AnalysisReport

CGD's data reconstruction effort - Adam Phillips (9/10/08)

In August 2008 CGD started a project to identify and correct errors in the post-processed **daily/monthly** Columbia run data.
The corrected daily data is located on hurricane: /datalocal/nrcm-cgd/asphilli/Columbia-daily
The corrected monthly data is also located on hurricane: /datalocal/nrcm-cgd/asphilli/Columbia-monthly

There are a few caveats to keep in mind when using the corrected data:

1. The data reconstruction effort detailed below corrected the most obvious errors in the post-processed monthly/daily data. There is no way to be sure that all errors have been eliminated from every field.
2. When analyzing the CTTOP, CPTOP, and CTAUCLD variables, keep in mind that when there is no cloud, the fields will have -9999 (=_FillValue /missing_value). This effect is highly compounded when averaging over an entire month. It is recommended that if you wish to analyze these fields that you do not use the monthly data. There are also three dates that should be avoided when looking at the daily data for these variables, 2003-11-06, 2005-03-25, and 2005-03-26, as the data on these dates is highly suspect. For more information about what additional corrections had to be taken for these variables please see below.
3. The Time variable has been renamed to time for the monthly data, in accordance with netCDF conventions.

Data Reconstruction Details

Errors in the original post-processed daily data were initially noticed when looking at global average timeseries. Spikes in the timeseries were evident across most daily variables. In some cases the spikes were small, while in others the values approached 1e20. To automatically identify the dates where there were problems, a two step script was developed. The first step involved identifying dates where the global average timeseries value was 5-10 times greater (or less) than the timeseries mean. The second step involved identifying dates where the derivative of the timeseries exceeded a realistic value. These two tests yielded a list of 69 dates that had outliers in one of more variables:

2000-06-28, 2000-08-06, 2000-08-31, 2000-09-30, 2001-01-31, 2001-03-31, 2001-10-24, 2003-04-24, 2003-09-08, 2003-11-04, 2003-11-05, 2003-11-25, 2003-12-09, 2003-12-14, 2004-01-14, 2004-02-11, 2004-02-28, 2004-03-22, 2004-03-31, 2004-04-23, 2004-05-01, 2004-05-24, 2004-06-01, 2004-06-21, 2004-06-26, 2004-07-17, 2004-07-31, 2004-08-21, 2004-08-31, 2004-09-16, 2004-09-22, 2004-09-23, 2004-09-30, 2004-10-23, 2004-10-31, 2004-11-19, 2004-12-12, 2005-01-03, 2005-01-09, 2005-01-19, 2005-02-11, 2005-03-02, 2005-03-09, 2005-03-25, 2005-04-01, 2005-04-16, 2005-05-01, 2005-05-15, 2005-05-23, 2005-05-24, 2005-05-26, 2005-05-29, 2005-06-01, 2005-06-24, 2005-07-16, 2005-07-19, 2005-08-01, 2005-08-05, 2005-08-10, 2005-08-23, 2005-08-31, 2005-09-20, 2005-09-30, 2005-10-23, 2005-11-04, 2005-11-17, 2005-11-30, 2005-12-12, 2005-12-13

The list of days includes all of the days listed by Cindy below, along with an additional 6 days. Note that errors were seen across many variables; they were not confined solely to the accumulated fluxes. Spatial plots of the identified days were examined, and the errors ranged from being extremely subtle to grids that were totally incorrect. Due to the fact that it was not possible to ensure that every variable for a particular date was an outlier or not, it was determined that **all** variables would be replaced for the 69 dates through standard linear interpolation.

Other steps that were taken:

*   In the original post-processed data, the CPTOP, CTTOP, and CTAUCLD variables were set to -9999. when a cloud was not present, but the proper _FIllValue / missing_value attributes were not set. _FillValue and missing_value attributes were thus added for the CPTOP, CTTOP, and CTAUCLD variables in every daily/monthly netCDF file. Furthermore, values less than zero were identified on 2003-11-06 and 2005-03-26, and these values were set to missing.
*   Negative values in RAINC and RAINNC were found on these dates: 2001-08-24, 2001-08-25, 2005-05-15, 2005-05-16, 2005-05-23, 2005-05-24, 2005-05-25, 2005-05-26, 2005-08-05, 2005-08-06. For these dates, RAINC and RAINNC were linearly interpolated from surrounding data.
*   New monthly averages were calculated from the corrected daily data and placed on hurricane.
*   As of 3/5/10, the Time and Times variables are now correct for both daily and monthly data. The Time variable has been renamed to time in accordance with netCDF conventions for the monthly data only.

Report on data with problems - Cindy Bruyere

*   A number of files for the Columbia runs are corrupt.
    All these problems were due to transfer errors between the Columbia and NCAR computers.
    As the files appear to have the right size, these problems were only discovered during the analysis step.

    For these times, the wrfout files were re-created and placed on the web with extension _interpolated.
    These files were re-created by interpolating from the day before and after the missing time.
    As the accumulated fluxes could not be re-created in this way, the fluxes were set to -9999., in the new wrfout files and two additional files with extensions _6hFLX.nc were created. These FLX files contain estimates of the 6 hour accumulated flux fields before and after the missing time period.

    Below is a list of times that were re-created:
    *   2000-06-28_18
    *   2000-08-06_18
    *   2000-08-31_18
    *   2001-01-31_18
    *   2001-03-31_18
    *   2003-04-24_12
    *   2003-09-08_06 - 2003-09-08_21
    *   2003-11-04_18 - 2003-11-05_21
    *   2004-03-22_12
    *   2004-04-01_00
    *   2004-09-17_00

- 2004-09-23_00
- 2004-10-31_18
- 2005-03-26_00 - 2005-03-26_12
- 2005-05-16_00
- 2005-05-16_06
- 2005-05-24_00
- 2005-05-26_00
- 2005-05-29_18
- 2005-05-30_00
- 2005-08-01_12
- 2005-08-06_00
- 2005-08-06_12
- 2005-08-11_00
- 2005-08-31_18
- 2005-11-04_18
- 2005-12-12_12 - 2005-12-13_06

- Restart related problems
  A second type of problem was discovered in the 06Z files directly after a restart.
  The reason for this is not 100% clear, but it seems as if the radiation computations are temporarily corrupt just after a restart.

  These problems:
  - Only appear after Nov 25, 2003
  - Only affects the accumulated flux data at the specified times

    Below is a list of the times affected:
  - 2003-11-25_06
  - 2003-12-09_06
  - 2003-12-14_06
  - 2004-02-28_06
  - 2004-04-23_06
  - 2004-05-01_06
  - 2004-05-24_06
  - 2004-06-01_06
  - 2004-06-21_06
  - 2004-06-26_06
  - 2004-07-17_06
  - 2004-07-31_06
  - 2004-08-21_06
  - 2004-08-31_06
  - 2004-09-30_06
  - 2004-10-23_06
  - 2004-11-19_06
  - 2004-12-12_06
  - 2005-01-03_06
  - 2005-01-09_06
  - 2005-01-19_06
  - 2005-02-11_06
  - 2005-03-02_06
  - 2005-03-09_06
  - 2005-04-01_06
  - 2005-04-16_06
  - 2005-05-01_06
  - 2005-06-01_06
  - 2005-06-24_06
  - 2005-07-16_06
  - 2005-07-19_06
  - 2005-09-20_06
  - 2005-09-30_06
  - 2005-10-23_06
  - 2005-11-17_06
  - 2005-11-30_06

Please send any reports concerning problems with data on mass store to bruyerec@ucar.edu