

National Park Service



Standard Operating Procedure 3.3.3

Acquiring and Preparing Mean Daily Flow Data for use in Calculation of Hydrologic Metrics as a part of the IMD Environmental Settings Protocol

Please cite this as:

DeVivo, J. C. 2018. Acquiring and Preparing Mean Daily Flow Data for use in Calculation of Hydrologic Metrics as a part of the IMD Environmental Settings Protocol. Inventory & Monitoring Division Standard Operating Procedure NPS/IMD/SOP—3.3.3. National Park Service, Fort Collins, Colorado.

Summary

Each standard operating procedure (SOP) should have a summary that describes its content.

Revision Log

[illegible]

Contents

	Page
Summary	1
Revision Log.....	1
Overview.....	3
Preparation	3
Procedures.....	5
Conducting Annual Updates of PORs to be Used for Analysis	6
Identify New Stations of Interest.....	6
Acquire and Update Metadata and Mean Daily Flow data for Stations of Interest.....	7
Update Station Qualification Codes in the Master Database	8
Update Daily Flow and POR data in the Master Database.....	8
Creating Custom Periods of Record for Analysis	8
Literature Cited	9
Appendix A. RMD File Output POR Analysis.....	10
Appendix B. RMD File Output: Acquiring Daily Flow Data to be used for Hydrologic Analysis.....	16

Overview

Calculation of metrics generally follows three steps. This SOP covers the calculation of metrics and QA/QC of resultant data. Note that only some procedures QA/QC procedures have been developed at the time of protocol publication.

1. Reviewing and updating stations of interest and periods of record of interest
2. Calculating metrics for each period of record (See SOPs 3.3.1 and 3.3.2; DeVivo 2018a, b)
 - a. Storing metric data in protocol database where calculations succeed
 - b. Storing null data and qualification flag for instances where calculations fail
3. Quality Assurance and Control (See SOPs 3.3.1 and 3.3.2)
 - a. For periods of record where calculations succeed, evaluation of whether analysis code reproduces published results (from USGS) and flagging of data incorrectly reproduced. Data will be flagged by metric across all stations and periods of record.
 - b. For periods of record where calculations fail, evaluation of the period of record and one or more of the following:
 - i. Truncating of the period of record to remove data gaps at the beginning or end. The period of record is qualified as differing from the published period of record with an explanation.
 - ii. Splitting of the period of record to periods before and after discontinuities.
 - iii. Estimation or interpolation of missing data. Metrics calculated using interpolated or estimated data are qualified as based on incomplete data records.
 - c. For each period of record, metrics are also qualified based on underlying data adequacy, which will be used to inform trend analyses. This will be based on a combination of the duration of the period of record, the type of calculations used in deriving metrics (seasonal averages vs. rolling averages), and the amount of estimated vs. measured data included in the analysis.

Preparation

Required software:

- R and R Studio. The procedures below (and scripts for calculating methods) are based on R version 3.5.1 and may need to be updated based on your local installation.

Required R Packages. Please install with all dependencies:

- dataRetrieval (<https://github.com/USGS-R/dataRetrieval>). Collection of functions to help retrieve U.S. Geological Survey (USGS) and U.S. Environmental Protection Agency (EPA) water quality and hydrology data from web services. USGS web services are discovered from National Water Information System (NWIS) <<https://waterservices.usgs.gov/>> and <<https://waterdata.usgs.gov/nwis>>. Both EPA and USGS water quality data are obtained from the Water Quality Portal <<https://www.waterqualitydata.us/>>.
- RODB (C)C (<https://cran.r-project.org/web/packages/RODBC/index.html>). An ODBC database interface.

Before you begin

- You will need to set up an ODBC connection to allow the RODB (C)C R package to read and write data to the JCD_Hydro database on the INPNISCVDBNRSSD\IMDGIS SQL Server instance.

Procedures

For each station of interest, the period of record for flow data is analyzed to identify data suitable for analysis. In general, periods of record must span at least two water years and not contain any missing data. For any given station, particularly those with long periods of record, there may be many “sub” periods of record where there are continuous data spanning two water years. For all continuous records, a series of PORs for analysis are identified beginning with the first two years, and then annually-augmented periods as appropriate for the period of record.

For example, for site number 0105400 (Androscoggin River near Gorham, NH, the period of record is from 10/1/1913 to 11/13/2018. The period of record contains two discontinuities (Figure 1):

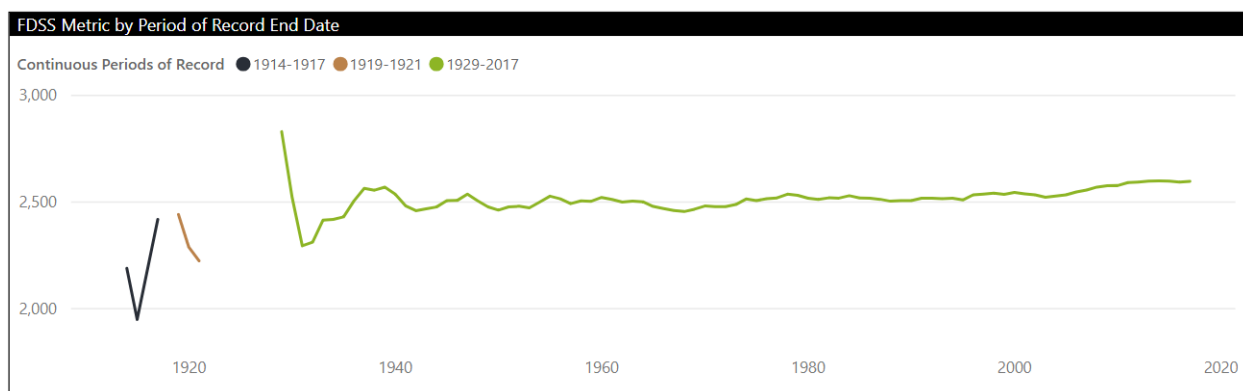


Figure 1. Mean daily streamflow for Androscoggin River near Gorham, NH.

The period of record contains the following continuous water-years of data, termed “Base PORs”

Start Date	End Date
1913-10-01	1917-09-30
1918-10-01	1921-09-30
1928-10-01	2017-09-30

Each Base POR is then used to generate several Analysis PORs, which start with the first day of the first water year, and end in incrementally-increasing end-of-water-year dates (only those for the 1913-1917 Base POR is shown):

Start Date	End Date
1913-10-01	1914-09-30
1913-10-01	1915-09-30
1913-10-01	1916-09-30
1913-10-01	1917-09-30

Scripts (in R) for analyzing periods of record are available at https://github.com/joe-devivo/ESHydrology/blob/master/vignettes/POR_Analysis.Rmd (Table 1 and **Error! Reference source not found.**). Prior to protocol publication, periods of record have been analyzed for all stations within 10km of parks; these stations and periods of record will need to be updated on an annual basis.

Table 1. Scripts to analyze periods of record for hydrologic metric calculations.

Script Name	Description
POR_Analysis.Rmd	Scripts to conduct initial population of data tables with periods of record for analysis and to perform annual updates
AcquireDailyFlowData.Rmd	Scripts to acquire daily mean flow data from USGS, interpolate missing data, and evaluate the data record for suitability for use in the IMD Environmental Settings Protocol

Generally, each year the following must occur:

1. Update Stations table with any new stations of interest.
2. Update Data Available table from USGS services. Update flags for stations that have or do not have discharge data available.
3. Evaluate station PORs for data gaps and derive Base PORs for each station.
4. Compare new Base PORs to those in the master database. Comparisons should be based on station_no and AnalysisStartDate (or start year). if new End Year is > old End Year, update the Base POR (don't replace; it needs the same ID)
5. For all new Base PORs and all Base PORs that have changed, construct intermediate intervals.
6. Add any Base PORs not already in the AnalysisPORs table.

Conducting Annual Updates of PORs to be Used for Analysis

IMD maintains a database of stream gage stations relevant to the IMD Environmental Setting protocol. Each year, the protocol lead must query networks to ensure the database is accurate, review station metadata and data availability, and if appropriate, add new periods of record to the analysis queue.

Identify New Stations of Interest

1. In January of each year, a data call to networks should go out requesting any additions to the list of stations for which hydrologic metrics are calculated.
2. Requests for new stations can be entered using a google form [here](#).
3. Prior to calculating metrics, download the responses and unzip the file to get the *.csv file:

The screenshot shows a Google Form interface. At the top, there are two tabs: 'QUESTIONS' and 'RESPONSES', with 'RESPONSES' being the active tab and having a '1' next to it. Below the tabs, it says '1 response'. There are two buttons: 'SUMMARY' (highlighted in purple) and 'INDIVIDUAL'. Below these buttons is a section titled 'Who has responded?' with a table showing one response. The table has a header 'Email' and a row with the email 'joe_devivo@nps.gov'. Below this is a section titled 'Station ID' with a table showing one response. The table has a header 'Station ID' and a row with the value '01199900'. A dropdown menu is open on the right side of the 'RESPONSES' tab, showing options: 'Get email notifications for new responses', 'Select response destination', 'Unlink form', 'Download responses (.csv)' (highlighted), 'Print all responses', and 'Delete all responses'.

4. See section 3.1 of the POR_Analysis.Rmd file for instructions on how to enter new station data into the database.
5. Once added, you can delete all responses in the google form.

Acquire and Update Metadata and Mean Daily Flow data for Stations of Interest

1. See sections 2 and 3 of the AcquireDailyFlowData.Rmd file for procedures to acquire and update station metadata.
2. See section 4 of the AcquireDailyFlowData.Rmd file for procedures to download daily flow data. This process will save a local .Rdata file so that if you are interrupted during processing you will not need to requery USGS for the data.
3. See section 5 of the Acquire DailyFlowData.Rmd file for QC procedures to prepare the flow data for analysis. This includes
 - a. Eliminating stations where all flows measured zero
 - b. Interpolating missing data for small data gaps (three days or less when such gaps are less than 40% of the data record in total), and
 - c. Conducting an evaluation of the length and gaps between periods of record at each station to identify those suitable for analysis. Stations (and daily flow data) for non-analyzable periods of record are removed from the dataset.

Update Station Qualification Codes in the Master Database

1. See section 6 of the `AcquireDailyFlowData.Rmd` file for procedures to update station qualifiers in the master database. This includes:
 - a. Flagging any stations that should be newly-classified as having daily flow data
 - b. Flagging any stations that should be newly-classified as having data suitable for the ES Protocol.
 - c. Unflagging any stations that should no longer be classified as having data suitable for the ES Protocol.

Update Daily Flow and POR data in the Master Database

See section 7 of the `AcquireDailyFlowData.Rmd` file for procedures to update mean daily flow data and POR data. This includes:

1. Writing daily flow data to the database
2. Writing daily flow metadata to the database to support documentation needs in data visualization and exploration applications
3. Writing new Base PORs and PORs for Analysis to the Database
4. Saving a local file of the fully-QCed data frame for use in data analysis and reporting.

Creating Custom Periods of Record for Analysis

Other periods of interest may be added depending on park interest/need (such as in cases where there is interest in comparing metric values before and after events). Standard procedures to be developed for Version 2.0.

Literature Cited

DeVivo, J. C. 2018a. Calculation of Hydrologic Indicator Tools metrics and Fundamental Daily Streamflow Statistics. Inventory & Monitoring Division Standard Operating Procedure NPS/IMD/SOP—3.3.1. National Park Service, Fort Collins, Colorado.

DeVivo, J. C. 2018b. Calculation of Indicators of Hydrologic Alteration metrics. Inventory & Monitoring Division Standard Operating Procedure NPS/IMD/SOP—3.3.2. National Park Service, Fort Collins, Colorado.

Appendix A. RMD File Output POR Analysis

This vignette is current as of 15 November 2018. For the latest version, see:

https://github.com/joe-devivo/ESHydrology/blob/master/vignettes/POR_Analysis.Rmd

1. Before you begin

You will need to set up an ODBC connection named “Hydro” to allow the RODBC R package to read and write data to the JCD_Hydro database on the INPNISCVDBNRSSDSQL Server instance.

You will also need the following R packages installed:

- dataRetrieval (<https://github.com/USGS-R/dataRetrieval>). Collection of functions to help retrieve U.S. Geological Survey (USGS) and U.S. Environmental Protection Agency (EPA) water quality and hydrology data from web services. USGS web services are discovered from National Water Information System (NWIS) <https://waterservices.usgs.gov/> and <https://waterdata.usgs.gov/nwis>. Both EPA and USGS water quality data are obtained from the Water Quality Portal <https://www.waterqualitydata.us/>.
- RODBC (<https://cran.r-project.org/web/packages/RODBC/index.html>). An ODBC database interface.

2. Constructing Initial Periods of Record to Use for Analysis

2.1. Pull station and POR data from the master database

The following pulls data from the master hydrology database

```
rm(sites_to_process)

ch <- odbcConnect("Hydro")
siteInfo<-sqlFetch(ch,"dbo.vw_Stations_With_Daily_Discharge_Data",as.is = c(TRUE), stringsAsFactors = FALSE)
sites_to_process<- sqlFetch(ch,"dbo.vw_Stations_To_Analyze_for_POR_continuity",as.is = c(TRUE), stringsAsFactors = FALSE)
MasterPORs<- sqlFetch(ch,"dbo.tlu_PORs")
odbcClose(ch)
colnames(MasterPORs)[5] <- "AnalysisStartDate"
colnames(MasterPORs)[6] <- "AnalysisEndDate"
```

2.2 Create Base Periods of Record within PORs

The following breaks any station’s POR into sub-PORs that do not have any missing daily values, and retains any sub-PORs with two or more years worth of data. Base PORs are trimmed to coincide with water year start and end dates.

```
p=0
q=nrow(sites_to_process)

repeat
{
```

```

site_no<-sites_to_process[p+1,]

# get daily flow data for a site
dailyQ <- readNWISdv(siteNumber = site_no, parameterCd = "00060")

# get first and last date of any data gaps, generate data gap table. If it contains any records, the full POR is not usable and must be split into sub-PORs for analysis
startgaps<-dailyQ[diff(dailyQ$Date)>1, ]
endgaps<-dailyQ[c(1, diff(dailyQ$Date))>1, ]
dfgaps <-data.frame( start= startgaps$Date, end = endgaps$Date)

# get first and last date of full POR
startdate<-min(dailyQ$Date)
enddate<-max(dailyQ$Date)

if (nrow(dfgaps)==0) {

PORs<-rbind(data.frame(start=startdate, end=enddate))

} else if (nrow(dfgaps)==1) {

# construct first sub-POR
PORs<-rbind(data.frame(start=startdate, end=min(dfgaps$start)))

# construct last sub POR
startPORs<-dailyQ[c(1, diff(dailyQ$Date))>1, ]
PORs<-rbind(PORs,data.frame(start=max(startPORs$Date), end=enddate))

} else {

# construct first sub-POR
PORs<-rbind(data.frame(start=startdate, end=min(dfgaps$start)))

# get intermediate sub-PORs if there is more than one data gap

x<-1
y<-nrow(dfgaps)

repeat
{
PORs<-rbind(PORs,data.frame(start=dfgaps$end[x], end=dfgaps$start[x+1]))
x=x+1
if(x==y)
{
break
}
}
}

```

```

# construct last sub POR
startPORs<-dailyQ[c(1, diff(dailyQ$Date))>1, ]
PORs<-rbind(PORs,data.frame(start=max(startPORs$Date), end=enddate))

}
rm(POR_out)
POR_out<-cbind(site_no,POR_begin_date=as.character(PORs$start),POR_end_date=
as.character(PORs$end),Year_Type="water",AnalysisStartYear="",AnalysisEndYear=
"", AnalysisStartDate="", AnalysisEndDate="", ShortTest="")

# if start date is a leap year then assign start year based on day of year
POR_out[,7]<-ifelse(

  leap_year(as.Date(POR_out[,2]))== TRUE,

  ifelse(
    yday(POR_out[,2]) > 275,
    paste(year(POR_out[,2])+1,"-10-01", sep=""),
    paste(year(POR_out[,2]),"-10-01", sep="")
  ),

  ifelse(
    yday(POR_out[,2]) > 274,
    paste(year(POR_out[,2])+1,"-10-01", sep=""),
    paste(year(POR_out[,2]),"-10-01", sep="")
  )
)

#repeat for end dates.
POR_out[,8]<-ifelse(

  leap_year(as.Date(POR_out[,3]))== TRUE,

  ifelse(
    yday(POR_out[,3]) < 275,
    paste(year(POR_out[,3])-1,"-09-30", sep=""),
    paste(year(POR_out[,3]),"-09-30", sep="")
  ),

  ifelse(
    yday(POR_out[,3]) < 274,
    paste(year(POR_out[,3])-1,"-09-30", sep=""),
    paste(year(POR_out[,3]),"-09-30", sep="")
  )
)

# Get Analysis Water Years for based on Analysis Start and End Dates
POR_out[,5]<-as.character(waterYear(POR_out[,7]))

```

```

POR_out[,6]<-as.character(waterYear(POR_out[,8]))

# Remove any PORs with less than two years of data (maybe need to change this
to three years)POR

POR_out[,9]<-ifelse(
  as.integer(POR_out[,6])-as.integer(POR_out[,5])+1 < 2,
  "Fail",
  "Pass"
)

POR_out<-subset(POR_out,POR_out[,9]=="Pass")

if (nrow(POR_out)>0) {

# Write PORs to master POR table

POR_out<-merge(POR_out,siteInfo,by="site_no")
POR_out[,8]<-as.Date(POR_out[,8])
POR_out[,7]<-as.Date(POR_out[,7])
POR_out<-merge(POR_out,MasterPORs,by=c("AnalysisStartDate","AnalysisEndDate")
)
PORs_to_write<-subset(POR_out,select=c(10,54))

} else {
print("station POR is too short to do anything with")
temp<-cbind(site_no,POR_ID=22953)
temp<-merge(temp,siteInfo,by="site_no")
PORs_to_write<-subset(temp,select=c(3,2))

}

ch <- odbcConnect("Hydro")
sqlSave(ch,PORs_to_write,"dbo.vw_site_no_X_PORs",append=TRUE,rownames=FALSE)
odbcClose(ch)
p=p+1
if(p==q)
{
break
}
}
}

```

2.3 Create Analysis PORs

The following creates PORs to be used for analysis from each base POR.

```

ch <- odbcConnect("Hydro")
siteInfo<-sqlFetch(ch,"dbo.vw_Stations_With_Daily_Discharge_Data",as.is = c(TRUE), stringsAsFactors = FALSE)
MasterPORs<- sqlFetch(ch,"dbo.tlu_PORs")
sites_X_PORs<- sqlFetch(ch,"dbo.vw_PORs_To_Use_For_Hydro_Metrics")
odbcClose(ch)

AnalysisPorsOut<-cbind(sites_X_PORs[1],sites_X_PORs[2],sites_X_PORs[3])

ch <- odbcConnect("Hydro")
sqlSave(ch,AnalysisPorsOut,"dbo.vw_site_no_X_PORs_for_Analysis",append=TRUE,rownames=FALSE)
odbcClose(ch)

```

3. Annual Updates to PORs

Each year, (after the turn of the water year), the PORs to be analyzed are updated as follows:

1. Update Stations table with any new stations of interest.
2. Update Data Available table from USGS services. Update flags for stations that have or do not have discharge data available.
3. Evaluate station PORs for data gaps and derive Base PORs for each station.
4. Compare new Base PORs to those in the master database. Comparisons should be based on station_no and AnalysisStartDate (or start year). if new End Year is > old End Year, update the Base POR (don't replace; it needs the same ID)
5. For all new Base PORs and all Base PORs that have changed, construct intermediate intervals.
6. Add any Base PORs not already in the AnalysisPORs table.

The following sections are to be completed in version 2.0...

3.1 Update Stations table with any new stations of interest

Temporary Instructions: Code to be developed upon migration of database to final staging location.

3.2. Update Station Metadata and Available Daily Flow Data

Temporary Instructions: Code to be developed upon migration of database to final staging location.

This will query any requested stations that are not already in the database, retrieve the station metadata from the USGS/EPA Water Quality Portal, and append the metadata for those stations back to the database. Do not proceed to Step 2 until all stations have been successfully added.

Note that if any station IDs are invalid the script will not work, and therefore adding new stations may require some troubleshooting if errors are encountered. Contact the appropriate network program manager if information provided does not allow for resolution of proper ID names.

Link Stations of Interest to Parks

1. Option A. Run Part 2 of the UpdateStationsAndPORs script. This will use data provided by network & park staff to append park X site number combinations to the master data table. This is the suggested method in cases where multiple sites are being added at once.
2. Option B. Manually enter combinations using the HydroData Database. To do this, open the Link Station to Park form and enter the appropriate park/station combination.

3.3. Update Base POR table with New Base PORs

Temporary Instructions: Code to be developed upon migration of database to final staging location based on code in section 2.2 above.

3.4. Update Analysis POR table with New Analysis PORs

Temporary Instructions: Code to be developed upon migration of database to final staging location based on code in section 2.3 above.

4. Quality Assurance and Quality Control Procedures

Planned quality control procedures include the following (to be developed in Version 2.0):

1. QC check and qualification flagging for any PORs that contain missing data in the prior water year. This identifies PORs to be split into sub-PORs and for possible estimation of missing values for any gaps of 1 day or less.
2. QC check and qualification flagging for any PORs that are too short to run metrics on.
3. QC check and qualification flagging for any PORs based on length of record.
4. QC check and qualification flagging for any PORs that contain estimated values from USGS
5. QC check and qualification flagging for any PORs that contain estimated values from NPS

Appendix B. RMD File Output: Acquiring Daily Flow Data to be used for Hydrologic Analysis

This vignette is current as of 15 November 2018. For the latest version, see:

<https://github.com/joe-devivo/ESHydrology/blob/master/vignettes/AcquireDailyFlowData.Rmd>

1. Before you begin

This vignette describes how to pull daily flow data for all stations relevant to the national IMD Environmental Settings protocol. Its primary use is to download and archive the daily flow data used in metric calculation to allow for reproducibility of results, but sections 2-5 can be used to update data used to drive data visualization and exploration applications.

1.1 Before you begin

You will need to set up an ODBC connection named “Hydro” to allow the RODBC R package to read and write data to the JCD_Hydro database on the INPNISCVDBNRSSDSQL Server instance.

You will also need the following R packages installed:

- dataRetrieval (<https://github.com/USGS-R/dataRetrieval>). Collection of functions to help retrieve U.S. Geological Survey (USGS) and U.S. Environmental Protection Agency (EPA) water quality and hydrology data from web services. USGS web services are discovered from National Water Information System (NWIS) <https://waterservices.usgs.gov/> and <https://waterdata.usgs.gov/nwis>. Both EPA and USGS water quality data are obtained from the Water Quality Portal <https://www.waterqualitydata.us/>.
- RODBC (<https://cran.r-project.org/web/packages/RODBC/index.html>). An ODBC database interface.
- smwrBase (<https://github.com/USGS-R/smwrBase>). Collection of functions used to prepare time-series hydrologic data for analysis. Used to interpolate missing data.
- waterData (<https://cran.r-project.org/web/packages/waterData/index.html>). Collection of functions to import daily hydrologic data from USGS web services.

2. Pull station data from the master database

The following pulls the current site information data from the master hydrology database. The Station IDs are used in later steps to query USGS for daily flow data.

```
# Pull list of sites associated with the national protocol

ch <- odbcConnect("Hydro")
mstr_sitesWithDailyDischargeData<-sqlFetch(ch,"dbo.vw_Stations_With_Daily_Discharge_Data",as.is = c(TRUE), stringsAsFactors = FALSE)
mstr_siteInfo<- sqlFetch(ch,"dbo.tbl_siteInfo",as.is = c(TRUE), stringsAsFactors = FALSE)
mstr_stationQuals<-sqlFetch(ch,"dbo.xtab_site_no_X_station_qualifiers",as.is
```

```
= c(TRUE), stringsAsFactors = FALSE)
odbcClose(ch)
```

3. Process Station Information

The following examines the metadata for all stations in the database to determine which have daily flow data available, and updates the XX table with appropriate qualifiers.

```
# Get daily flow period of record information for all statins
sitesWithDailyDischargeData <- whatNWISdata(siteNumber = mstr_siteInfo$site_no,
parameterCd="00060", statCd ="00003")
sitesWithDailyDischargeData<-unique(subset(sitesWithDailyDischargeData, data_type_cd == 'dv',
select=-c(loc_web_ds)))
```

```
# Update master database tables with any changes
```

4. Get Daily Flow data

```
AllDailyFlowData<- setNames(data.frame(matrix(ncol = 5, nrow = 0)), c("agency_cd",
"site_no", "Date", "X_00060_00003", "X_00060_00003_cd"))
```

```
stationcount<-nrow(sitesWithDailyDischargeData)
for (i in 1:stationcount) {
  dailyQ <- readNWISdv(siteNumber = sitesWithDailyDischargeData[i,2], parameterCd = "00060",
startDate = sitesWithDailyDischargeData[i,21], endDate = sitesWithDailyDischargeData[i,22])
  dailyQ <-subset(dailyQ,select=1:5)
  colnames(dailyQ)<-c("agency_cd", "site_no", "Date", "X_00060_00003", "X_00060_00003_cd")
  AllDailyFlowData<-rbind(AllDailyFlowData,dailyQ)
}
```

```
#write data frame to file so that you don't have to do this again!
save(AllDailyFlowData,file="AllDailyFlow.RData",compress=TRUE)
```

5. Quality Assurance and Quality Control Procedures

QA/QC of mean daily flow is performed primarily by data providers. The following procedures evaluate, augment, and qualify source data prior to use in analysis as a part of the IMD Environmental Settings Protocol.

5.1. Missing Data Handling

Periods of record are examined for missing data, and any data gaps of two days or less are interpolated using the “fillMiss” function within the waterData package (Ryberg and Veccia 2017).

From the package documentation:

The method used to interpolate missing values is based on `tsSmooth` constructed using `StructTS` on `x` with type set to `trend`. The smoothing method basically uses the information (slope) from two values previous to missing values and the two values following missing values to smoothly interpolate values accounting for any change in slope. Beauchamp (1989) used time-series methods for synthesizing missing streamflow records. The group that is used to define the statistics that control the interpolation is very simply defined by span rather than the more in-depth measures described in Elshorbagy and others (2000).

If the data have gaps rather than missing values, then `fillMissing` will return a vector longer than `x` if `Dates` is given and the return data cannot be inserted into the original data frame. If `Dates` is not given, then the gap will be recognized and not be filled. The function `insertMissing` can be used to create a data frame with the complete sequence of dates.

The “span” parameter is set to 3 resulting in simple linear interpolation to replace missing values.

```
# Convert the AllDailyFlowData to a format compatible with the waterData package
# load("AllDailyFlow.RData")
AllDailyFlowData2<-AllDailyFlowData[c(2,4,3,5)]
colnames(AllDailyFlowData2)<-c("staid", "val", "dates", "qualcode")

#delete any rows where the maximum flow is 0 for the station (this causes an
error and the sites aren't useful for analysis anyway).
my.max <- function(x) ifelse( !all(is.na(x)), max(x, na.rm=T), NA)
zerotest<-AllDailyFlowData2 %>% group_by(staid) %>% summarise(Value = max(my.
max(val)))
zerotest<-subset(zerotest, Value==0)
AllDailyFlowData2<-anti_join(AllDailyFlowData2,zerotest)
maxvaluesbystation<-AllDailyFlowData2 %>% group_by(staid) %>% summarise(Value
= max(my.max(val)))

# Initialize dataframe to store flow data with interpolated values
AllInterpolatedFlowData<- setNames(data.frame(matrix(ncol = 4, nrow = 0)), c(
"staid", "val", "dates", "qualcode"))

# Begin interpolations
stationcount<-nrow(maxvaluesbystation)

for (i in 1:stationcount) {
  # subset AllDailyFlowData by station and order by date
  temp<-subset(AllDailyFlowData2,staid==as.character(maxvaluesbystation[i,1])
)

  #interpolate missing data for gaps of two days or less
  temp2<-fillMiss(temp, block=3, pmiss=40)

  # Append missing data to daily flow data frame
  AllInterpolatedFlowData<-rbind(AllInterpolatedFlowData,temp2)
}
```

```
#write data frame to file so that you don't have to do this again!
save(AllInterpolatedFlowData,file="InterpolatedDailyFlow.RData",compress=TRUE
)
```

5.2. Remove Stations with PORs Less than Two Years

Metrics can only be calculated on periods of record of two water years or more. The following trims the data to the start and end of the first and last water year in the record and removes any stations (and associated data) from the data set. The trimmed periods of record are termed “Base PORs” for the protocol.

```
# load("InterpolatedDailyFlow.RData")

my.max <- function(x) ifelse( !all(is.na(x)), max(x, na.rm=T), NA)
maxvaluesbystation<-AllInterpolatedFlowData %>% group_by(staid) %>% summarise
(Value = max(my.max(val)))
stationcount<-nrow(maxvaluesbystation)

# Initialize POR table
rm(tempBasePORs)
tempBasePORs<- setNames(data.frame(matrix(ncol = 9, nrow = 0)), c("staid", "P
OR_begin_date", "POR_end_date", "Year_Type", "AnalysisStartYear", "AnalysisEn
dYear", "AnalysisStartDate", "AnalysisEndDate", "ShortTest"))

# Trim PORs to start and end of water years
for (i in 1:stationcount) {

  # subset InterpolatedDailyFlowData by station and order by date
  temp<-subset(AllInterpolatedFlowData,staid==as.character(maxvaluesbystation
[i,1]))

  # get first and last date of full POR
  startdate<-min(temp$dates)
  enddate<-max(temp$dates)

  rm(POR_out)
  POR_out<-cbind(maxvaluesbystation[i,1],POR_begin_date=as.character(startdat
e),POR_end_date=as.character(enddate),Year_Type="water",AnalysisStartYear="",
AnalysisEndYear="", AnalysisStartDate="", AnalysisEndDate="", ShortTest=""
)

  # if start date is a leap year then assign start year based on day of year
  POR_out[,7]<-ifelse(
    leap_year(as.Date(POR_out[,2]))== TRUE,
    ifelse(
```

```

      yday(POR_out[,2]) > 275,
      paste(year(POR_out[,2])+1,"-10-01", sep=""),
      paste(year(POR_out[,2]),"-10-01", sep="")
    ),

    ifelse(
      yday(POR_out[,2]) > 274,
      paste(year(POR_out[,2])+1,"-10-01", sep=""),
      paste(year(POR_out[,2]),"-10-01", sep="")
    )
  )
)

#repeat for end dates.

POR_out[,8]<-ifelse(
  leap_year(as.Date(POR_out[,3]))== TRUE,

  ifelse(
    yday(POR_out[,3]) < 275,
    paste(year(POR_out[,3])-1,"-09-30", sep=""),
    paste(year(POR_out[,3]),"-09-30", sep="")
  ),

  ifelse(
    yday(POR_out[,3]) < 274,
    paste(year(POR_out[,3])-1,"-09-30", sep=""),
    paste(year(POR_out[,3]),"-09-30", sep="")
  )
)

# Get Analysis Water Years for based on Analysis Start and End Dates
POR_out[,5]<-as.character(waterYear(POR_out[,7]))
POR_out[,6]<-as.character(waterYear(POR_out[,8]))

# Flag Base PORs with less than two years of data

POR_out[,9]<-ifelse(
  as.integer(POR_out[,6])-as.integer(POR_out[,5])+1 < 2,
  "Fail",
  "Pass"
)

# Append POR to tempPORs table
tempBasePORs<-rbind(tempBasePORs,POR_out)
}

# Delete Base PORs with less than two years of data

```

```

BasePORs<-subset(tempBasePORs,tempBasePORs[,9]=="Pass")
BasePORsTooShort<-subset(tempBasePORs,tempBasePORs[,9]=="Fail")
StatioinsWithoutBasePORs<-as.data.frame(unique(BasePORsTooShort$staid))
colnames(StatioinsWithoutBasePORs)<-"staid"

# Remove Flow Data for sites with BasePORs too short
InterpolatedDailyFlowData2<-anti_join(AllInterpolatedFlowData,BasePORsTooShort)

```

5.3. Data Gap Analysis

Metrics can only be calculated on continuous periods of record. The following inspects the data for each station, identifies data gaps, and breaks the period of record into sub-PORs of two water years or more for use in metric calculation (these are termed AnalysisPORs). In cases where stations do not have any continuous AnalysisPORs of two years or more, the station and associated data are removed from the dataset.

The resultant data set should be published with an accompanying data report.

```

# load("InterpolatedDailyFlow.RData")

my.max <- function(x) ifelse( !all(is.na(x)), max(x, na.rm=T), NA)
maxvaluesbystation<-InterpolatedDailyFlowData2 %>% group_by(staid) %>% summarise(Value = max(my.max(val)))
stationcount<-nrow(maxvaluesbystation)

# Initialize POR table
rm(AnalysisPORs)
AnalysisPORs<- setNames(data.frame(matrix(ncol = 9, nrow = 0)), c("staid", "POR_begin_date", "POR_end_date", "Year_Type", "AnalysisStartYear", "AnalysisEndYear", "AnalysisStartDate", "AnalysisEndDate", "ShortTest"))

# Perform gap analysis
for (i in 1:stationcount) {

  # subset InterpolatedDailyFlowData by station and order by date
  temp<-subset(InterpolatedDailyFlowData2,staid==as.character(maxvaluesbystation[i,1]))
  temp<-temp[order(temp$dates),]

  # get first and last date of any data gaps, generate data gap table. If it contains any records, the full POR is not usable and must be split into sub-PORs for analysis
  startgaps<-temp[diff(temp$dates)>1, ]
  endgaps<-temp[c(1, diff(temp$dates))>1, ]

  if (nrow(startgaps)==nrow(endgaps)) {

    dfgaps <-data.frame( start= startgaps$dates, end = endgaps$dates)
  }
}

```

```

} else if (nrow(startgaps) > nrow(endgaps)) {
  dfgaps <-data.frame( start= head(startgaps$dates,nrow(endgaps)), end = en
d gaps$dates)

}

# get first and last date of full POR
startdate<-min(temp$dates)
enddate<-max(temp$dates)

# construct sub-PORs
rm(PORs)
if (nrow(dfgaps)==0) {

  PORs<-rbind(data.frame(start=startdate, end=enddate))

} else if (nrow(dfgaps)==1) {

  # construct first sub-POR
  PORs<-rbind(data.frame(start=startdate, end=min(dfgaps$start)))

  # construct last sub POR
  startPORs<-temp[c(1, diff(temp$dates))>1, ]
  PORs<-rbind(PORs,data.frame(start=max(startPORs$dates), end=enddate))

} else {

  # construct first sub-POR
  PORs<-rbind(data.frame(start=startdate, end=min(dfgaps$start)))

  # get intermediate sub-PORs if there is more than one data gap

  x<-1
  y<-nrow(dfgaps)

  repeat
  {
    PORs<-rbind(PORs,data.frame(start=dfgaps$end[x], end=dfgaps$start[x+1])
)
    x=x+1
    if(x==y)
    {
      break
    }
  }

  # construct last sub POR
  startPORs<-temp[c(1, diff(temp$dates))>1, ]
  PORs<-rbind(PORs,data.frame(start=max(startPORs$dates), end=enddate))

```

```

}

rm(POR_out)
POR_out<-cbind(maxvaluesbystation[i,1],POR_begin_date=as.character(PORs$start),POR_end_date=as.character(PORs$end),Year_Type="water",AnalysisStartYear="",AnalysisEndYear="", AnalysisStartDate="", AnalysisEndDate="", ShortTest="")

# if start date is a leap year then assign start year based on day of year
POR_out[,7]<-ifelse(

  leap_year(as.Date(POR_out[,2]))== TRUE,

  ifelse(
    yday(POR_out[,2]) > 275,
    paste(year(POR_out[,2])+1,"-10-01", sep=""),
    paste(year(POR_out[,2]),"-10-01", sep="")
  ),

  ifelse(
    yday(POR_out[,2]) > 274,
    paste(year(POR_out[,2])+1,"-10-01", sep=""),
    paste(year(POR_out[,2]),"-10-01", sep="")
  )
)

#repeat for end dates.
POR_out[,8]<-ifelse(

  leap_year(as.Date(POR_out[,3]))== TRUE,

  ifelse(
    yday(POR_out[,3]) < 275,
    paste(year(POR_out[,3])-1,"-09-30", sep=""),
    paste(year(POR_out[,3]),"-09-30", sep="")
  ),

  ifelse(
    yday(POR_out[,3]) < 274,
    paste(year(POR_out[,3])-1,"-09-30", sep=""),
    paste(year(POR_out[,3]),"-09-30", sep="")
  )
)

# Get Analysis Water Years for based on Analysis Start and End Dates
POR_out[,5]<-as.character(waterYear(POR_out[,7]))
POR_out[,6]<-as.character(waterYear(POR_out[,8]))

```

```

POR_out[,9]<-ifelse(
  as.integer(POR_out[,6])-as.integer(POR_out[,5])+1 < 2,
  "Fail",
  "Pass"
)

# Append PORs to tempPORs table
AnalysisPORs<-rbind(AnalysisPORs,POR_out)

}

# Remove any AnalysisPORs with less than two years of data
AnalysisPORs<-subset(AnalysisPORs,AnalysisPORs[,9]=="Pass")

# Identify Stations with no viable AnalysisPORs
StatioinsWithBasePORs<-as.data.frame(unique(BasePORs$staid))
colnames(StatioinsWithBasePORs)<-"staid"
StationsWithAnalysisPORs<-as.data.frame(unique(AnalysisPORs$staid))
colnames(StationsWithAnalysisPORs)<-"staid"
StationsWithoutAnalysisPORs<-anti_join(StatioinsWithBasePORs,StationsWithAnal
ysisPORs)

# Remove Flow data for stations without viable AnalysisPORs
InterpolatedDailyFlowData3<-anti_join(InterpolatedDailyFlowData2,StationsWith
outAnalysisPORs)

#write data frame to file so that you don't have to do this again!
save(InterpolatedDailyFlowData3,file="FlowDataForMetricCalculations.RData",co
mpress=TRUE)

```

6. Update Period of Record Information

Each year, the master database must be updated with new periods of record for newly-identified stations of interest, and for active stations as new data are collected.

6.1 New Base PORs

To be developed when preparing 2018 water year data.

6.2 New Analysis PORs

To be developed when preparing 2018 water year data.

7. Update Station Qualification Codes in Master Database

Because the stations used in the IMD Environmental Settings Protocol are a subset of those contained in the broader monitoring station inventory (also maintained by IMD), monitoring stations are flagged based on their suitability for use in the Environmental Settings Protocol.

7.1 Flag any stations that should be newly classified as having daily flow data

To be developed when preparing 2018 water year data.

7.2 Flag any stations that should be newly classified as having data suitable for the ES Protocol

To be developed when preparing 2018 water year data.

7.3 Unflag any stations that should not be classified as having data suitable for the ES Protocol

To be developed when preparing 2018 water year data.

8. Export daily flow and POR Data to master database

Data from the 2017 water year are already in the master database. Procedures will be updated to conform to final staging database when preparing 2018 water year data for analysis.

Three components: * Write daily flow data to database * Update daily flow metadata table to support visualization documentation * Update BasePORs Table * Update AnalysisPORs Table

9. References

Ryberg, K.R. and Vecchia, A.V., 2012, waterData—An R package for retrieval, analysis, and anomaly calculation of daily hydrologic time series data, version 1.0: U.S. Geological Survey OpenFile Report 2012–1168, 8 p. (Also available at <http://pubs.usgs.gov/of/2012/1168/>)