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

By Cédric H. David (cedric.david@jpl.nasa.gov),
16 Jun 2011, updated 13 Apr 2012, 05 Jun 2012, 21 Jan 2015

Goal

Download streamflow measurements from USGS NWIS gages 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). Along the way, shapefiles to be used in ArcGIS are also created.

Requirements

The ArcGIS software.
The HydroExcel tool (http://his.cuahsi.org/hydroexcel.html)
The HydroGET tool for ArcGIS (http://his.cuahsi.org/hydroget.html).

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.

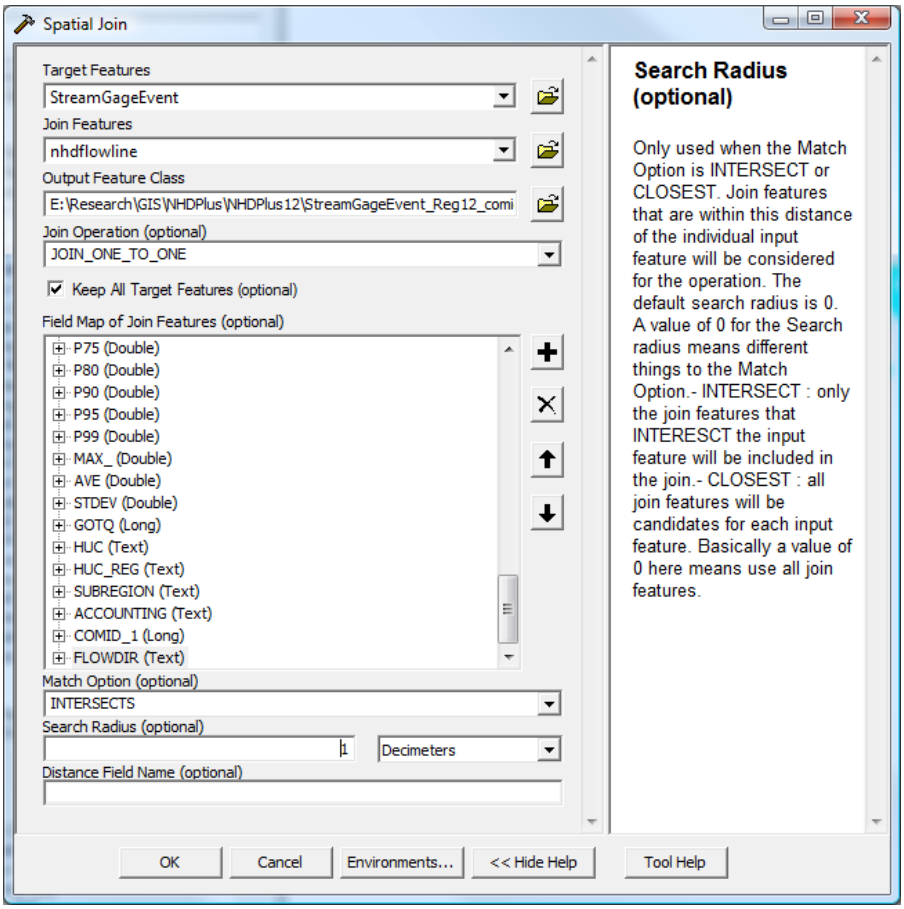
Region 12, 2000-2007

Download the file StreamGageEvent.shp and NHDFlowline.shp for Region 12 from the NHDPlus website (http://www.horizon-systems.com/nhdplus/data.php), rename as StreamGageEvent_Reg12.shp (there are 773 gages in this file) and NHDFlowline_Reg12.shp (there are 74,615 reaches in this file).

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_Reg12.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.shp, add **COMID** and **FLOWDIR** from

© 2015 Cédric H. David
All Rights Reserved

NHDFlowline_Reg12.shp. Save as StreamGageEvent_Reg12_comid.shp. All 773 stations should have a COMID associated to them.

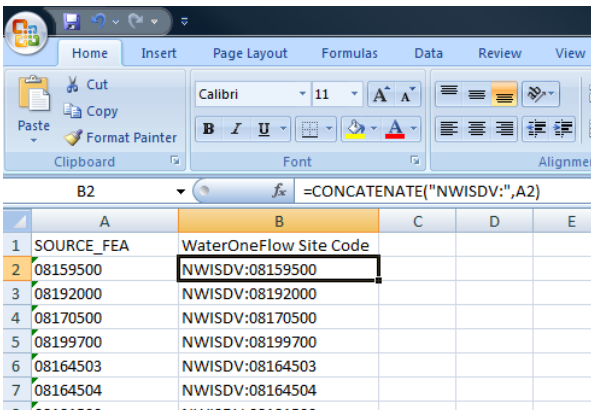


RAPID can only run on those reaches which have known flow direction, and therefore we need to select only the stations located on these reaches. Select the stations that have **FLOWDIR**=”with digitized” and export the shapefile as StreamGageEvent_Reg12_comid_withdir.shp, there should be 756 stations.

One could start downloading data using HydroGET here. However, HydroGET has some issues when downloading data for several years and for several hundreds of stations. It seems that when downloading data for large number of stations HydroGET sometimes interprets lags in response from USGS as errors and skips to the next station. This triggers a variable number of total records downloaded depending on network/webservice status and prevents repeatability of the process. Instead, we’ll use HydroExcel to decrease the number of stations by selecting only the stations that have a potential for full record.

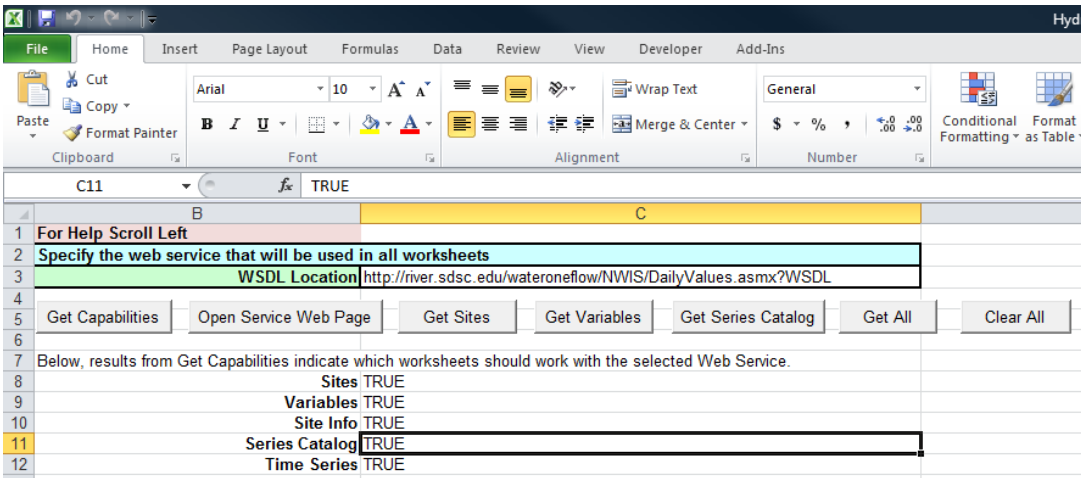
Open StreamGageEvent_Reg12.dbf in Excel, save as .xlsx, clean up to only keep SOURCE_FEA and create the corresponding site codes through concatenation. For Region 12, there should be 773 stations.

© 2015 Cédric H. David
All Rights Reserved

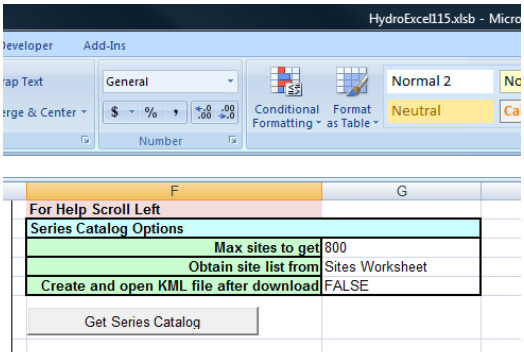


Now open HydroExcel.

Make sure to set the WSDL Location (in the Data Source worksheet) to the WSDL corresponding to USGS the Daily Values Service.

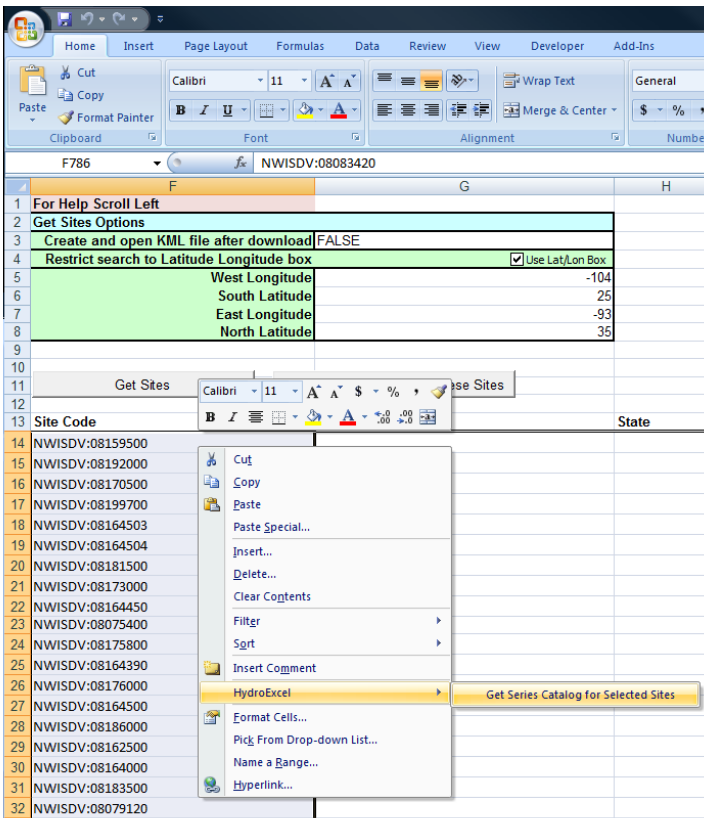


Make sure to change the max number of sites in the “Series Catalog” worksheet to match the number of gages to be used.

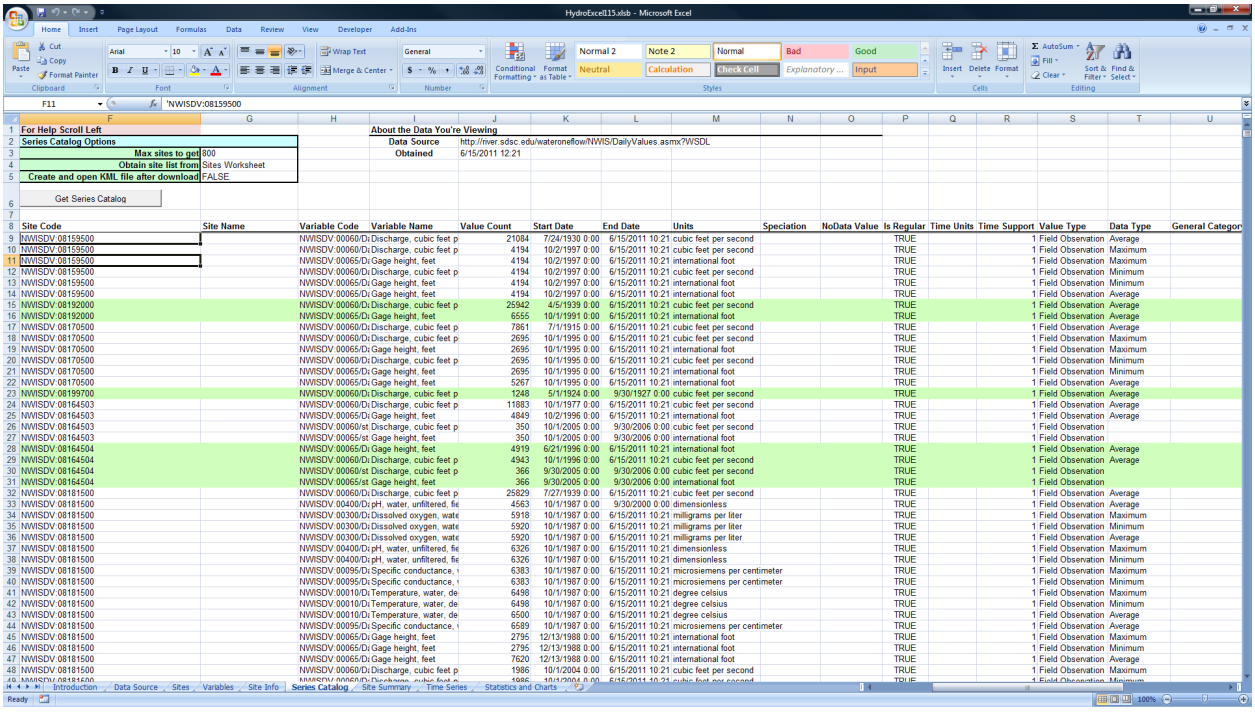


Copy all the site codes in the “Sites” worksheet. Select all of them and “Get Series Catalog for Selected Sites” through right-clicking.

© 2015 Cédric H. David
All Rights Reserved



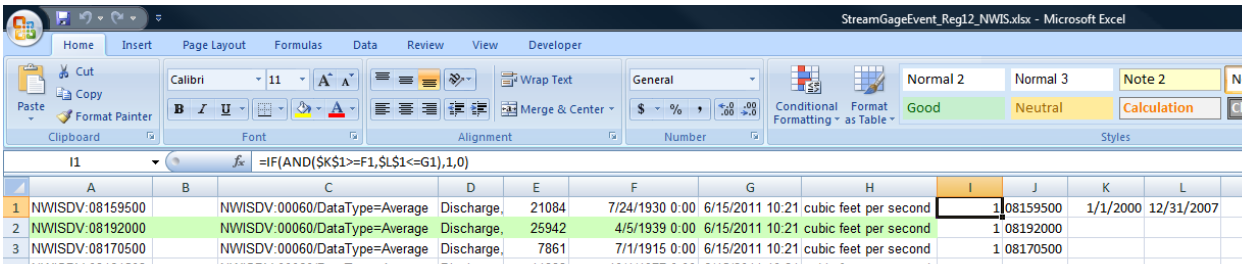
After a few minutes of getting information online, the following is obtained:



Copy this table back in the previous Excel file.

© 2015 Cédric H. David
All Rights Reserved

Back into the old file. Sort the Variable Code field and keep only “NWISDV:00060/DataType=Average”, there should be 618 stations.

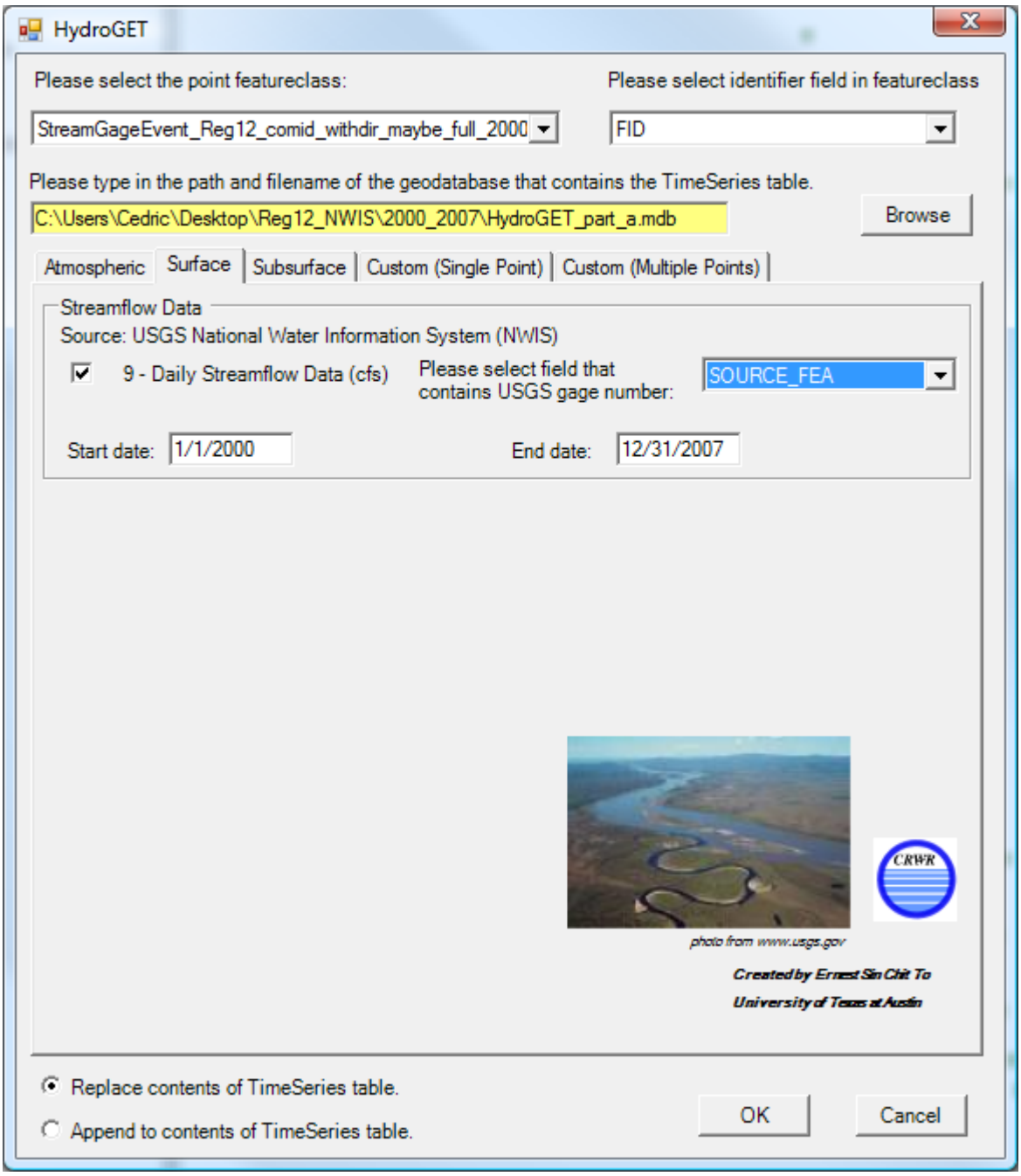


	A	B	C	D	E	F	G	H	I	J	K	L
1	NWISDV:08159500		NWISDV:00060/DataType=Average	Discharge,	21084	7/24/1930 0:00	6/15/2011 10:21	cubic feet per second	1	08159500	1/1/2000	12/31/2007
2	NWISDV:08192000		NWISDV:00060/DataType=Average	Discharge,	25942	4/5/1939 0:00	6/15/2011 10:21	cubic feet per second	1	08192000		
3	NWISDV:08170500		NWISDV:00060/DataType=Average	Discharge,	7861	7/1/1915 0:00	6/15/2011 10:21	cubic feet per second	1	08170500		

Use an “if” statement to figure out which stations have a period that started before 01/01/2000 and ended after 12/31/2007. Only these can hope to have a full record for the period of interest (granted there is no data gaps). There should be 316 stations. Determine the NWIS number of the corresponding stations and select them in a new shapefile called StreamGageEvent_Reg12_comid_withdir_maybe_full_2000_2007.shp. There should be 315 stations (one of the stations that may be full is not on a reach with known flow direction: 08067070).

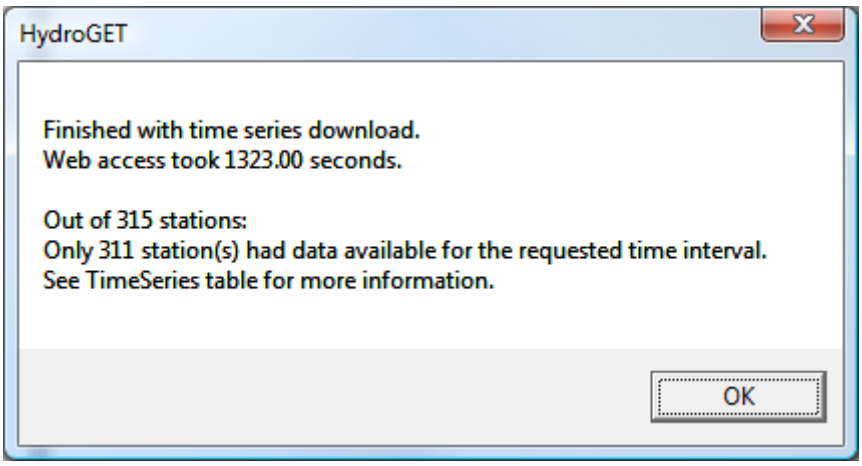
Now use HydroGET to download data. Run HydroGET (pause the visualization in ArcMap prior to starting HydroGET for faster run). In HydroGET, select **FID** as the identifier field in feature class, and **SOURCE_FEA** for the field that contains the USGS gage number (the NWIS ID for each gage is stored in the feature class called SOURCE_FEA). FID is better than COMID for this procedure because if two stations are on the same reach and one is retired, there will be more records than days on the COMID (2923 or more datapoints instead of 2922 for 8 years).

© 2015 Cédric H. David
All Rights Reserved

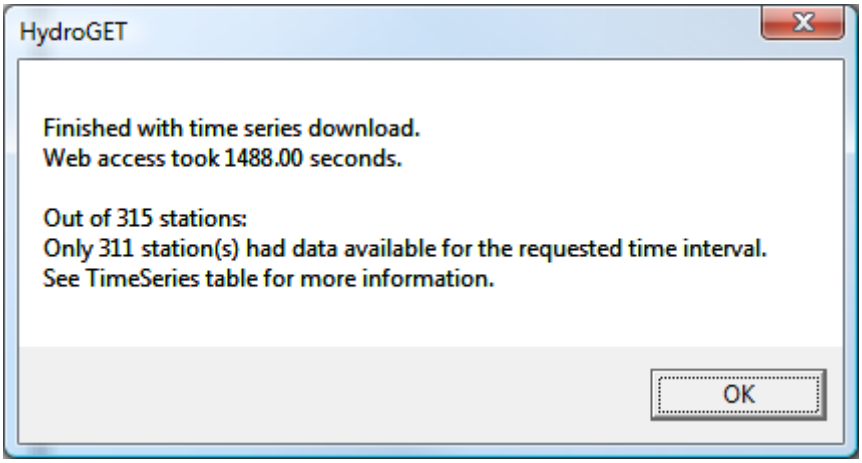


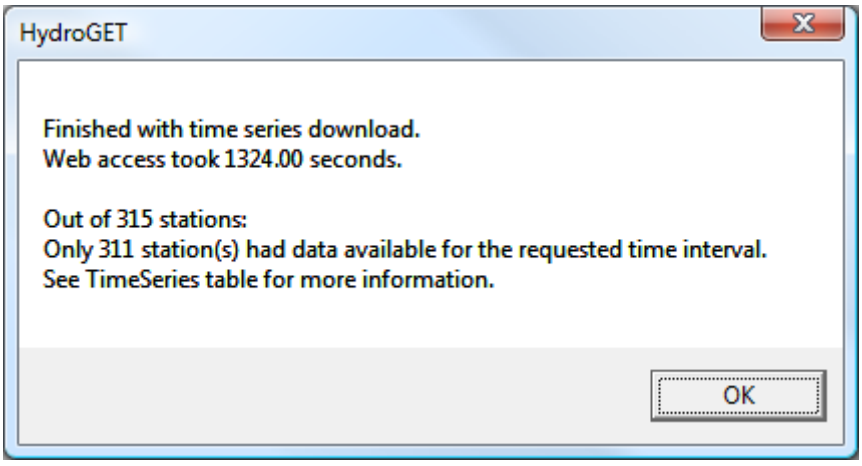
Once HydroGET is done, a message box should pop up (see below). The results of HydroGET are that out of the 315 gages in StreamGageEvent_Reg12_comid_withdir_maybe_full_2000_2007.shp, 311 stations have data and that the download time was 1323 seconds.

© 2015 Cédric H. David
All Rights Reserved



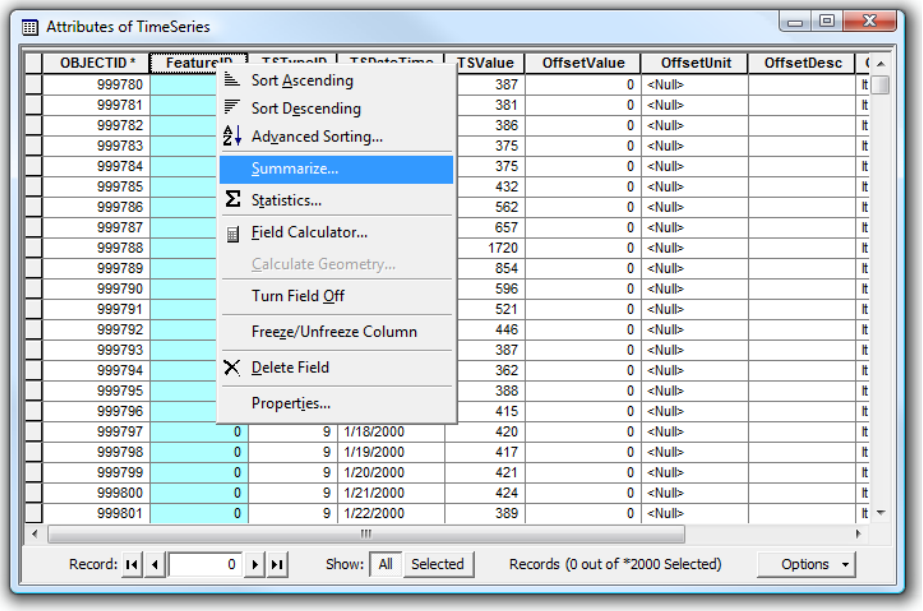
HydroGET downloaded all measurements in a table called TimeSeries which is located inside the HydroGET.mdb file. Open TimeSeries table and add the table to ArcGIS. One can see the total of records downloaded by looking at the number of records in the table. In this particular case 843,764 data points were downloaded. This process was repeated 3 times on 15-16 June 2011 and the same number of records was obtained each time:

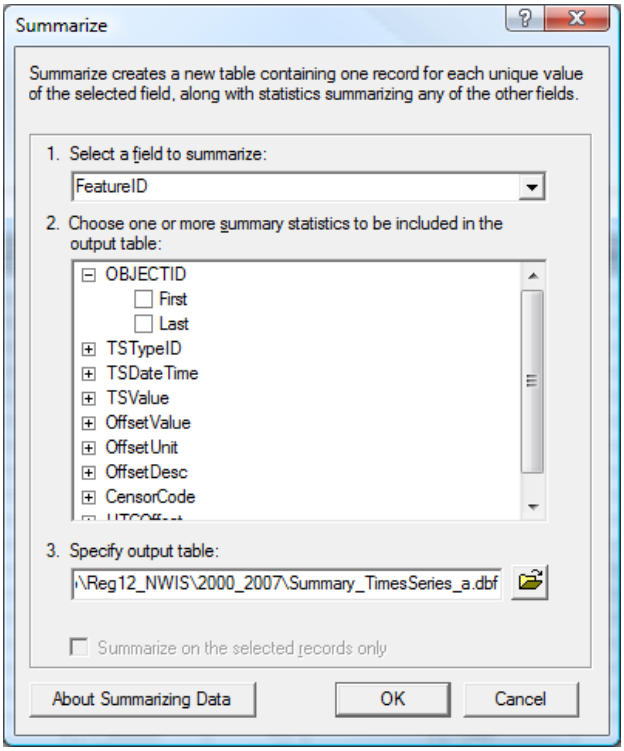




© 2015 Cédric H. David
All Rights Reserved

Let's summarize the TimeSeries table to see how many records were downloaded for each gaging station. Open TimeSeries table in ArcGIS and summarize the FeatureID field with the average of TSValue, save as Summary_TimesSeries_a.dbf.

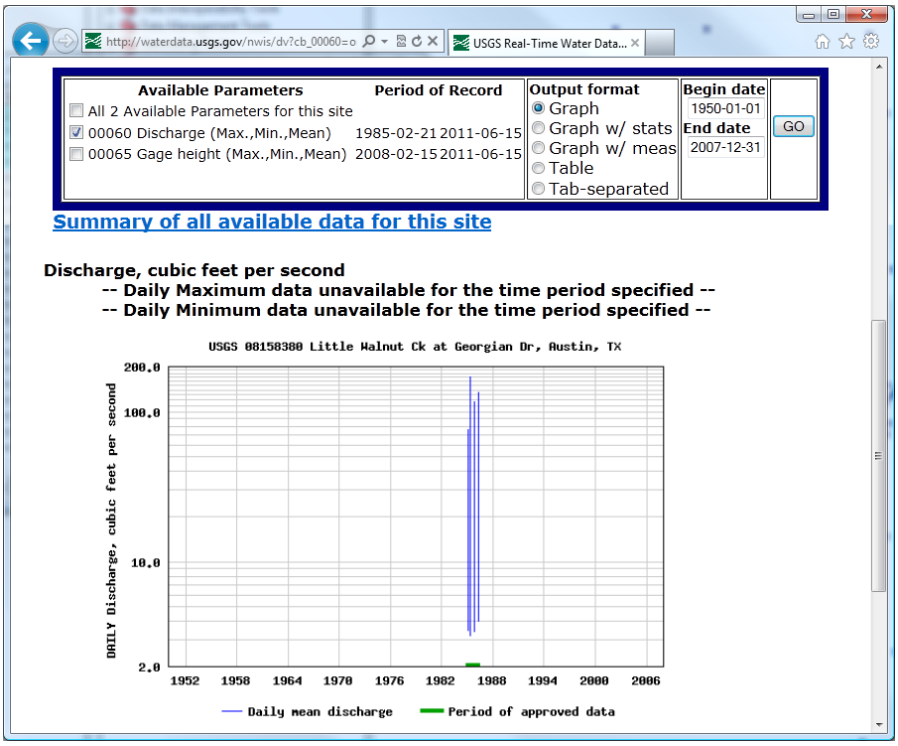




Join StreamGageEvent_Reg12_comid_maybe_full_2000_2007.shp with Summary_TimesSeries_a.dbf based on FID and FeatureID. One can also check on the USGS NWIS website that all 4 stations that

© 2015 Cédric H. David
All Rights Reserved

didn't have data (08158380, 08110000, 08130500 and 08052780) have a complete data gap between 2000-01-01 and 2007-12-31. Following is an example for 08158380:

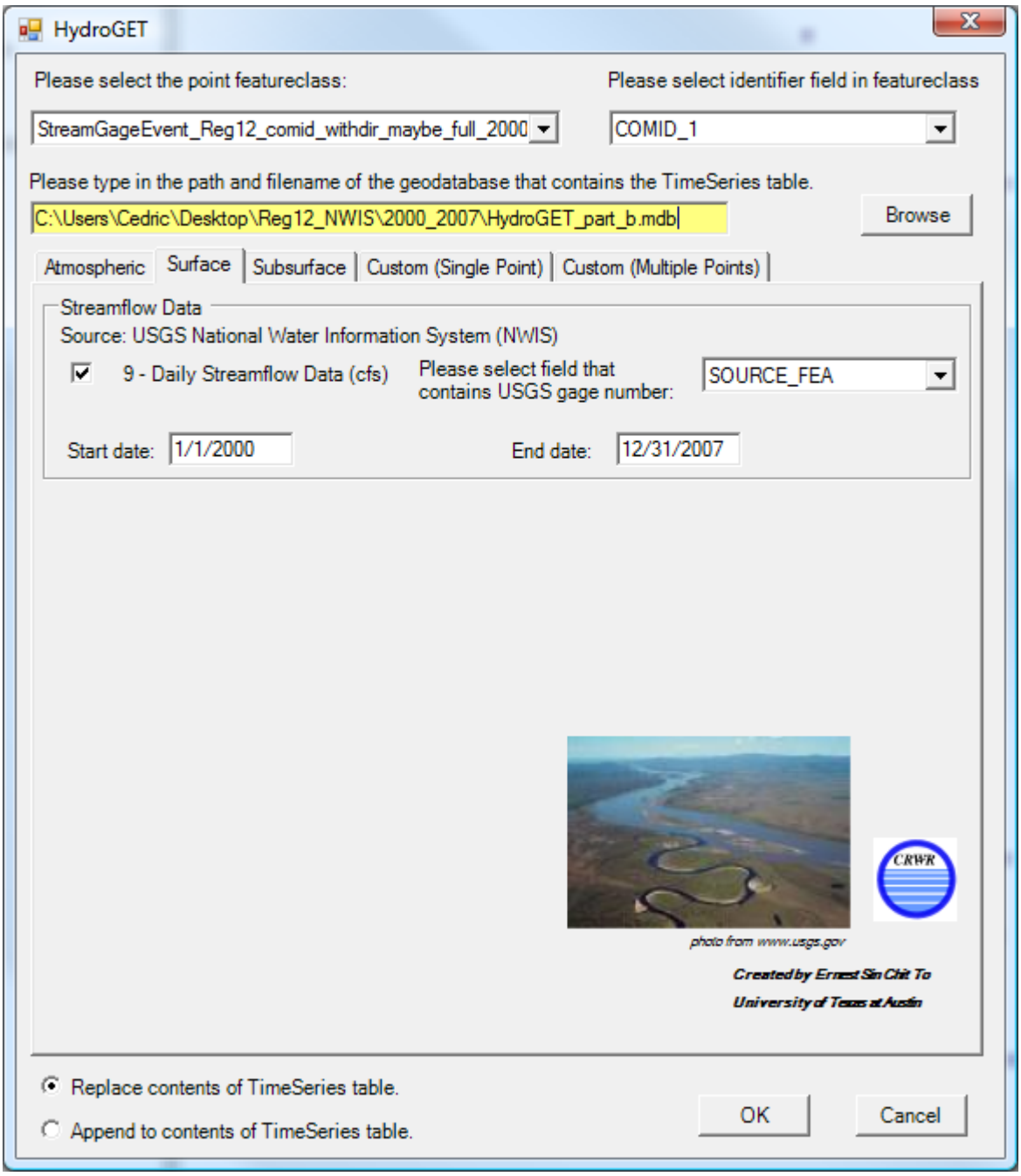


It is important to check that the stations with no data actually have no data on the USGS website. Otherwise, it may be that HydroGET skipped the stations due to network response time.

Select the gages that have a full record (Count_FeatureID=2922), there should be 248 stations. Remove all joins, export as StreamGageEvent_Reg12_comid_full_2000_2007.shp. In this case, all 248 stations are located on separate river reaches so one doesn't have to remove stations to fit the data model used in RAPID.

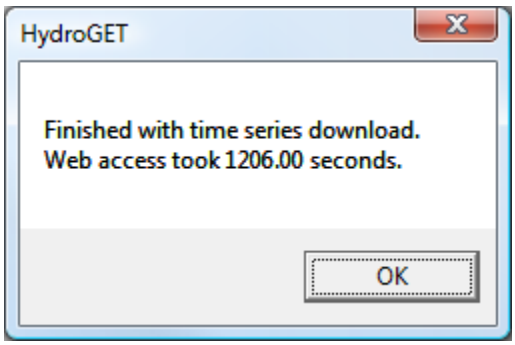
Use HydroGET again, but here the first field is **COMID_1** as the identifier field in feature class, and **SOURCE_FEA** for the field that contains the USGS gage number.

© 2015 Cédric H. David
All Rights Reserved



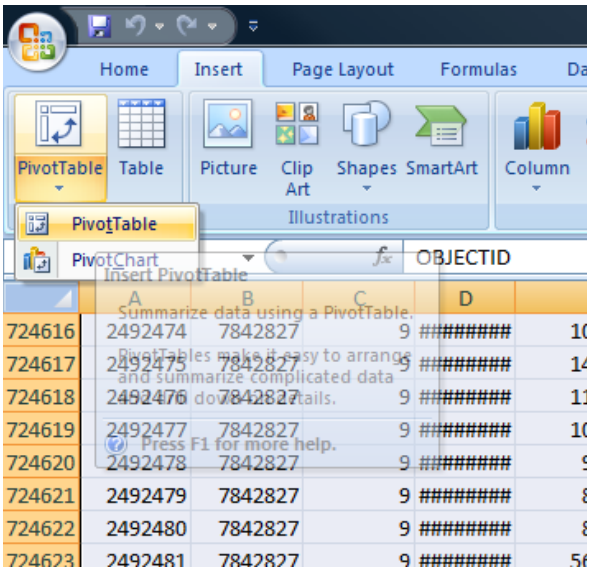
There should be **no stations without data this time.**

© 2015 Cédric H. David
All Rights Reserved

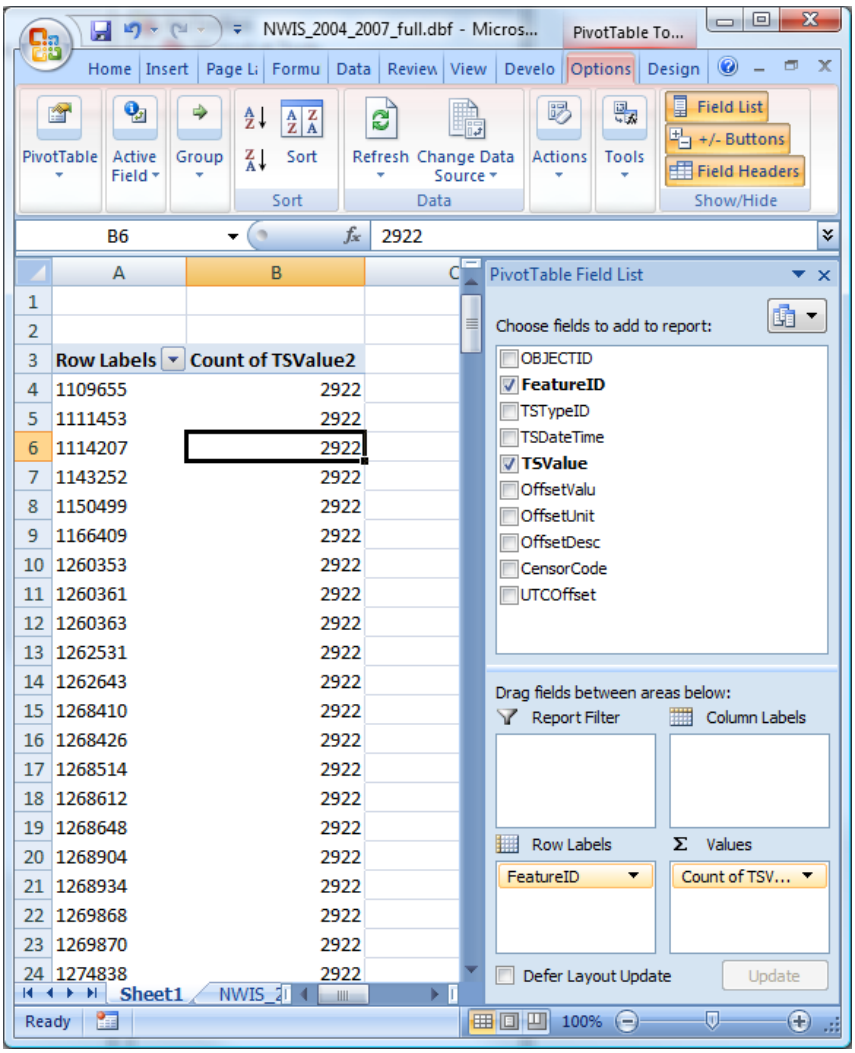


One can check that 724,656 records (248 x 2922) were downloaded, as expected!!!

Save TimeSeries table as NWIS_2004_2007_full.dbf. Open in Excel and save as xlsx. **Make sure the number of rows does not exceed the maximum** (1,048,576 in Excel 2007). In this case 724,656 records are fine. Do a first pivot table to check that all records are complete.

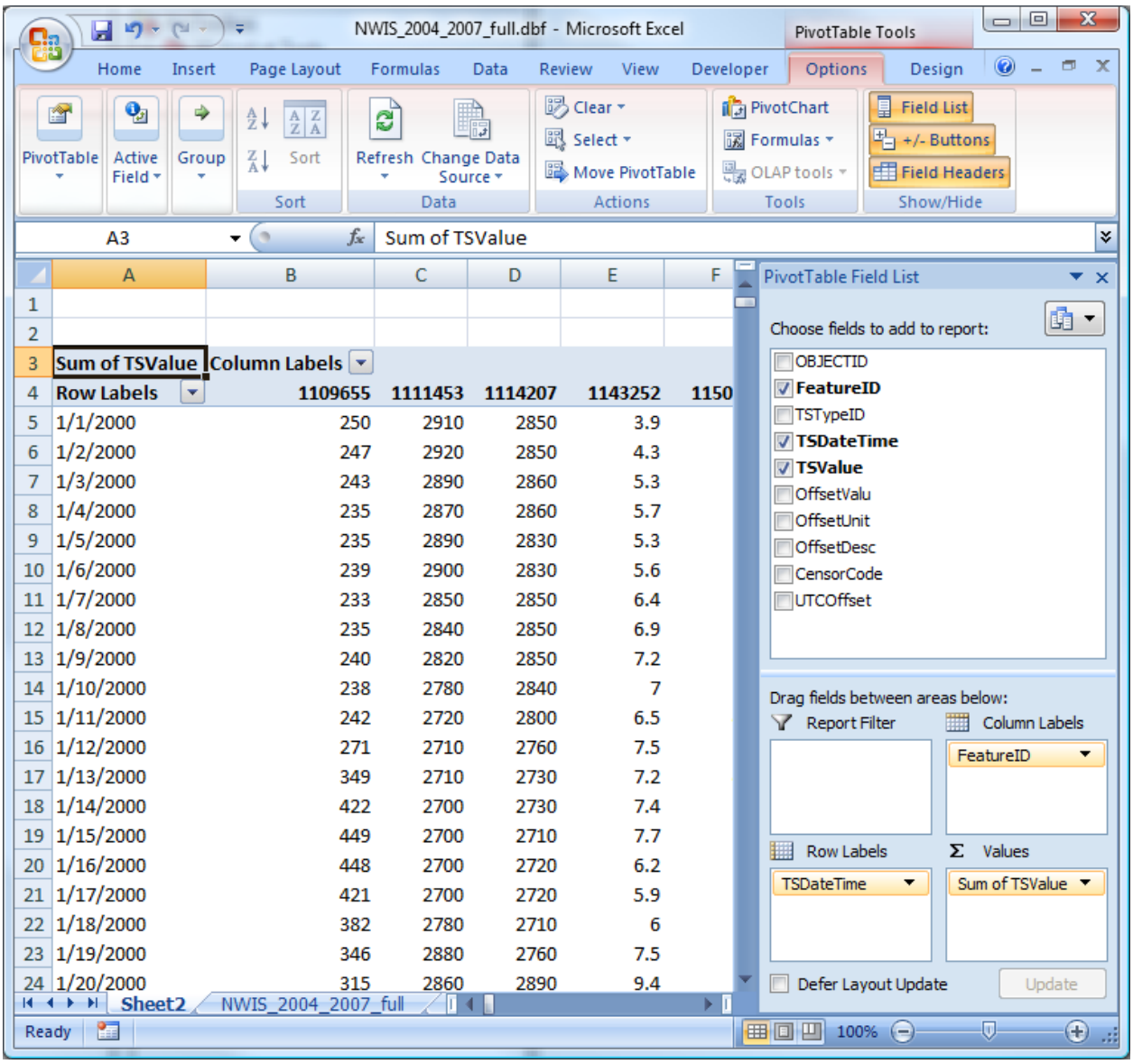


© 2015 Cédric H. David
All Rights Reserved



No do another pivot table for observed streamflow:

© 2015 Cédric H. David
All Rights Reserved



Sum of TSValue	Column Labels				
Row Labels	1109655	1111453	1114207	1143252	1150
1/1/2000	250	2910	2850	3.9	
1/2/2000	247	2920	2850	4.3	
1/3/2000	243	2890	2860	5.3	
1/4/2000	235	2870	2860	5.7	
1/5/2000	235	2890	2830	5.3	
1/6/2000	239	2900	2830	5.6	
1/7/2000	233	2850	2850	6.4	
1/8/2000	235	2840	2850	6.9	
1/9/2000	240	2820	2850	7.2	
1/10/2000	238	2780	2840	7	
1/11/2000	242	2720	2800	6.5	
1/12/2000	271	2710	2760	7.5	
1/13/2000	349	2710	2730	7.2	
1/14/2000	422	2700	2730	7.4	
1/15/2000	449	2700	2710	7.7	
1/16/2000	448	2700	2720	6.2	
1/17/2000	421	2700	2720	5.9	
1/18/2000	382	2780	2710	6	
1/19/2000	346	2880	2760	7.5	
1/20/2000	315	2860	2890	9.4	

Copy and transpose the list of COMIDs (the first row) into a new file called gage_id_Reg12_2000_2007_full.csv. Copy all observed streamflow in a new csv file called Qobs_Reg12_2000_2007_full.csv, multiply each measurement by 0.3048^3 to convert from ft^3/s to m^3/s.

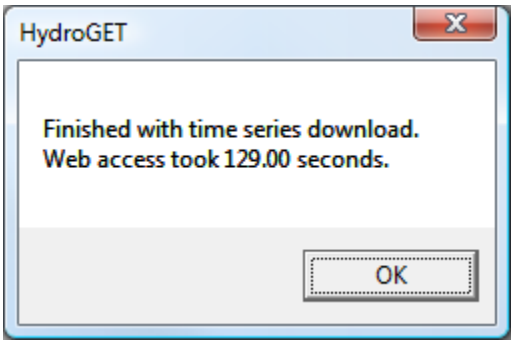
That's it!

© 2015 Cédric H. David
All Rights Reserved

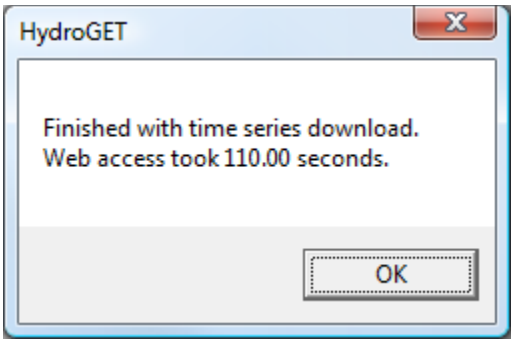
San Antonio and Guadalupe Basins, 2004-2007

The procedure here is similar to the one presented above, except for an additional step that consists in selecting the gages that are located in the San Antonio and Guadalupe Basins prior to running HydroExcel and HydroGET in order to save time. From the file StreamGageEvent.shp for Region 12 (there are 773 gages in this file), select the gages that intersect with subbasin_San_Guad.shp (there should be 84), and export selected features as StreamGageEvent_San_Guad.shp. Associate each of the 84 stations to an NHDPlus **COMID** and **FLOWDIR** through Spatial Join using 1 decimeter threshold. 81 stations should be on reaches with known flow direction.

Out of the 84 stations located in the San Antonio and Guadalupe Basins, 43 potentially have a full record between 2004-01-01 and 2007-12-31 (determined with HydroExcel). All of these are gages located on reaches with known flow direction. All of these actually have data (60,977 data points downloaded with HydroGET):



This process was repeated 3 times on 16 June 2011 and the same number of data points was downloaded each time. There should be 36 stations with full record (1461 days), for a total of 52596 records.



© 2015 Cédric H. David
All Rights Reserved

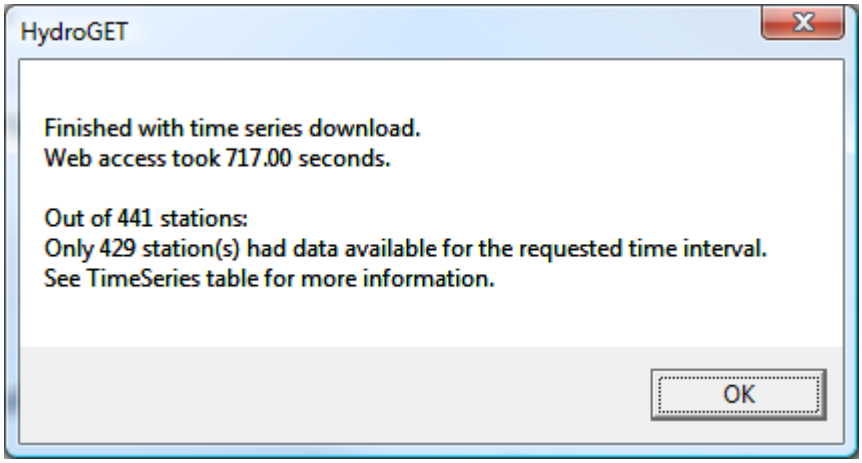
Region 07, 2004

The procedure here is identical to the first one presented in this document. From the file StreamGageEvent.shp for Region 07, associate each of the 1,277 stations to an NHDPlus **COMID** and **FLOWDIR** through Spatial Join using 1 decimeter threshold. 1,251 stations should be on reaches with known flow direction. Out of these 1,251 stations :

- 165 stations have "Site Info Error: No site info was returned" in HydroExcel. These sites do not have daily values recorded.
- 1,005 have NWIS daily value for average discharge from HydroExcel.
- Others have daily values recorded but not for average discharge from HydroExcel.
- If the message "Site Info Error: An error occurred while calling the service. Timeout expired. The timeout period elapsed prior to completion of the operation or the server is not responding." appears in HydroExcel, rerun the Series Catalog part for these stations.

Out of the 1,005 stations that have NWIS daily values for average discharge, 441 may have a full record for the year 2004.

Using HydroGET on these 441 stations leads to:

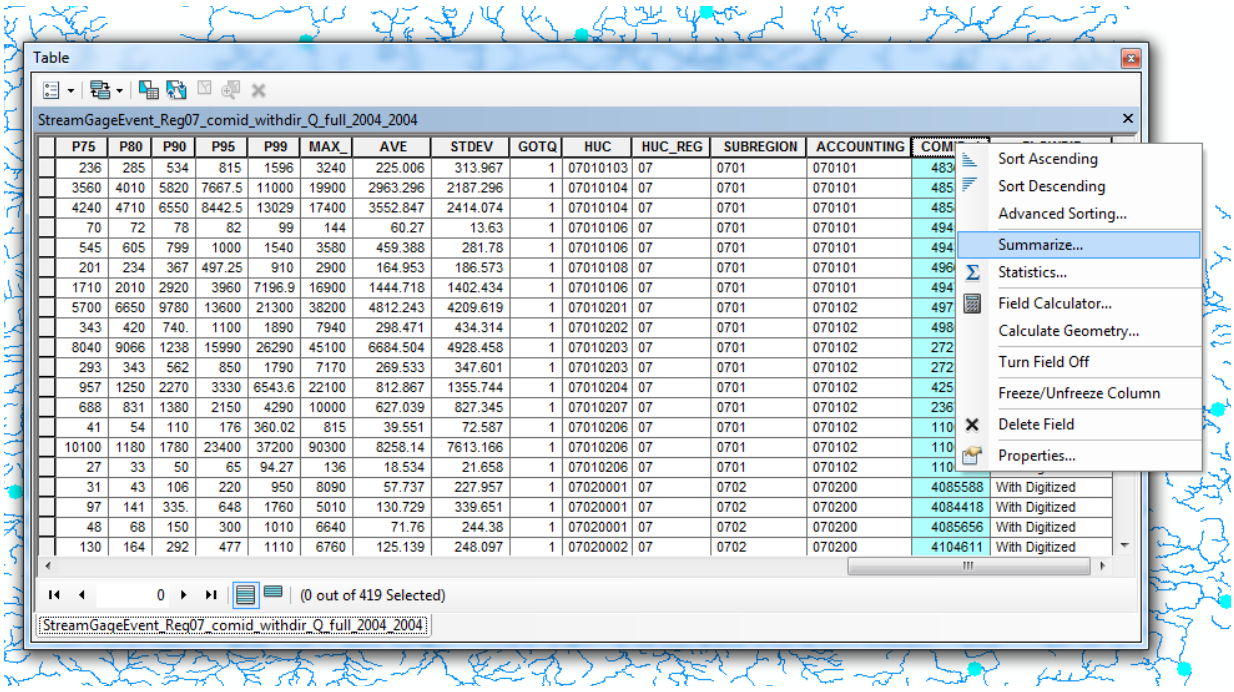


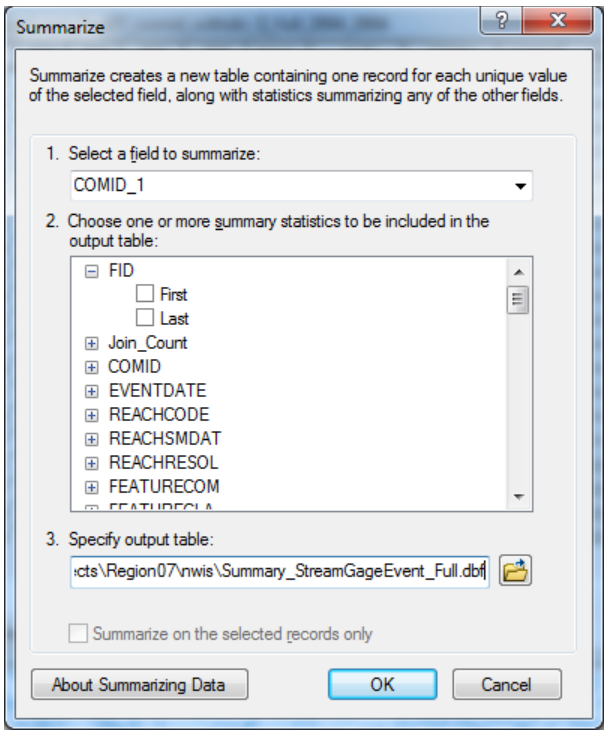
155,229 data points in the TimeSeries downloaded with HydroGET (checked with two separate downloads on 13 April 2012 that results are the same). Summarizing the TimeSeries table obtained by HydroGET based on Feature ID (Summary_TimeSeries_a.dbf), one can check that out of these 441 stations:

- 12 stations have no data at all
- 10 stations have partial data
- 419 stations have full record for 2004

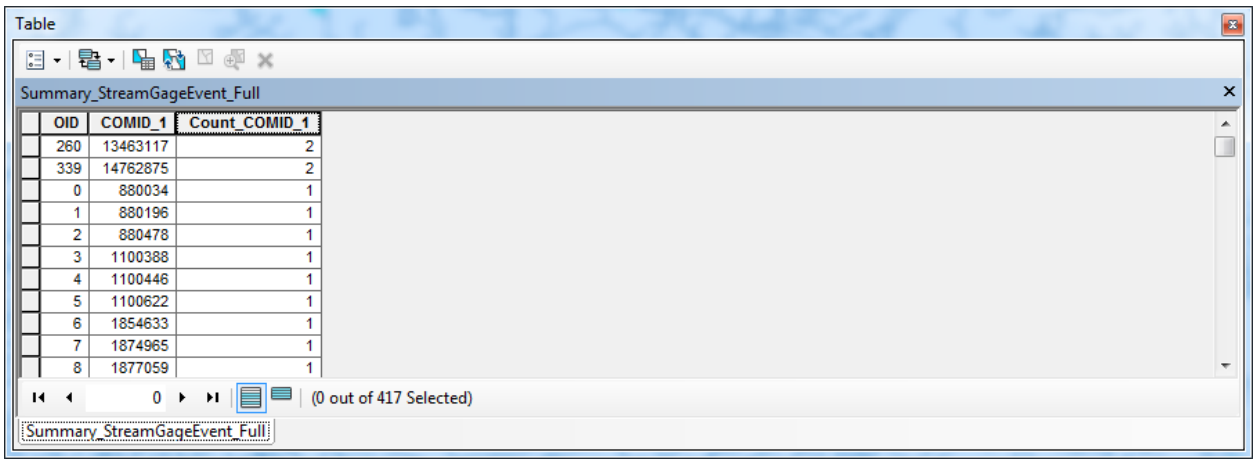
Let's make sure that each river reach has a maximum of 1 station on them to follow the data model in RAPID. Select all stations with full data record and export in a new shapefile called "StreamGageEvent_Reg07_comid_withdir_Q_full_2004_2004.shp". Summarize the COMID_1 field:

© 2015 Cédric H. David
All Rights Reserved

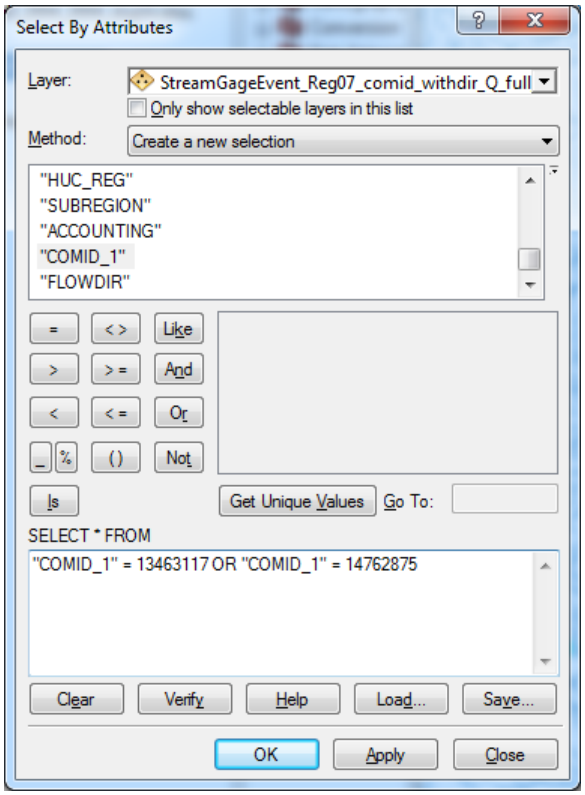




© 2015 Cédric H. David
All Rights Reserved



River reaches located at COMID=13463117 and at COMID=14762875 have two stations with full one-year data record each. Let's see where these stations are located:



© 2015 Cédric H. David
All Rights Reserved

Table

StreamGageEvent_Reg07_comid_withdir_Q_full_2004_2004

SOURCE_FEA	FEATUREDET
05535000	http://waterdata.usgs.gov/nwis/nwisman/?site_no=05535000
05535070	http://waterdata.usgs.gov/nwis/nwisman/?site_no=05535070
05551675	http://waterdata.usgs.gov/nwis/nwisman/?site_no=05551675
05551700	http://waterdata.usgs.gov/nwis/nwisman/?site_no=05551700

1 (4 out of 419 Selected)

StreamGageEvent_Reg07_comid_withdir_Q_full_2004_2004

Table

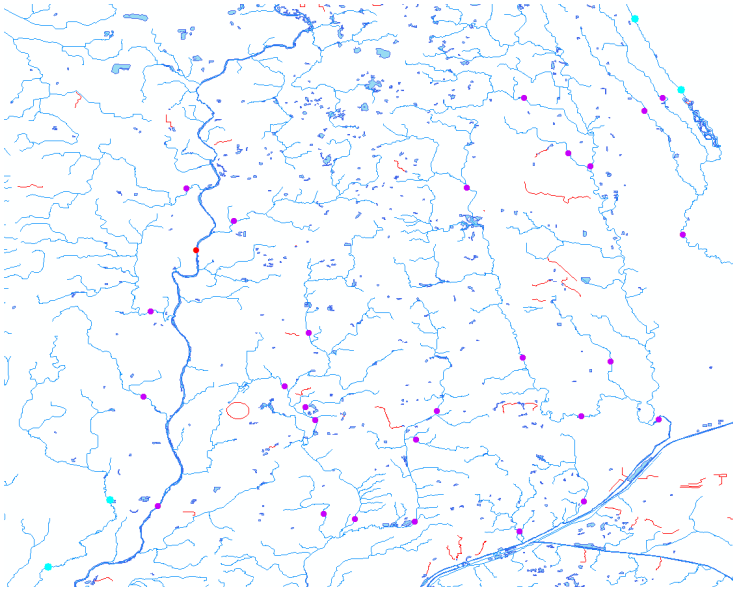
StreamGageEvent_Reg07_comid_withdir_Q_full_2004_2004

SUBREGION	ACCOUNTING	COMID_1	FLOWDIR
0712	071200	13463117	With Digitized
0712	071200	13463117	With Digitized
0712	071200	14762875	With Digitized
0712	071200	14762875	With Digitized

1 (4 out of 419 Selected)

StreamGageEvent_Reg07_comid_withdir_Q_full_2004_2004

Stations 05535000 and 05535070 are on the same river reach (COMID=13463117), so are stations 05551675 and 05551700 (COMID=14762875).

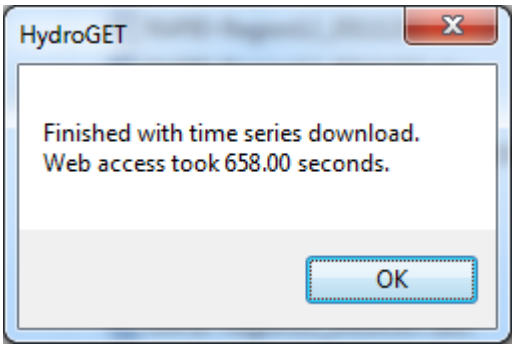


These stations don't seem to be influenced by reservoirs or divergences. In the data model used for RAPID, only one gage can be associated to a COMID. The stations that are furthest downstream in each river reach (05535070 and 05551700) are selected here since RAPID flows are computed at the outlet of each river reach. Export (with the two stations removed) in a new shapefile called

© 2015 Cédric H. David
All Rights Reserved

“StreamGageEvent_Reg07_comid_withdir_Q_full_2004_2004_no_duplicates.shp”. There should be 417 stations in this file.

Run HydroGET again:



152,622 data points in the TimeSeries downloaded with HydroGET.

Further information

RAPID website: http://rapid-hub.org/
RAPID source code: https://github.com/c-h-david/rapid/

© 2015 Cédric H. David
All Rights Reserved