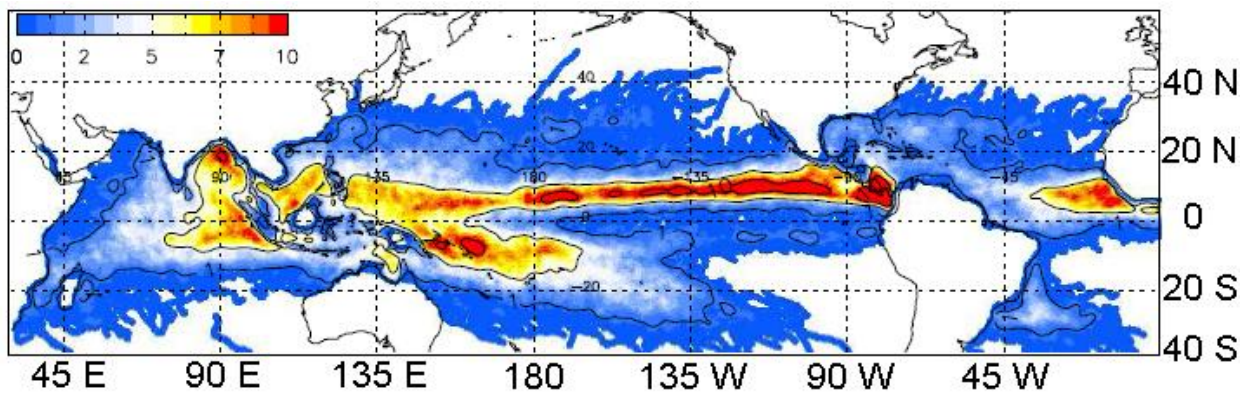


Global Tropical Cloud Cluster (TCC) Dataset

v01r01



Christopher C. Hennon
Chip Helms

UNC Asheville
August 2011

Table of Contents

1. The TCC Algorithm

- a. Data*
- b. Methodology*
- c. Settings*

2. Data Formats

- a. NetCDF*
- b. CSV*

3. Software to Work with the Data

4. Output Variables

5. Known Bugs and Issues

6. Support

7. Availability and Citation

8. References

1. The TCC Algorithm

All tropical cyclones (TCs) form from areas of sustained convection over warm tropical waters. Thousands of these convective “seedlings”, or tropical cloud clusters (TCCs), form in any given year yet only a small fraction develop into TCs. This algorithm was developed to identify TCCs and track their evolution in both movement and intensity.

a. Data

We use two datasets to create the TCC data: GridSat (Knapp et al. 2011) and IBTrACS (Knapp et al. 2010).

GridSat is a near-global satellite dataset that provides a homogeneous set of IR (and other channels not used here) observations. GridSat uses geostationary data from 32 satellite platforms, including GOES (1-13), Metsat (2-9), GMS (1-5), and others. The homogeneity comes from the application of numerous calibration efforts; essentially, GridSat can be thought of as a satellite reanalysis. GridSat version v01r01 was used in this version of the TCC data.

IBTrACS is a global best-track collection from the world’s tropical cyclone forecast centers. We use IBTrACS to identify TCCs that have developed into or are already TCs. IBTrACS version v03r02 was used in this version of the TCC data.

b. Methodology

There have been many criteria used to objectively determine the nature of a TCC. This algorithm uses criteria based on the work of Lee (1989) and adopted by Hennon and Hobgood (2003). Figure 1 shows a flow chart of the process we employ to identify TCCs. More details can be found in Hennon et al. (2011).

The objective criteria for TCC identification are:

1. Size and Intensity – Candidates are first identified by their IR brightness temperature (T_b) and size. A 3-hourly GridSat scene is filtered such that only pixels meeting a pre-determined T_b threshold pass through. The thresholds were determined by an analysis of samples of T_b from each basin’s TC season; the values used are shown in Table 1. Groups of acceptable pixels are then analyzed to determine a maximum radius from the geometric center of the group. This radius must be at least 1° (~ 111 km) in order for the group to be considered large enough to be a TCC. Furthermore, below-threshold cloud pixels must make up an area of at least $34,800 \text{ km}^2$ ($\sim 90\%$ of the area of a 1° radius circle) to filter out non-circular convective structures.
2. Location – The algorithm filters out candidate TCCs that form poleward of 30° latitude to minimize the inclusion of baroclinic systems into the data. We believe that a small number of legitimate TCCs may be lost. We also apply a land mask to eliminate any cluster centered over land, since tropical cyclogenesis cannot proceed there.

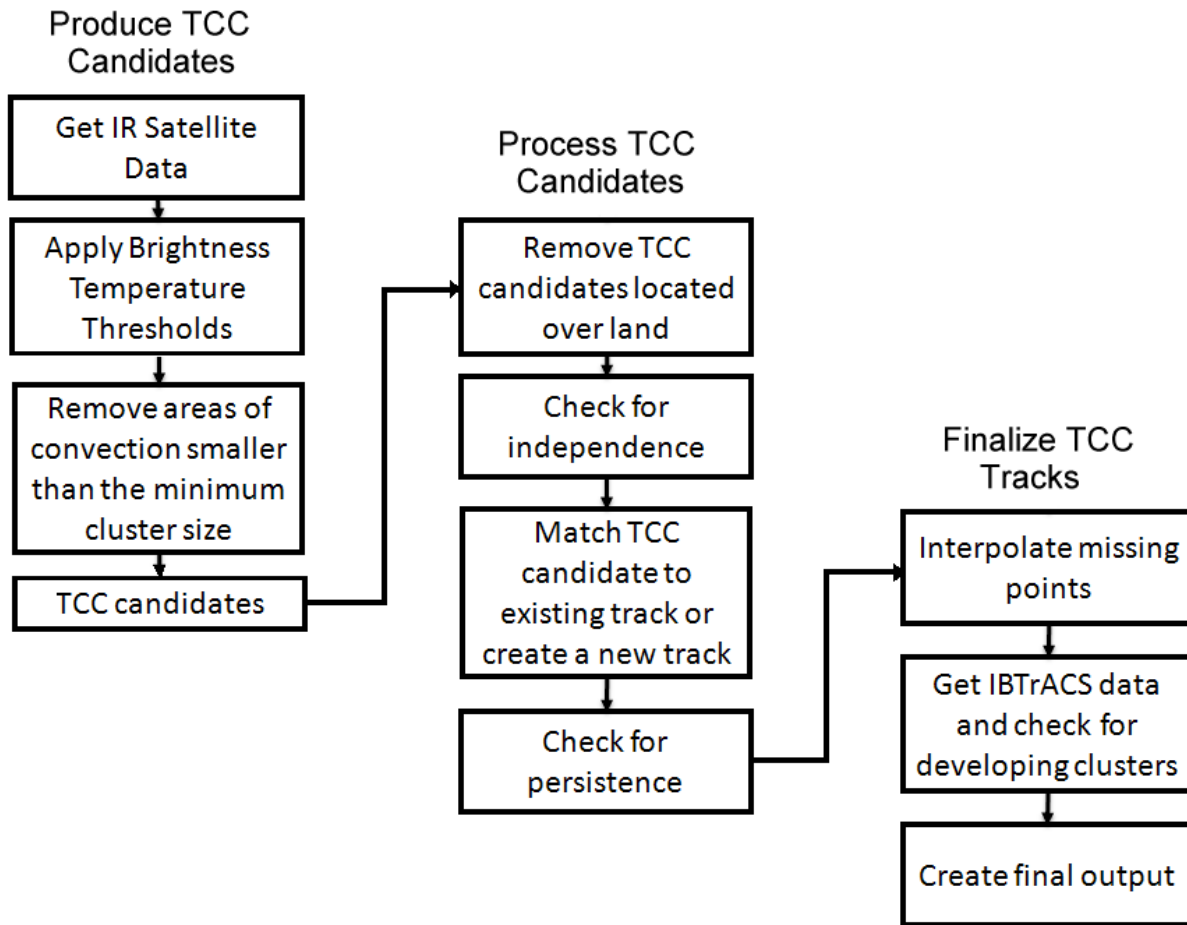


Figure 1. Procedure for the TCC algorithm. Flow chart proceeds left to right.

3. Independence - TCCs can be complex systems with multiple and distinct areas of intense convection. It is common to find several TCC candidates within close proximity to one another. In these situations, the algorithm considers all of the convective maxima subsets of the largest (“parent”) cluster, as long as they are within 1200 km of the parent. If a candidate is more than 1200 km away, then it is considered a distinct TCC.

4. Persistence – Once a TCC candidate is identified, the algorithm searches back in time to determine if the current candidate is new or had formed previously. A search radius is used (see Table 2) to find the candidate at a previous time. Candidates are allowed to “disappear” for up to 12 hours before regenerating somewhere within a pre-determined search radius to allow for convective fluctuations common in these systems. A candidate must persist for at least 24 hours in order to be counted as a TCC; this is to eliminate diurnal and other shorter time scale convection that may not be part of an organized synoptic-scale disturbance. Exceptions to this are made if a candidate TCC forms into a tropical cyclone within 24 hours of the candidate’s formation. In this case, the TCC will appear in the dataset.

Several post-processing steps are then performed, including: the determination of the development status, interpolation of missing TCC track points, quality control checks, and track smoothing (optional). Users are encouraged to read Hennon et al. (2010) for caveats, assumptions, and more details on the algorithm.

c. Settings

The following settings were used for the TCC data (v01r01):

Basin Boundaries

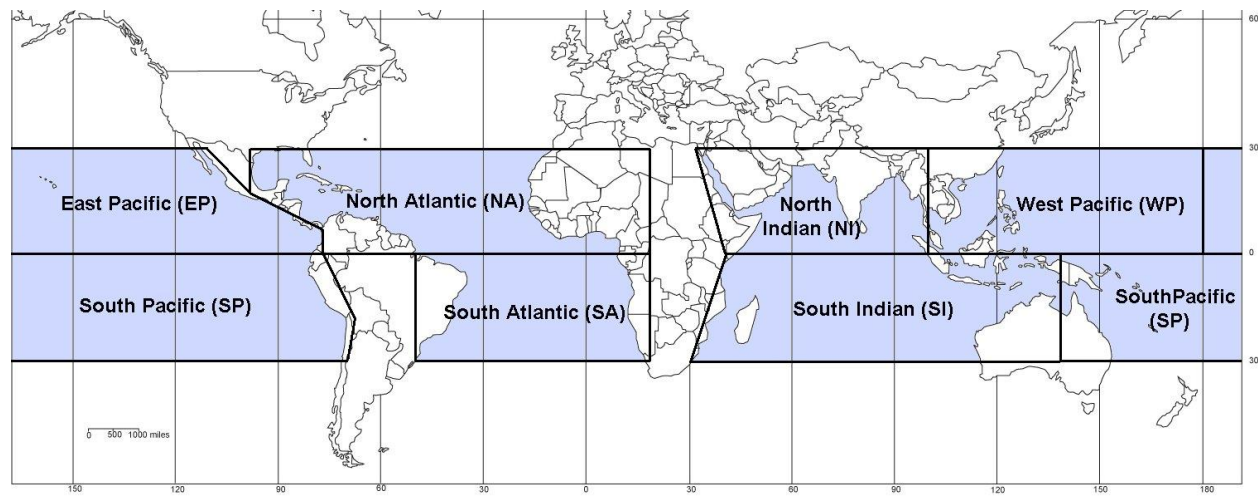


Figure 2. Basin boundaries used in the TCC algorithm.

Brightness temperature thresholds

TABLE 1. Brightness temperature (T_b) thresholds by basin

<i>Basin</i>	<i>Threshold (K)</i>
North Atlantic (NA)	224
South Atlantic (SA)	227
East Pacific (EP)	228
South Pacific (SP)	221
West Pacific (WP)	219
North Indian (NI)	218
South Indian (SI)	221

Algorithm Search Radii

TABLE 2. Algorithm search radii (km) for detecting the previous location of the current TCC candidate.

<i>Hours since last fix</i>	<i>Circular search radius</i>
3	450
6	550
9	600
12	650

2. Data Formats

The algorithm creates the following output data formats:

1. NetCDF file (1982-2009)
2. NetCDF files by individual year (1982-2009)
3. Comma Separated Variable (CSV) files by year (text, 1982-2009)

The NetCDF files contain all output variables created by the algorithm. The CSV files contain a subset of those data. The individual year files contain TCCs whose *first* fix time occurred in that year.

a. NetCDF

NetCDF was developed by Unidata, a division of the University Corporation for Atmospheric Research (UCAR). It is a data model for array-oriented scientific data access and is completely machine independent. NetCDF is popular in the atmospheric science community and is also widely used in other disciplines such as oceanography and physics (Brown et al. 1993). See <http://www.unidata.ucar.edu/software/netcdf/> for more information and downloads.

The “ncdump” utility can be used to quickly view the information and data contained in a netCDF file. To view the header information for the file “year.2005.cluster.nc”, type:

➤ `ncdump -h year.2005.cluster.nc`

This will return the following (abbreviated):

```
netcdf year.2005.cluster {
dimensions:
    nclus = UNLIMITED ; // (1564 currently)
    record = 300 ;
    ncharsn = 13 ;
    ncharnm = 36 ;
variables:
    int clusno(nclus) ;
        clusno:long_name = "Cluster number" ;
    char clus_sn(nclus, ncharsn) ;
        clus_sn:long_name = "Cluster serial number" ;
        clus_sn:key = "YYYYJJJHTTNNN\n",
            "YYYY - year of first position of the cluster\n",
            " JJJ - day of year of first position of the cluster\n",
            "   H - hemisphere in which the cluster began (N=Northern,
S=Southern, E=Equator)\n",
            "   TT - rounded absolute value of the initial latitude of
the cluster (range: 0-90)\n",
            "   NNN - rounded value of the initial longitude of the
cluster (range: 0-359, 0=prime meridian)" ;
    double time(nclus, record) ;
        time:long_name = "Modified Julian Day" ;
```

The ‘dimensions’ show the number of clusters in the file (“nclus”, 1564), the number of records or times allowed for each cluster (“record”, 300 max) and some character limits for the variable names. The ‘variables’ are then listed, along with some key information for each including names, formats, scale factors, and descriptions.

To view a particular variable, use the ‘-v’ argument with ncdump:

➤ `ncdump -v storm_name year.2005.cluster.nc`

This will return the header again, followed by the “storm_name” for each cluster:

data:

```
storm_name =
    "n/a",
    "n/a",
    "n/a",
    "n/a",
    "n/a",
    "n/a",
    "n/a",
    "n/a",
    "KERRY",
    "n/a",
    "n/a",
    "n/a",
```

In the example above, we can see that the first 8 TCCs in 2005 did not have an associated name with them (since they did not develop into a TC), but that the 9th TCC eventually developed in TC KERRY.

See Section 4 for a list of available variables in the NetCDF data.

b. CSV

Comma Separated Variable (CSV) format is a common way of structuring text data to be read by software or imported into statistical software for processing. CSV files do not contain the entire suite of output variables (see section 4), but may be more convenient to work with. An example of one track:

```
Global Tropical Cloud Clusters -- Version: v01r01
Data derived from GridSat (v01r01) and IBTrACS (v03r02)
clusno,clus_sn,time,lat,lon,wlat,wlon,devflag,initial_basin,in_season,lifespa
n,storm_sn,storm_name,genesis_basin,interp_flag,pixel_count,t_avg,t_min,t_std
,rad_avg,wgt_rad_max,avg_cloud_height,direction,wgt_direction,speed,wgt_speed
,qc_flag,qc_flag_interp,qc_flag_dev
1, 2004361S05074, 2004122600, -5.3, 74.0, -5.4, 74.3, 0, 6, 1,
30, xxxxxxxxxxxxxxx, n/a, 14, 0, 487, 210.2, 189.2, 5.8, 110.1,
227.7, 14.05, 119.9, 108.6, 16.1, 18.9, 0, 0, 0
1, 2004361S05074, 2004122603, -5.7, 74.7, -5.7, 75.2, 0, 6, 1,
30, xxxxxxxxxxxxxxx, n/a, 14, 1, -999, -999.9, -999.9, -999.9, -999.9, -
999.9, -99.99, 122.9, 107.6, 16.6, 19.9, 0, 0, 0
1, 2004361S05074, 2004122606, -6.2, 75.4, -6.0, 76.2, 0, 6, 1,
30, xxxxxxxxxxxxxxx, n/a, 14, 1, -999, -999.9, -999.9, -999.9, -999.9, -
999.9, -99.99, 122.9, 107.6, 16.6, 19.8, 0, 0, 0
1, 2004361S05074, 2004122609, -6.6, 76.1, -6.3, 77.1, 0, 6, 1,
30, xxxxxxxxxxxxxxx, n/a, 14, 0, 3681, 209.3, 193.9, 5.3, 312.2,
634.1, 14.15, 81.5, 99.5, 22.7, 13.0, 0, 0, 0
1, 2004361S05074, 2004122612, -5.9, 77.4, -6.1, 76.8, 0, 6, 1,
30, xxxxxxxxxxxxxxx, n/a, 14, 0, 1752, 209.3, 191.0, 5.9, 194.7,
421.4, 14.17, 7.0, 315.1, 26.7, 11.4, 0, 0, 0
1, 2004361S05074, 2004122615, -5.8, 76.2, -5.5, 76.3, 0, 6, 1,
30, xxxxxxxxxxxxxxx, n/a, 14, 0, 1325, 208.5, 194.1, 5.6, 158.2,
288.4, 14.27, 298.8, 318.5, 17.1, 12.0, 0, 0, 0
1, 2004361S05074, 2004122618, -5.3, 76.3, -5.2, 76.0, 0, 6, 1,
30, xxxxxxxxxxxxxxx, n/a, 14, 0, 1592, 209.5, 197.7, 4.9, 154.5,
286.1, 14.11, 288.4, 283.0, 21.1, 14.2, 0, 0, 0
1, 2004361S05074, 2004122621, -5.3, 74.7, -5.2, 75.0, 0, 6, 1,
30, xxxxxxxxxxxxxxx, n/a, 14, 0, 1439, 209.6, 197.7, 4.2, 172.1,
357.4, 14.10, 270.1, 265.9, 16.9, 14.1, 0, 0, 0
1, 2004361S05074, 2004122700, -5.3, 74.6, -5.3, 74.6, 0, 6, 1,
30, xxxxxxxxxxxxxxx, n/a, 14, 0, 2046, 209.5, 193.4, 4.8, 185.4,
321.9, 14.12, 270.0, 254.7, 7.0, 11.4, 0, 0, 0
1, 2004361S05074, 2004122703, -5.3, 74.0, -5.5, 73.9, 0, 6, 1,
30, xxxxxxxxxxxxxxx, n/a, 14, 0, 1853, 212.1, 200.6, 3.5, 156.5,
288.9, 13.72, 286.0, 288.5, 22.1, 20.8, 0, 0, 0
1, 2004361S05074, 2004122706, -4.7, 72.5, -4.7, 72.8, 0, 6, 1,
30, xxxxxxxxxxxxxxx, n/a, 14, 0, 831, 212.2, 200.1, 3.5, 127.7,
260.3, 13.70, 291.8, 306.1, 32.2, 27.2, 0, 0, 0
```


3. Software to Work with the Data

There are several IDL programs available to search the data and display tracks. An IDL license is required. At some point in the future, we plan to make IDL .sav programs available, which can be run without an IDL license.

Please send an email to Chris Hennon (chennon@unca.edu) if you are interested in the software.

4. Output Variables

The NetCDF data contain the full suite of output variables (54 in all). The CSV files contain a subset of the NetCDF data; those rows are shaded (29 in all). Variables are listed in the order that they appear in the NetCDF files.

<i>Variable Name</i>	<i>Description</i>	<i>Units</i>
clusno	Cluster number. '1' is the first TCC in the time period.	
clus_sn	Cluster serial number. A unique identifier, comprised of the following: YYYYJJJHTTNNN where: YYYY – Year of first position of TCC JJJ – Julian day of first position of TCC H – Hemisphere (N,S,E=Equator) TT – Rounded latitude of first position (0-90) NNN – Rounded longitude of first position (0-359)	
time	Julian day for each fix	
lat	Latitude of the geometric center of the TCC (south is negative)	°N
lon	Longitude of the geometric center of the TCC (west is negative)	°E
wlat	Latitude of the TCC weighted towards the most intense convection (south is negative)	°N
wlon	Longitude of the TCC weighted towards the most intense convection (west is negative)	°E
conv_lat	Latitude of the coldest cloud pixel in the TCC (south is negative)	°N
conv_lon	Longitude of the coldest cloud pixel in the TCC (west is negative)	°E
lat_for_mapping	Latitude of the TCC that was smoothed with the times on either side using a 3-point running mean (south is negative)	°N
lon_for_mapping	Longitude of the TCC that was smoothed with the times on either side using a 3-point running mean (west is negative)	°E
devflag	Indicates whether the TCC developed into a tropical cyclone. 0 = No, 1 = Yes	

initial_basin	<p>A number that gives the ocean basin corresponding to the first time in the TCC track (see Fig. 2 for basin geography) boundaries)</p> <p>0 = North Atlantic (NA) 1 = South Atlantic (SA) 2 = West Pacific (WP) 3 = East Pacific (EP) 4 = South Pacific (SP) 5 = North Indian (NI) 6 = South Indian (SI) 14 = Missing (MM) 15 = Undefined (usually for TCC on the equator)</p> <p>Numbers 7-13 are currently not used.</p>	
num_basin	Number of different basins the TCC existed over in its lifetime	
basin	The ocean basin that the TCC is over at the given track time. See 'initial_basin' for key.	
season	Year (YYYY format) based on the season that the TCC formed. Northern hemisphere season begins 1 January. Southern hemisphere season begins 1 July of the prior year.	
in_season	<p>Flag that indicates whether a TCC formed in the normal TC season for that basin.</p> <p>0 = Formed outside of the normal season 1 = Formed within the dates of the normal season</p> <p>The seasons are defined as:</p> <p>NA: June – November SA: No active season (TCCs always flagged '0') WP: April – January EP: May – November SP: November – April NI: April – December SI: November – April</p> <p>Official seasons that include only a portion of the month (e.g. 15 May) are expanded to include the entire month.</p>	
lifespan	Time between first and last fixes for the TCC. The last fix may be due to dissipation or it may be the genesis time for a developing TCC.	hours
record_count	Number of 3-hourly fixes (or interpolated fixes) for the TCC	
interpolation_count	Number of 3-hourly fixes that were interpolated rather than determined from a satellite image.	
storm_sn	For developing TCCs, this is the storm serial number of the tropical cyclone from the IBTrACS database	
storm_name	For developing TCCs, this is the storm name of the tropical cyclone from the IBTrACS database	
genesis_time	For developing TCCs, the time of genesis (Julian Day)	

genesis_lat	For developing TCCs, the latitude of the matched IBTrACS storm at the time of genesis (south is negative)	°N
genesis_lon	For developing TCCs, the longitude of the matched IBTrACS storm at the time of genesis (west is negative)	°E
genesis_basin	For developing TCCs, a number that corresponds to the basin in which genesis occurred. See 'initial_basin' for the key.	
genesis_dist	How far the TCC fix was from the matching IBTrACS storm at the time of genesis. Ideally this should be as close to zero as possible.	km
storm_dist	Average distance from the post-genesis TCC track to the matching storm track in IBTrACS. The post-genesis TCC tracks are not included in the database (the genesis point is the last point in a developing TCC track)	km
interp_flag	Flag showing whether the current fix was from the satellite image or interpolated from surrounding points. Track points are interpolated when the TCC does not meet T_b threshold requirements at that time. Other variables are not calculated when a fix is interpolated. 0 = Fix from satellite data 1 = Interpolated fix	
pixel_count	Number of satellite pixels that make up the TCC. This variable is one indication of size.	
t_avg	Average IR brightness temperature for the TCC	K
t_min	The minimum IR brightness temperature for the TCC	K
t_std	The standard deviation of IR brightness temperature for the TCC.	K
t_med	The median IR brightness temperature for the TCC	K
t_5per	5 th percentile IR brightness temperature. 95% of TCC pixels are warmer than this value.	K
t_10per	10 th percentile IR brightness temperature. 90% of TCC pixels are warmer than this value.	K
rad_max	Largest distance from the 'lat'/'lon' fix (e.g. geometric center) to the edge of the TCC around azimuth	km
rad_min	Smallest distance from the 'lat'/'lon' fix (e.g. geometric center) to the edge of the TCC around azimuth	km
rad_avg	Average distance from the 'lat'/'lon' fix (e.g. geometric center) to the edge of the TCC around azimuth. This is generally a good indication of the size of the TCC	km
wgt_rad_max	Largest distance from the 'wlat'/'wlon' fix (e.g. weighted center) to the edge of the TCC around azimuth	km
wgt_rad_min	Smallest distance from the 'wlat'/'wlon' fix (e.g. weighted center) to the edge of the TCC around azimuth	km
wgt_rad_avg	Average distance from the 'wlat'/'wlon' fix (e.g. weighted center) to the edge of the TCC around azimuth. This is generally a good indication of the size of the TCC	km
conv_rad_max	Largest distance from the 'conv_lat'/'conv_lon' fix (e.g. convective center) to the edge of the TCC around azimuth	km
conv_rad_min	Smallest distance from the 'conv_lat'/'conv_lon' fix (e.g. convective center) to the edge of the TCC around azimuth	km

conv_rad_avg	Average distance from the 'conv_lat'/'conv_lon' fix (e.g. convective center) to the edge of the TCC around azimuth. This is generally a good indication of the size of the TCC	km
cloud_height	Maximum derived cloud top height for the TCC. These are found by comparing the minimum T_b to an idealized tropical profile (McClatchey et al. 1972). We have low confidence in the accuracy of this parameter and it should be used with caution.	km
avg_cloud_height	Average derived cloud top height for the TCC from McClatchey et al. (1972). We have low confidence in the accuracy of this parameter and it should be used with caution.	km
direction	Direction that the TCC is moving toward in degrees off north (clockwise)	deg
wgt_direction	As in 'direction' but with the weighted centers	deg
speed	Forward speed of the TCC	kt
wgt_speed	Forward speed of the TCC using the weighted centers	kt
qc_flag	General quality control flag 0 = No flag set (everything looks fine) 1 = Abnormally large TCC (> 900 km) 2 = TCC crossed the Equator 3 = Both conditions 1 and 2 occurred	
qc_flag_interp	Quality control flag concerning interpolation issues 0 = No flag set 1 = TCC contains 3 consecutive interpolated points 2 = TCC has at least 50% of its points interpolated 3 = Both conditions 1 and 2 occur	
qc_flag_dev	Quality control flag for developing TCCs 0 = No flag set (everything good) 1 = Post-genesis TCC track at least 500 km from matching IBTrACS track 2 = TCC is at least 500 km from IBTrACS genesis point at genesis time 3 = Both conditions 1 and 2 occur	

5. Known Bugs and Issues (Status)

- A gap in geostationary coverage of the Indian Ocean produce artificially low global counts of TCCs prior to 1998 (not resolvable)
- The algorithm identifies about 75% of developing TCCs, missing 25%. This occurs for a number of reasons; in particular, "quick developers" (identified as a TC with little or no TCC phase) probably make up a large portion of the missing cases. Users should

consult the IBTrACS database as the developing “ground truth” data if they plan to use these data for TC genesis study. (No current plans to enhance the detection algorithm)

This section will be updated as new reports come in. Please send all discovered problems to Chris Hennon (chennon@unca.edu)

6. Support

This dataset and associated software is not actively supported, although there are plans to update the dataset with the latest year as input data become available. However, Attempts will also be made to correct reported problems with the data and subsequent updates will be issued. To report a problem or ask questions about the data, contact Chris Hennon (chennon@unca.edu).

7. Availability and Citation

TCC data are freely available. But *please* reference Hennon et al. (2011) in any publications or other work that use the dataset. Also, we would like to know if you are using the TCC data so that we can 1) justify further development, and 2) notify you of any bug fixes or updates. Please send an email to Chris Hennon (chennon@unca.edu) to let us know.

Data can be downloaded from:

<http://facstaff.unca.edu/chennon/research/tcc.shtml>

8. References

Hennon, C.C., C.N. Helms, K.R. Knapp, and A.R. Bowen, 2011: An Objective Algorithm for Detecting and Tracking Tropical Cloud Clusters: Implications for Tropical Cyclogenesis Prediction. *In Press, Journal of Atmospheric and Oceanic Technology*. Early release copy available from <http://facstaff.unca.edu/chennon/research/tcc.shtml>.

Hennon, C.C., and J.S. Hobgood, 2003: Forecasting tropical cyclogenesis over the Atlantic Basin using large-scale data. *Monthly Weather Review*, **131**, 2927-2940.

Knapp, K.R., and coauthors, 2011: Globally Gridded Satellite (GridSat) Observations for Climate Studies. *In Press, Bulletin of the American Meteorological Society*.

Knapp, K.R., M.C. Kruk, D.H. Levinson, H.J. Diamond, and C.J. Neumann, 2010: The International Best Track Archive for Climate Stewardship (IBTrACS): Unifying tropical cyclone best track data. *Bulletin of the American Meteorological Society*, **91**, 363-376.

Lee, C.S., 1989: Observational analysis of tropical cyclogenesis in the Western North Pacific. Part I: Structural evolution of cloud clusters. *Journal of the Atmospheric Sciences*, **46**, 2580-2598.