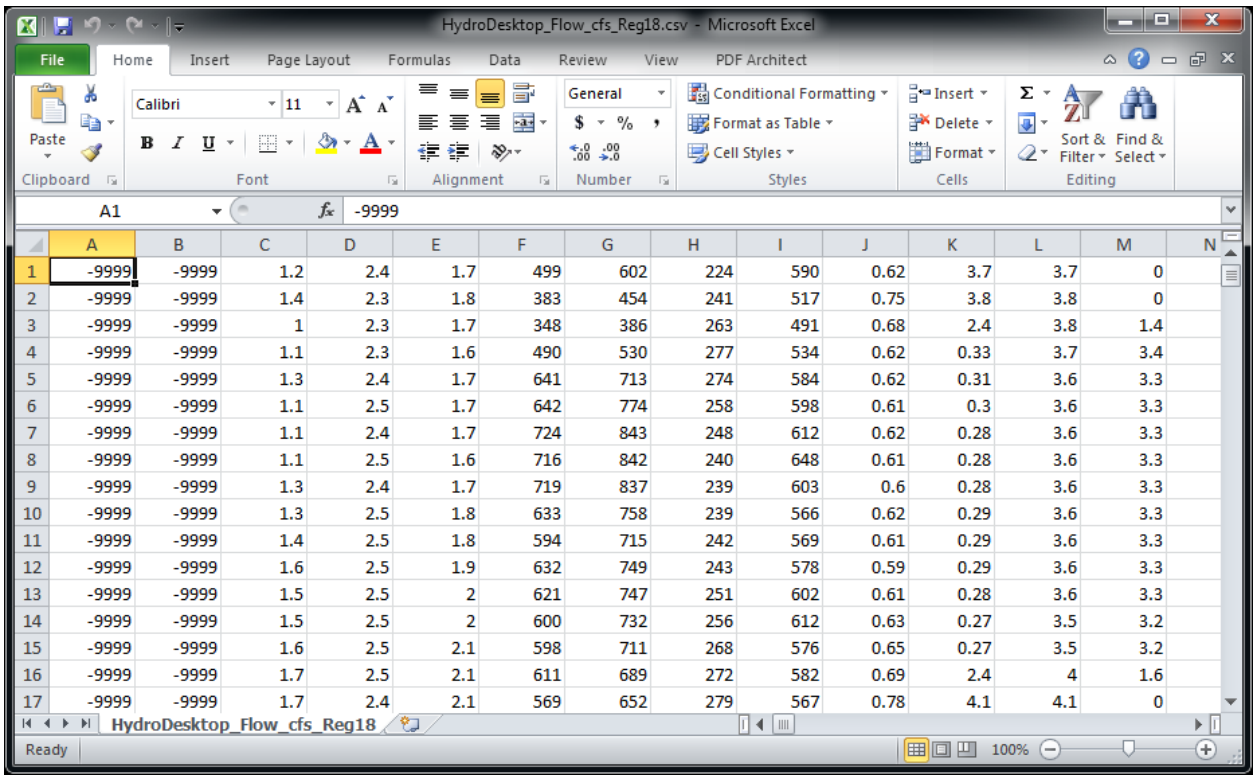
We can now combine pivot table 1 for the three parts into a single file called HydroDesktop_SOURCE_FEA_Datacount_Reg18.csv and pivot table 2 for the three parts into a single file called HydroDesktop_Flow_cfs_Reg18.csv. In the latter file, replace missing values by -9999. Make sure the vertical ordering of HydroDesktop_SOURCE_FEA_Datacount_Reg18.csv is the same as the horizontal ordering of HydroDesktop_Flow_cfs_Reg18.csv. Also make sure the horizontal ordering of HydroDesktop_Flow_cfs_Reg18.csv is chronological.

HydroDesktop_SOURCE_FEA_Datacount_Reg18.csv



	A	B	C	D	E	F	G	H	I
1	9527590	1470							
2	9527600	466							
3	10251300	2922							
4	10251375	639							
5	10254050	2896							
6	10254670	1369							
7	10254730	2922							
8	10254970	2922							
9	10255550	2922							
10	10255810	1487							
11	10256500	2922							
12	10256501	2922							
13	10256550	2922							
14	10257499	2922							
15	10257500	2922							
16	10257501	2922							
17	10257548	2922							
18	10257549	2922							
19	10257550	2922							
20	10257600	2557							
21	10257720	2922							
22	10258000	2922							
23	10258500	2922							
24	10259000	2922							
25	10259050	2922							
26	10259100	2922							
27	10259200	2922							
28	10259300	2922							
29	10259540	2913							
30	10260500	2922							
31	10260550	2192							

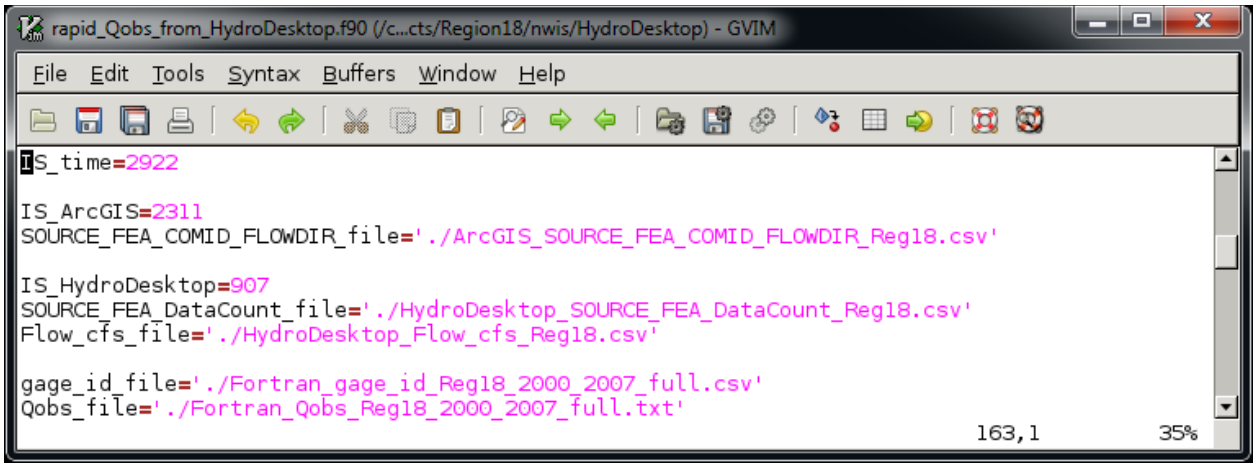
HydroDesktop_Flow_cfs_Reg18.csv



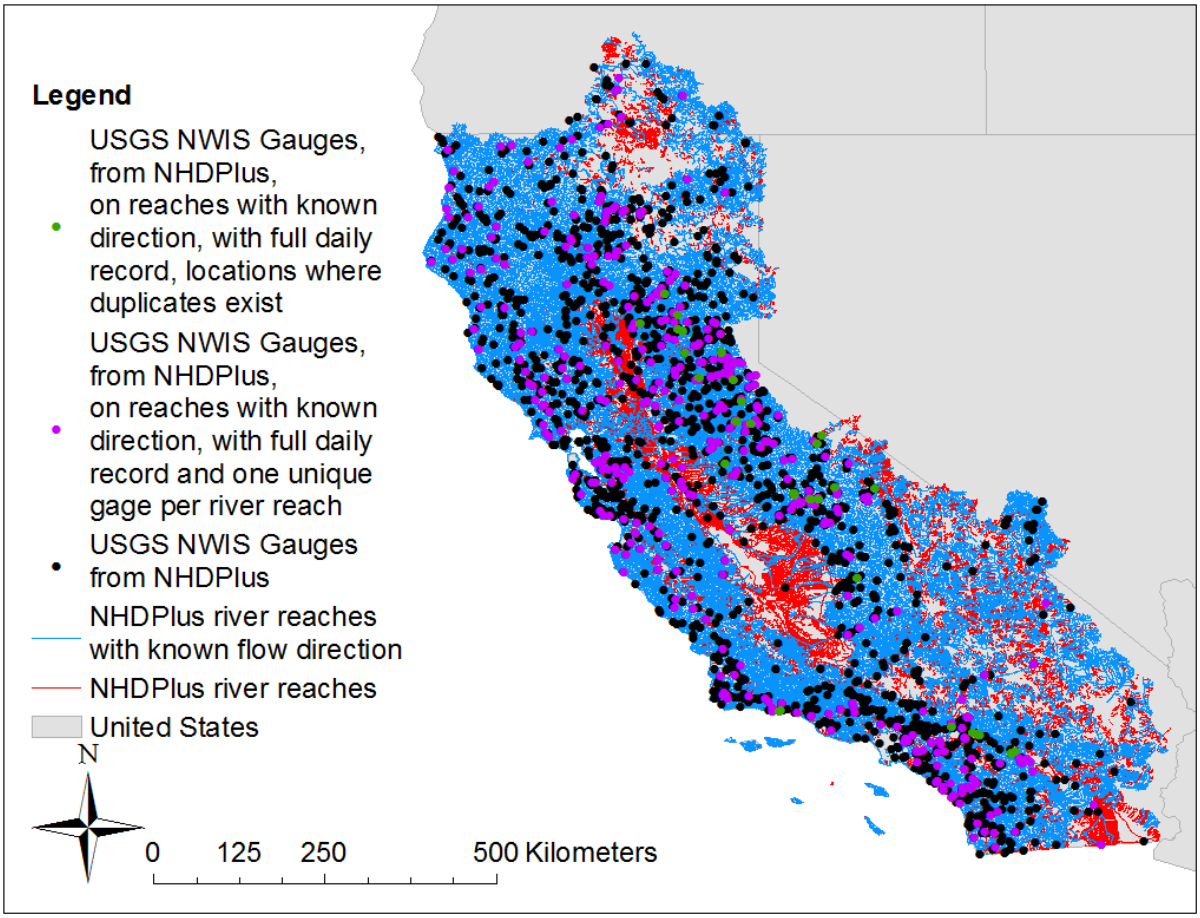
	A	B	C	D	E	F	G	H	I	J	K	L	M
1	-9999	-9999	1.2	2.4	1.7	499	602	224	590	0.62	3.7	3.7	0
2	-9999	-9999	1.4	2.3	1.8	383	454	241	517	0.75	3.8	3.8	0
3	-9999	-9999	1	2.3	1.7	348	386	263	491	0.68	2.4	3.8	1.4
4	-9999	-9999	1.1	2.3	1.6	490	530	277	534	0.62	0.33	3.7	3.4
5	-9999	-9999	1.3	2.4	1.7	641	713	274	584	0.62	0.31	3.6	3.3
6	-9999	-9999	1.1	2.5	1.7	642	774	258	598	0.61	0.3	3.6	3.3
7	-9999	-9999	1.1	2.4	1.7	724	843	248	612	0.62	0.28	3.6	3.3
8	-9999	-9999	1.1	2.5	1.6	716	842	240	648	0.61	0.28	3.6	3.3
9	-9999	-9999	1.3	2.4	1.7	719	837	239	603	0.6	0.28	3.6	3.3
10	-9999	-9999	1.3	2.5	1.8	633	758	239	566	0.62	0.29	3.6	3.3
11	-9999	-9999	1.4	2.5	1.8	594	715	242	569	0.61	0.29	3.6	3.3
12	-9999	-9999	1.6	2.5	1.9	632	749	243	578	0.59	0.29	3.6	3.3
13	-9999	-9999	1.5	2.5	2	621	747	251	602	0.61	0.28	3.6	3.3
14	-9999	-9999	1.5	2.5	2	600	732	256	612	0.63	0.27	3.5	3.2
15	-9999	-9999	1.6	2.5	2.1	598	711	268	576	0.65	0.27	3.5	3.2
16	-9999	-9999	1.7	2.5	2.1	611	689	272	582	0.69	2.4	4	1.6
17	-9999	-9999	1.7	2.4	2.1	569	652	279	567	0.78	4.1	4.1	0

Run a Fortran program

Compile and run the program rapid_Qobs_from_HydroDesktop.f90, the source code as well as the example files used to run it (see below) are available on the RAPID website.



Final results



Further information

RAPID website: http://rapid-hub.org/

RAPID source code: https://github.com/c-h-david/rapid/