User's guide to the Stratospheric Water and OzOne Satellite Homogenized (SWOOSH) data set

Sean M. Davis (sean.m.davis@noaa.gov), Karen H. Rosenlof

Purpose:

The primary goal of this document is to offer a brief overview of the SWOOSH data set and explain the naming convention for variables contained therein.

Overview:

The primary SWOOSH product is a monthly-mean zonal-mean gridded data set containing ozone and water vapor data from the SAGE-II/III, UARS HALOE, UARS MLS, and Aura MLS satellite instruments, spanning 1984 to present. For both ozone and water, corrections are applied to the SAGE, HALOE, and UARS MLS data to force agreement with the Aura MLS measurements. The corrections are additive offsets that vary with latitude and height, but not with time, and are determined from coincident observations closely matched in space and time during the instrument overlap time periods.

The data set contains gridded monthly-mean water vapor and ozone concentrations from SAGE, HALOE, and MLS data, as well as the uncorrected versions from SAGE, HALOE, and UARS MLS (for reference). Data from each instrument are quality-controlled, as described in the relevant literature. In addition, unphysical SAGE H2O data have been removed using a 3-sigma filter. An empirical correction devised by us was applied to the MLS data to remove unphysical oscillations in the high latitude data at levels between 100 and 215 hPa.

A combined product containing data from all instruments is also provided. The combined product is constructed by taking a weighted mean of the available satellite measurements in the given lat/height/time bin. The combined product comes in several flavors, including one that contains missing data, and several types that have been filled in based on different algorithms.

In addition to the gridded monthly-mean zonal-mean water vapor and ozone concentration values, ancillary information such as the standard deviation, number of data points, and mean profile uncertainty is stored for each variable type.

Furthermore, for each variable type the data are provided on two types of horizontal grids: geographic and equivalent latitude. The SWOOSH data set contains numerous variables, and for ease of use, a logical naming convention has been devised. This naming convention is described below.

Data set resolution:

- Time grid:
  - All products are monthly-mean (1984 – present)
- Horizontal grid:
  - 2.5°, 5°, and 10° zonal-mean (also contains equivalent latitude-gridded data)

    - 30° (lon) x 10° (lat), 20° x 5°
- Vertical grid:
    - 31 pressure levels, 316 -1 hPa (Same as Aura MLS v4 products)
    - 21 isentropic levels (300 – 650 K)

File description:

The data are provided as both an IDL save file (.sav) and in netCDF format (.nc). The same variable names are present in each file, and missing data are stored as IEEE NaNs.

The files are named as follows:

swoosh-version number-startyyyymm-endyyyymm-gridtype-horizontalresolution-numberverticallevels.sav (or *.nc*)

EXAMPLE: swoosh-v02.0-198401-201204-latpress-10deg-L31.nc means

SWOOSH version 2.0
Starting time: 198401
Ending time: 201204
Grid: zonal-mean, vertical pressure grid
10 degree latitude resolution
31 vertical levels

Variable naming convention:

All variables are stored as arrays of [(lon),lat, level, time]. The naming convention is:

[Instrument] *(correction)(fill)(seasonalcycle/anomaly)* ***[species][quantity]*** *(grid)*

NOTE: Categories in [ ] brackets are present in every variable name, those in ( ) parentheses are optional.

TABLE 1. Possible values for each category (optional categories in orange)

Instrument	**Correction**	**Fill**	**Seasonal cycle /anomaly**	**Species**	**Quantity**	**Grid**
sage2	raw	eqfill	seas	h2o	q	eq
sage3		anomfill	anom	o3	n	
haloe		eqfillanomfill			sigma	
uarsmls					meanunc	
mls						
combined						
hirdls*						
ace*						

* HIRDLS and ACE data not included in combined product.

1. ***Instrument***
   - The instrument source (or combination of instrument sources).

2. ***Correction***
   - raw - (e.g., sage2rawh2oq) The variable contains the original, uncorrected data. If this is omitted (e.g., sage2h2oq), then the variable has been corrected to line up with the MLS data during the overlap period. Corrections are additive offsets that vary with pressure and latitude.

3. ***Fill***
   - Eqfill – When possible, missing values in the geographically-gridded variable are filled in using values from the corresponding equivalent latitude-gridded variable.
   - Anomfill – For each vertical level, the anomaly array is interpolated (in lat, time space) to fill in missing values. The anomaly array is then added back to the seasonal cycle variable to get the "anomfill" variable.
   - Eqfillanomfill – A combination of the two processes above (eqfill is performed first , then anomfill)

4. ***Seasonal cycle / anomaly***
   - Values are computed relative to data from the given instrument. E.g., combinedanomh2oq is computed from the combinedh2oq variable, and sageanomo3q is computed from the sageo3q variable.

5. ***Species***
   - This is either "h2o" or "o3"

6. ***Quantity***
   - q – mixing ratio (ppmv)
   - N – number of measurements in the bin
   - Sigma – standard deviation of data in the bin (ppmv)
   - Meanunc – mean of the profile uncertainties of data in the bin (ppmv)

7. ***Grid***
   - eq – equivalent latitude grid (blank if geographical grid)

Additional variables not described above:

Dimension variables:
lat, level, time, latbound, jultime, month, year, yrtime

The meaning of these variables is probably obvious, except perhaps for the time variables. The "time" variable is the month since 1/1984, the "jultime" variable is the date in IDL's Julian format, and "yrtime" is the fractional year (e.g., 2000.88 is 11/2000).

Offset variables:
E.g., haloeh2omeandiff, haloeh2omeandiffvslat, haloeo3meandiff, haloeo3meandiffvslat, sageh2omeandiff, sageh2omeandiffvslat, sageo3meandiff, sageo3meandiffvslat, offsetlat, uarsmlsh2omeandiff, uarsmlsh2omeandiffvslat

These variables give the offsets as a function of height (*meandiff variables), or as a function of height and latitutde (*meandiffvslat). The *meandiffvslat variables are what are actually used for the adjustments of the HALOE , UARS MLS, and SAGE data. The *meandiff variables are provided only for a quicklook as to how the offsets vary with height. For the *meandiffvslat variables, the latitude binning is 10°, and the bin centers are stored in the variable offsetlat.

Tropopause variables:
These are calculated from MERRA. See metadata in netCDF files for descriptions.

An Example (IDL):

To see how many SAGE-II H2O observations there were at 100 hPa in the grid containing 49°N during the month of 1/1985, do the following:

```
restore, '~/sd-data/projects/swoosh/savfiles/swoosh-v02.5-198401-
201512-latpress-10deg-L31.sav'
latind = VALUE_LOCATE( latbound, 50)
pind = WHERE( fix(level) EQ 100 )
tind = WHERE( (month eq 1) and (year eq 1990) )

IDL> print, sage2h2on[latind, pind, tind]

30
```

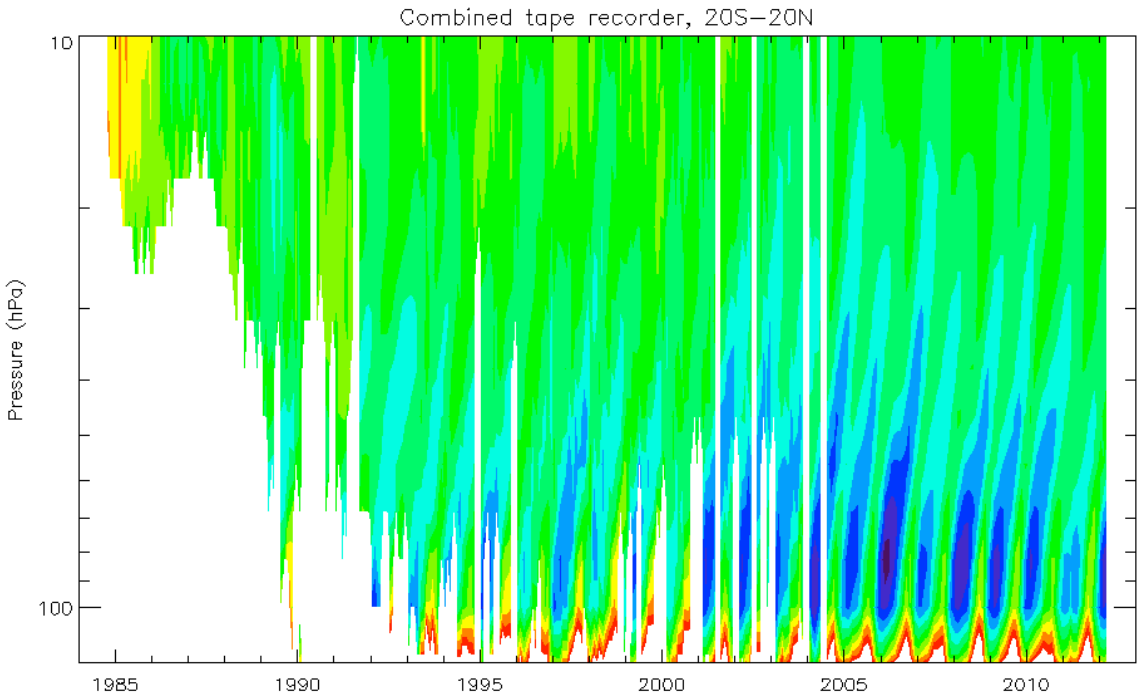
Now print both the corrected and uncorrected monthly-mean SAGE-II H2O values at this location

```
IDL> print, sage2h2oq[latind,pind,tind], sage2rawh2oq[latind,pind,tind]
      4.89006
      4.17777
```

Now plot the tropical tape recorder from the combined data set

```
tropicind = where( abs(lat) LT 20)
combinedttr= MEAN( combinedh2oq[tropicind,*,*], dimen=1, /nan)
contour,transpose(combinedttr), yrtime, level,
/ylog,/ystyle,yrange=[125,10.],xrange=[1984,2013],/xstyle,/cell_fill,le
vels=findgen(13)*.4+1.7,title='Combined tape recorder, 20S-
20N',ytitle='Pressure (hPa)',charsize=1.5
```

This should produce a plot that looks something like this (this plot produced with SWOOSH v2.0)



Caveats:

- This whole data set should be considered a "beta" product, and may still undergo significant changes in the future.
- The H2O product is most reliable between 100 hPa – 1 hPa, and is unlikely to change significantly in the future. Studies outside of this range should be treated with EXTREME caution, as there are known severe limitations of some of the instruments in the upper troposphere and mesosphere.
- Although extensive filtering has been done, the SAGE-II and HALOE data still seem to exhibit anomalous behavior around Pinatubo in the tropics, and SAGE-IImay have some issues in 1986-1987.

Changelog:

- 2012-01-10: Initial version released
- 2012-01-13: Fixed problem with *combinedanomh2oq, combinedanomo3q* variables
- 2012-01-27: Fixed data type of *yrtime* to be float in netCDF version
- 2012-01-27: Fixed data type of variables containing quantity "N" from floating-point to long-integer in netCDF version
- 2012-01-30: Fixed HALOE offsets (code was using obsolete version of offsets). New offsets are less than 10% different over range 0.1 hPa $\leq p \leq$ 100 hPa.
- 2012-02-01: Fixed problem with HALOE equivalent latitudes. All equivalent latitude-gridded HALOE variables are now fixed.

- 2012-02-01: Added offset variables that contain the additive offsets (as a fxn of height, latitude) applied to the HALOE and SAGE data
- 2012-02-02: Created isentropic gridded-data set
- 2012-02-03: Added additional SAGE aerosol filtering criteria for WV. All points in a profile below the point at which 1020 nm extinction ≥ 2 x 10^{-4} km^{-1} are removed. This filter removes anomalously high data around 10-50 hPa in 1986-1987, and anomalously high data around Pinatubo.
- 2012-02-07: Added additional HALOE aerosol filtering criteria to address anomalously low tropical values in 1992. For each profile, if NO-channel extinction ≥ 8 x 10^{-5} km^{-1} at 18 hPa, then this point and all points below are removed.
- 2012-02-08: Added UARS MLS WV data to help fill the gap during the Pinatubo time period
- 2012-04-30: Added minperbin in gridding process. Minperbin (currently running mperbin=10 for the 10° zonal-mean product, 5 for the 2.5° product, and 2 for the lon/lat grid) is the minimum number of profiles required per bin. If less than this number of profiles goes into the bin, the value is set to missing in that bin. This helps prevent a bin that only has a few profiles in it from dominating an average over a range of latitudes (e.g., when doing a 30S-30N average).
- 2012-06-04: Version 2.0 release
- 2013-02-13: Fixed bug in UARS MLS code. Added MERRA tropopause data to lon-lat and latpress versions of SWOOSH (lattheta version at a later date). Added HIRDLS H2O (not used in combined array) as a separate gridded variable.
- 2013-02-14: Version 2.1 release
- 2015-07-08: Version 2.2 released. Major change is that SAGE-II is now the reference measurement for ozone (previously was Aura MLS).
- 2015-08-27: Version 2.3 released. Removed HALOE beta angle correction bug in lattheta/latlon products. MERRA tropopause now saved as a structure in .sav versions. Now removing all HALOE data at and below 147 hPa, per discussion in SWOOSH paper.
- 2015-09-24: Version 2.4 released. Now using MLS v4. Bug fix -- Removed old MLS oscillation correction and made SAGE-II the reference in grid_swoosh_lonlat. SAGE-II WV data before 1986 now removed. Re-processing SWOOSH data with update to 7/2015. Including 2.5, 5, and 10 deg gridding.
- 2016-03-02: Version 2.5 released. 2 Major variable name changes -- 1) "sigma" variables are now called "stddev" to reduce confusion, 2) "meanunc" variables replaced with "rmssunc" (slightly different uncertainty calculation -- see SWOOSH manuscript). Also, metadata updates in netCDF versions of files.