Extended Atlantic Zone (XA) Monthly Production Using SNPP VIIRS chlor_a (OCI) Data

- author: George N. White III (Bedford Institute of Oceanography)

NOTE: For 2012–2018 VIIRS-N calibration is known to be poor. Improved calibration depends on accumulating sufficient buoy data and will trigger another full mission reprocessing.

This calculation updates the previous R2014.0.2 VIIRS OCI `chlor_a` production results to R2018.0 VIIRS OCI `chlor_a` and adds data for 2017 and 2018. SST data were R2016.0. A full mission calculation was required for 2017. A change in the storage format for SST required a change in the processing to R2016.1 starting with Sept. 2018.

Note that VIIRS SST software, calibration, and validation was done by Miami, but funding for this activity terminated in 2018.

Intel compilers were used on macOS El Capitan. In tests using an iMac Retina system, the `pp_region` program gave identical results in under half the time of GNU gfortran 7.0.

Introduction

This calculation uses 4km binned files, and extends previous calculations performed for AZMP to the Extended Atlantic Zone area. MODIS Aqua calibration in recent years is poor, and VIIRS is currently the preferred Ocean Colour sensor.

The calculation uses software developed at BIO and Dalhousie University. The general approach is that described in Platt et al (2008). Production values are computed using a numerical model for the spectrally-resolved underwater light field. Input parameters for variables not directly available from remote sensing for each satellite pixel are obtained using nearest-neighbor imputation.

PNG images

cppday monthly images

Time-series plots for AZMP Statistical Areas

Note that blank plots will be produced for areas where production could not be computed.

cppday time series

Details of the calculation

The basic methods have been documented in publications, so this document emphasizes the sources of data and practical considerations.

Resource requirements

The All Canadian Waters (AC) calculation performed in 2013 used SeaWiFS 9km binned data. For the AC region, the calculation used 394432 bins.

This calculation uses VIIRS 4km binned data. The Extended Atlantic Zone contains 342532 data bins, so somewhat fewer than for the AC region. Users wishing to use data for a specific area of interest may need to consider the resource impacts of the reduced bin size.

Excluding input data sets, the 2012-2016 VIIRS 4km XA calculation required 10 Gbytes for 5 years, or 2 Gbytes/year.

Input data sets

Citation for chlor_a and PAR data:

Feldman, G. C., C. R. McClain, Ocean Color Web, VIIRS Reprocessing 2018.0, NASA Goddard Space Flight Center. Eds. Kuring, N., Bailey, S. W. January, 2018. http://oceancolor.gsfc.nasa.gov/

Following the approach described in Platt et al (2008), input data sets required are:

- remotely sensed
  - chlor_a, used for imputation and in the production calculation
  - total daily PAR, used in the production calculation
  - SST, used for imputation
  - cloud fraction, used in the production calculation to partition PAR between diffuse and direct components.
- in situ
  - P–I parameters (PmB, alphaB) from BIOchem??
  - biomass profile parameters, B_0, h, z-m, and sigma from BIOchem??
- other
  - ETOPO1 bathymetry, used to exclude locations where the light at water-column depth exceeds 10% of surface light.

Note that NASA standard level-2 processing for R2018.0 uses the ice mask from the OISST ancillary files. The previous ice mask provided smooth edges while the new mask gives stepped edges and tends to mask more pixels.

Remotely sensed data sources

Where possible, the calculations use widely available public data sources in order to obtain the "network benefits" associated with heavily scrutinized data sets (e.g., we hope someone else will have noticed any problems before we encounter them).

chlor_a, SST, and PAR

Monthly level-3 4km binned files from NASA OBPG VIIRS reprocessing 2018.0 were used for chlor_a and PAR. The SST processing version was 2016.0 uo to Sept., 2018, and 2016.1 aftrwards. Extractions were done using SeaDAS 7.5 l3bindump.

Cloud

Cloud fraction data are used to partition the total PAR from MODIS Aqua into diffuse and direct components. Diffuse light is important when the sun is low, so could play are large role at high latitudes in the fall when there is less ice. The irradiance model, however is not reliable longer atmospheric path lengths associated with low sun elevations, so in practice the results that can computed with the current model are not highly sensitive to the choice of cloud fraction data.

MODIS Aqua Monthly CLD_FR data were downloaded in png format from http://neo.sci.gsfc.nasa.gov/view.php?datasetId=MYDAL2_M_CLD_FR:

"Cloud fraction is the portion of Earth's surface covered by cloud relative to the portion of Earth not covered by cloud. Cloud fraction is derived from the 1-km-pixel resolution cloud mask product made from radiance and reflectance measurements of Earth collected by the Moderate Resolution Imaging Spectroradiometer (MODIS) aboard NASA's Terra and Aqua satellites. The MODIS Cloud Mask product (MOD 35) is a daily, global product. The algorithm employs a series of visible and infrared threshold and consistency tests to

specify confidence that an unobstructed view of the Earth's surface is observed. An indication of shadows affecting the scene is also provided.

The cloud mask is generated by applying appropriate single field of view spectral tests to each pixel. These tests generally rely on thresholds and there are different algorithms for ocean, desert, land, etc. An individual confidence flag is assigned to each pixel test and these flags are combined to produce the final cloud mask flag. A group confidence value is assigned to the pixel and is based on this final flag, ranging from less than 0.66 (low confidence of an unobstructed view of the surface) to greater than 0.95 (high confidence that the pixel is clear). For values between 0.66 and 0.95, spatial and temporal continuity tests are further applied in order to determine whether the pixel is confident clear or confident cloudy. The cloud fraction is then calculated from 5 x 5-km cloud mask pixel groupings, i.e., given the 25 pixels in the group, the cloud fraction for the group equals the number of cloudy pixels divided by 25. This derived cloud fraction is a dataset in the MODIS cloud product (MOD06)."

http://eospso.gsfc.nasa.gov/atbd-category/47 http://modis.gsfc.nasa.gov/data/dataprod/pdf/MOD_06.pdf http://eospso.gsfc.nasa.gov/sites/default/files/atbd/atbd_mod06.pdf http://www.cossa.csiro.au/modis/nov02_ws/lecture6.pdf http://www.cossa.csiro.au/modis/nov02_ws/lecture6.pdf <http://views.cira.colostate.edu/data/Documents/Terra_MODIS_Level3/MOD08_D3.005/ Intro_to_Modis_Cloud_Products.pdf>

Hubanks, King, Platnick, and Pincus, 2008: MODIS Atmosphere L3 Gridded Product Algorithm Theoretical Basis Document. ATBD Reference Number: ATBD-MOD-30.

No MODIS Aqua PNG file was available for December, 2009, so the corresponding MODIS Terra file was used. November, 2009 CLD_FR files were compared between Aqua and Terra.

Outline of the calculations

For each monthly period, the `OPsetup` script is run in the directory $OP_HOME/data/YYYY/MM. The script determines the time-period of the calculation from the directory. It then extracts the required data from the various input files and generates a configuration file and a Makefile. Running make will, if all goes well, run a series of programs ending with the regional production calculation. Experience has shown that problems will occur. Input data sets are often spread across multiple systems, so a failure in the network or any one system will generally prevent the calculation from running. In most cases this should result in some helpful messages in the log files, but in some cases the problem that prevents a calculation from running may also block access to the log files.

The actual production calculation takes only a couple hours for each year, much less than the time required to download and extract the input data. In general, the calculations are dominated by I/O so care is needed to avoid I/O bottlenecks.

Experience has also shown that some input files may be modified as problems are found. The make program is used to track dependencies among files so that in most cases simply running make again will regenerate the affected files.

Areas for future improvement

The current calculation is designed to use only minimal changes from previous calculations. Among the many areas where improvements can be made:

- cross comparisons/validation of input data sets. NASA OBPG has done comparisons between sensors and reprocessing versions at global scales for the SeaWiFS chl data, but there can be regional impacts that these comparisons would not detect.
- the nearest neighbor imputation configuration should be reexamined for the newer data and larger area
- eliminate SeaDAS 6 dependencies. NASA has stopped development of SeaDAS 6 so it is unlikely to be usable on future platforms.
- replace Bird with a PAR model adapted to high zenith angles. A major effort will be needed to support ancillary data requirements for surface conditions (ice, waves) and atmospheric conditions.

ROI

The three AZMP regions are:

```
AClimit=[S=30,W=-179,N=85,E=-17]
AZlimit=[S=39,W=-71,N=62.5,E=-42]
XAlimit=[S=39,W=-95,N=82,E=-42]
```

The region of interest for this calculation is the Extended North Atlantic (XA) region. When extracting data with differing rasterizations or binning, these limits should be expanded slightly to ensure that pixels near the edge are not omitted.

Data sets

- Chlorophyll, PAR :: VIIRSN R2018.0 reprocessing

URL: https://oceancolor.gsfc.nasa.gov/reprocessing/r2018/viirs-snpp/

- SST :: VIIRSN R2016.0 and 2016.1 reprocessings (not separately documented)
- CLD_FR :: MODIS level-3 monthly

URL: http://neo.sci.gsfc.nasa.gov/view.php?datasetId=MYDAL2_M_CLD_FR

Computational notes

- platform: MacOSX Snow Leopard and Linux
- chl file format – using SeaDAS level-3 binned monthly HDF files

chlor_a, PAR, and SST are from the standard NASA R2018.0 processing global 4km files. Data were extracted for the AC ROI the SeaDAS 7.5 OCSSW Processing System `l3bindump` program.

- irradiance:

  PAR is taken from the standard NASA R2018.0 global processing, 4km level-3 binned files, and extracted using the same method as the chlor_a data.

  The PAR data is total PAR. Data from a single sensor are not considered reliable for time-binning periods shorter than a month.
  MODIS monthly cloud fraction data are used to partition PAR into diffuse and direct components.

- cloud fraction

  MODIS monthly CLD_FR data were be used. When using total PAR the effect on production when zenith angles are small is minimal. Since Bird's model is not reliable for high zenith, the impact is limited to already suspect values.

- visualization is done in post-processing for a number of reasons:
  - minimize the number of programs that must be installed on the system used for processing
  - reduce the effort needed to track changes in seadas image output (png top-down vs bottom up, license restrictions, etc.)
- Makefiles use GNU make

successful job completion for program XX is tracked using an empty file YYYYMMP_XX.ts (timestamp) created with a command of the form:

```
XX ... && touch YYYYMMP_XX.ts
```

Many of the programs are witten in Fortran where there has been no standard way to provide non-zero exit status, so in some cases a timestamp file may be produced after an error was encountered, but the software

attempts to handle most errors by generating a record with missing values and then continuing with the next record.

The Fortran programs use a version of the Slatex xerror handler. Xerorr maintains a count of each type of error and produces a summary table at the end of the log file for each job, for example:

```
$ tail 200207_pp.log
utime =     0.235E+04  stime =      0.165
utime per pixel =     0.920E-02  stime per pixel =      0.646E-06

          Error message summary
Library    Subroutine Message start             NERR     Level     Count
SeaWiFS    pp_day     light at bottom              6        -1    134243
SeaWiFS    pp_regio   pp_day error                29        -1     13684
SeaWiFS    pp_regio   sum bound                   27        -1        19
SeaWiFS    pp_regio   bad or missing data         30        -1     10849
```

Tools required (2018)

- environment modules – allows switching between different production calculations. Currently, lmod (lua modules) is supported by the US NSF.
- SeaDAS 7.5.3 OCSSW Processing System. Although the SeaDAS 7 GUI is available on Windows, the OCSSW processing system requires a unix-based OS (linux or macOS). Due to changes in storage formats, previous versions may not always produce correct results.
- Command-line utilities, including GNU awk, make, and sed
- Compilers for C, Fortran 9x, and C++. Intel compilers produce code that runs 2x faster than the GNU compilers
- git version control (used by the OCSSW Processing System)
- R
- tex

Environment modules

The environment modules package makes it simple to switch between different production caculations.

Example:

```
ambrosia:~ gwhite$ module avail OP
------------ /opt/modules/modulefiles ------------
------------ /Users/gwhite/privatemodules ------------
OP/AC/S2010.0  OP/GL/S2009.1  OP/default
OP/AZ/S5.1     OP/NWA/S2009.1 null
ambrosia:! gwhite$ module help OP/AC/S2010.0
----------------------------------------------------------------------
Module Specific Help for /Users/gwhite/privatemodules/OP/AC/S2010.0:

The AC/S2010.0 modulefile defines the default system paths and
environment variables needed to use the Ocean Production.
package with SeaWiFS reprocessing 2010.0 data for All Canadian Waters.
----------------------------------------------------------------------

ambrosia:~ gwhite$ module load OP/AC/S2010.0
```

```
ambrosia:~ gwhite$ cd $OP_HOME
ambrosia:S2010.0 gwhite$ which pp_region
/Users/gwhite/OP/AC/S2010.0/bin/pp_region
```

Compilers

The software has been tested on linux (amd64) using the GNU Compiler Collection. The production software is quite portable, but the
OCSSW Processing System is large and uses a number of system-specific extensions. In practice, the OCSSW Processing System is has often been necessary to compile from source on systems being used for the production calculations.

Software organization

Programs are named XX.bin in $PP_HOME/bin/$arch. Each XX.bin is run by a bash script, XX, in $PP_HOME/bin. These scripts perform consistency checks on the command-line arguments, ensure that the required files are present, and call the appropriate binary program for the architecture/OS with the appropriate arguments.

The environment modules system may be used to adjust paths and environment variables for the different production calculations.

The key programs are:

- ann_cp :: nearest-neighbor imputation for biomass profiles
- ann_pi :: nearest-neighbor imputation for Pmb and alpha
- nn_pre :: preprocessor used to prepare input for nearest-neighbor programs
- pp_fs :: used to scale biomass profile parameters to the satellite (surface) values
- pp_pre :: preprocessor used to prepare input for the production calculation
- pp_region :: regional primary production processor

Each calculation uses a master control file (created by the setup script) to provide the names of input and output files as well as parameters such as the date of the calculation.

Input data

The calculations are performed for each level-3 bin. NASA assigns each bin a unique number, so records are identified by bin number. For convenience the latitude and longitude are often included, but may also be determined from the bin number.

In principle, the mission-long PAR climatology would provide the complete list of bins, but this file is no longer produced, so the monthly climatology for September (`V20122452017273.L3b_MC_SNPP_PAR.nc`) was used.

For each input data set (monthly mean chl, SST, PAR, and CFR) values are extracted for each level-3 bin in the chl data to flat ASCII .csv files. These are then combined to create monthly LTDBI (location, temperature, depth, biomass, and irradiance) flat ASCII files.

In situ

There are often problems reading in situ data. The `isrd` program is intended to test the ability to read in situ data correctly:

The following example uses Intel ifc:

```
pippin:06 gwhite$ $OP_HOME/bin/Darwin/isrd.bin cp ctl/200206.ctl
pippin:06 gwhite$ echo $?
174
```

The `isrd.err` file is created during program statup before it has processed the control file. For regular processing, the control file defines the name of the log file.

```
pippin:06 gwhite$ cat isrd.err

Ocean Production: isrd $Revision: 1.1 $

                  2012-12-12  09:09:04.839


IEC60559  $Revision: $
 plusinf (1./0.):  Infinity                (IEC60559 infinity)
 nan (0./0.):  NaN                      (IEC60559 NaN)
 minusinf (log(0.)):  -Infinity                (IEC60559 -Infinity)
 plusinf .ne. plusinf =  F
 nan .ne. nan =  T
 minusinf .ne. minusinf =  F
 minusinf .ne. minusinf =  F
 (nan .lt. 0.d0) .or. (1.d0 .lt. nan) =  F
 .not. ((0.d0 .le. plusinf) .and. (plusinf .le. 1.d0)) =  T
 .not. ((0.d0 .le. nan) .and. (nan .le. 1.d0)) =  T

      isrd: ****** IEC 60559 NaN/Inf tests passed ******


isrd: finished processing control information
  errors logged to <isrd_cp.log>.
```

The control file was processed, so now the log file is used:

```
$ cat isrd_cp.log
Program: /Users/gwhite/OP/AC/S2010.0/bin/Darwin/isrd.bin

Ocean Production: isrd $Revision: 1.1 $

                  2012-12-12  09:09:04.839

    [...]
```

There are several tests for invalid data, but as the in situ data set grows it may be necessary to refine these tests.

chlor_a, SST, and PAR

For each year, `l3bindump` is used to extract the ROI from standard NASA OBPG monthly binned files to produce a monthly .csv files for each variable. These are named `$OP_HOME/data/<l3varname>/YYYYMMM_<l3varname>.csv`. Note that `OPsetup` assume that commas are used as separators. Data extracted using ESA GPT use tabs, so must be translated. This is easily done using `sed`:

```
sed -i-tab -e 's/^I/,/' *.csv
```

LDTBI Files

These files provide values for location (as Lon, Lat), depth, temperature, biomass, and irradiance (as PAR, CFR). For each month, a file named `YYYYMM_LDTBI.csv` is created by merging individual `.csv` files in the setup script. The `.csv` header is:

`Name,Longitude,Latitude,height,sst,chlor_a,par,CLD_FR`

Here, `Name` is the bin number and is used to that the format is compatible with BEAM position files.

Cloud fraction

Because cloud fraction is used only to split total PAR into diffuce and direct components, production results are not highly sensitive to the small changes in cloud fraction values. The MODIS monthly cloud fraction product (MYDAL2_M_CLD_FR) is derived from a 1-km resolution cloud mask, defined as `CLD_FR=N/25`, where `N` is number of pixels masked as cloud in a 5x5 box. Previous investigations have shown that this product correlates well with legacy cloud fraction data.

In 2016 the MODIS `CLD_FR` files were replaced with new versions. There are significant differences in many parts of the world, mostly southern hemisphere. New comparisons with legacy cloud fraction data could be useful.

Example URL:

http://neo.sci.gsfc.nasa.gov/servlet/RenderData?si=1711658&cs=gs&format=PNG&width=3600&height=1800

Output files

In general, `.csv` files contain data. There will be corresponding `.log` and `.ts` files. There are also `.err` files which are used to provide error reports from the startup phase of the processing programs, e.g., before a log file has been opened.

- Files in the `nn_cp` and `nn_pi` directories:
  - biomass profile inputation `YYYYMM-CP-YYYYMMDD.*`
  - P–I parameter imputation `YYYYMM-PI-YYYYMMDD.*`
- Files in the `pp` directory:
  - preprocessor files `YYYYMM_pp_pre.*` provide basic input to the regional production calculation. In the files created by the `pp_pre` processor, profile parameters are not yet scaled to the surface (chlsat) values. When using cloud, Idifday is always set to a missing value code.
  - scaled profile parameters `YYYYMM_pp_fs.*`
  - results of the production calculation `YYYYMM_pp.*`

Headers

The ASCII .csv file headers for the generated files are:

- production

  $ head -1 *.csv ==> 200206_pp.csv <== chlsur,cppday,zpday

- scaled biomass profile parameters

  ==> 200206_pp_fs.csv <== h_sc,B0_sc

- input data, including imputed values of in situ parameters

  ==> 200206_pp_pre.csv <== bin,lon,lat,depth,SST,chlsat,Itotday,Idifday,cloud,PmB,alpha,h,sigma,zm,rho

Processing

The input data are stored on Data-ambrosia (local data cache).

Setup

$ do-setups $(gseq 2012 2018)

Calculation

The `do-pp` script starts calculations for all months in the given year, using `nohup make &` in each monthly directory. This should be adapted to use GNU parallel, but at present, run one year at a time, e.g.:

```
$ do-pp 2012
```

The final target in the `Makefile` for each month is a NetCDF file:

`$OP_ROOT/data/YYYY/MM/pp/YYYYMM_cppday.nc`

NetCDF4 files with mapped `cppday` data are created by the `OPcsv2OPnc` script, which uses R rasterVis.

Visualization

The monthly images are generated in the `$OP_ROOT/images` directory using NCL (`OPnc2png-web.sh`) to generate PNG images suitable for web viewing:

```
$ cd $OP_ROOT/netcdf
$ do-nc $(gseq 2012 2018)
$ cd $OP_ROOT/images
$ do-png $(gseq 2012 2018)
```

Note: NCL is deprecated and only runs on POSIX (e.g., linux and MacOS) platforms – python+cartopy can replace NCL going forward and runs on both POSIX and Windows platforms.

Time-series and statistics

GPT StatisticsOP from ESA SNAP is used with shapefiles for the AZMP regions. The xml files that define each product and shapefiles are installed on ambrosia. For each netcdf file in `$OP_ROOT\netcdf` a file `YYYYMM_cppday_geometries.stats` is generated by the `doStats.sh` script. This script takes a list of netcdf files. For the current calculation, data prior to 2018 are not changed, so existing `.stats` files are used.

```
$ cd $OP_ROOT/StatisticsOP
$ doStats.sh  cppday.xml AZMP_StatisticalAreas/geometries \
cppday-2018.list | tee doStats-2018.log
$ tsStats.sh  cppday.xml AZMP_StatisticalAreas/geometries \
cppday.list > cppday.tsdata
```

RStudio is used to process the `tsplot.Rmd` file to create a PDF document with time-series plots of monthly means for each region.

Programming notes

Most programs are stored in the `$OP_ROOT/bin/<arch>` directory and called from wrapper scripts in the `$OP_ROOT/bin` directory. These scripts check the runtime architecture and call the appropriate binaries. They also perform checks on arguments and, when invoked with `-h` or `--help` as the first argument, provide a

help page modelled on unix man pages. These help pages also identify the locations of resources used by the scripts.