The Problem of Solar Contamination in GLM Data

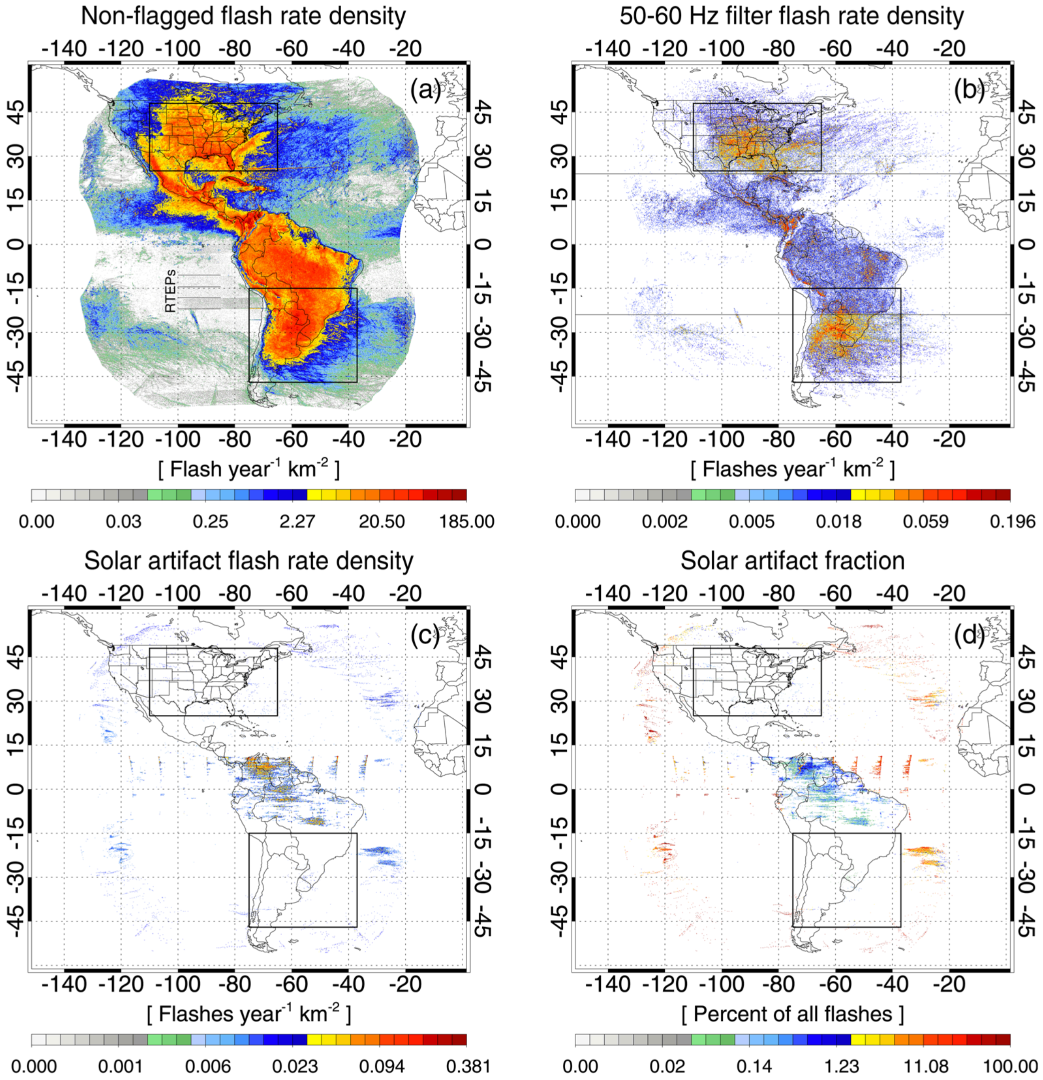
Michael Peterson

Background Reading

- The information in this presentation is summarized from DOI: https://doi.org/10.1117/1.JRS.14.032402

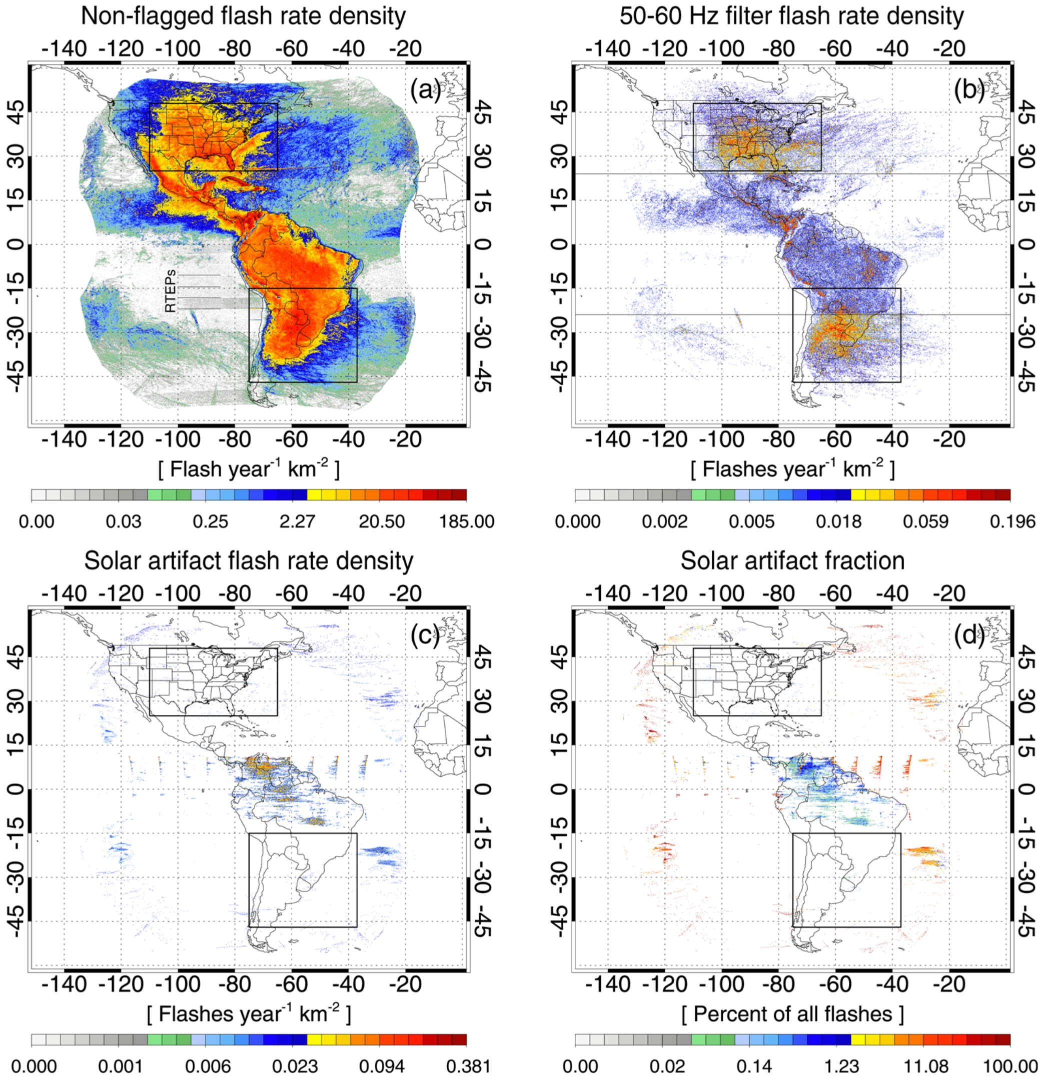
GLM Lightning Maps

- Panel (a) shows all lightning across the hemisphere seen by GLM
  - GLM detects ~1 million flashes per day
    - The edges of the figure show the limits of GLM's field of view
    - The GLM geographic extent is ~55 degrees in lat/lon from the satellite subpoint at 75.2 W
  - Most lightning occurs over land, with much lower flash activity over the oceans



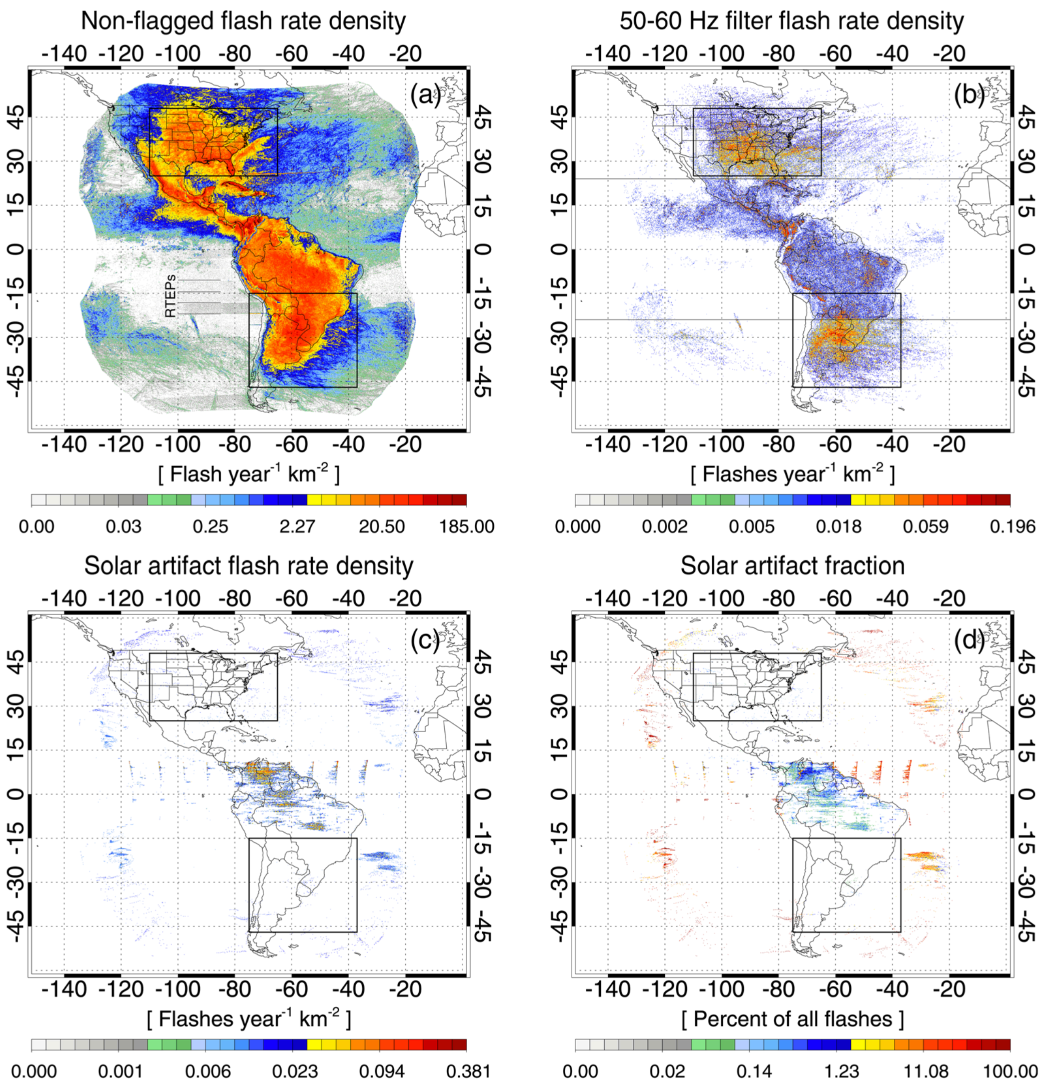
GLM Lightning Maps

- Panel (c) shows where solar contamination occurs
  - Not all solar contamination is included here – just the cases that are easy to find
  - The vertical banding and horizontal streaks near the equator are from sunlight reflecting off water and clouds directly below the satellite (usually at local noon)
  - The clusters at 30 N/S latitude near the edge of the field of view is from dawn / dusk glint off of seawater



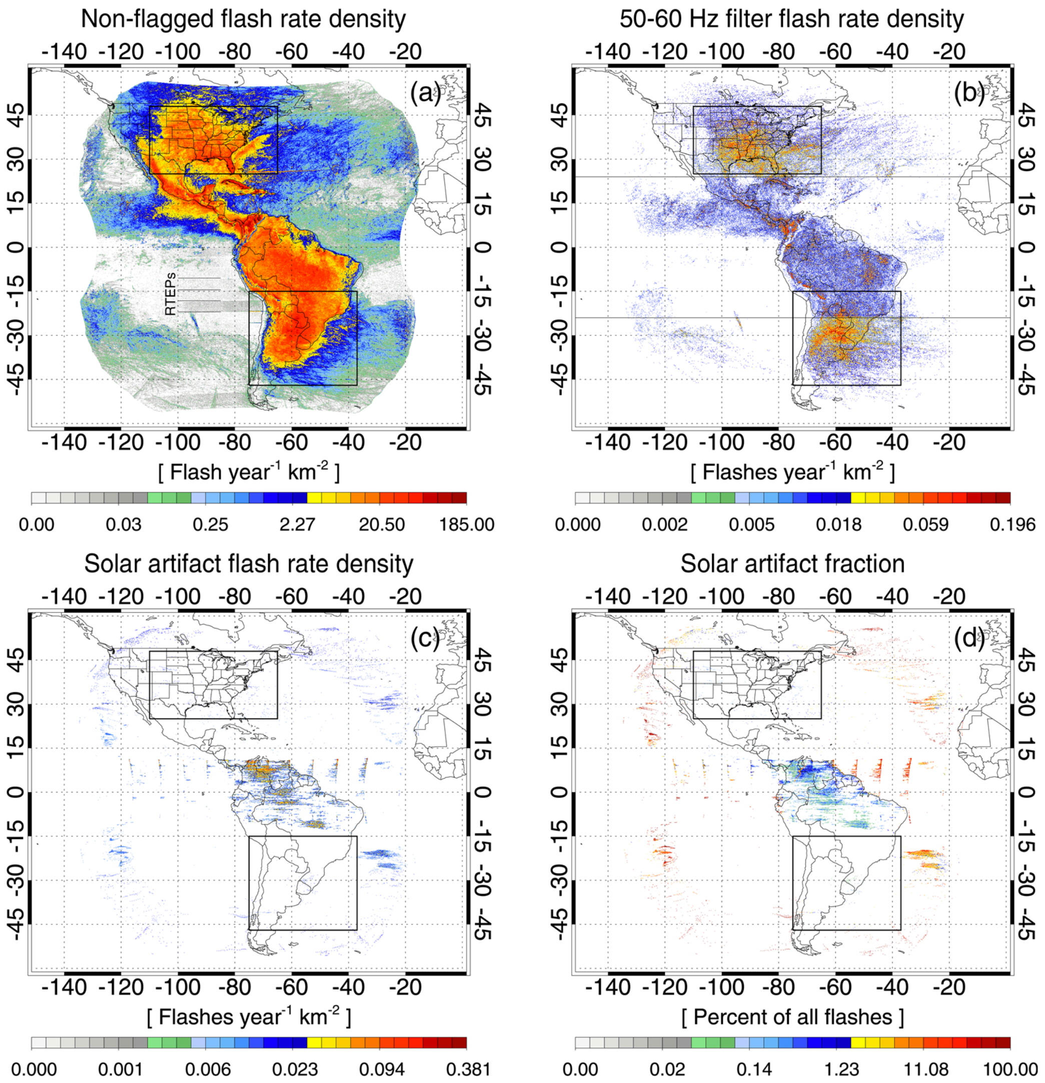
GLM Lightning Maps

- Solar contamination is very episodic. It only occurs when the Sun is in the right spot
    - Some days will have no false flashes from solar contamination. Other days will have dozens to hundreds within a ~1 hour period and none outside of this window



GLM Lightning Maps

- The problem with the artifact distribution in (c) is that a lot of the solar contamination occurs in regions with little lightning
- Thus, the oceanic contamination stands out in (a) and (d)

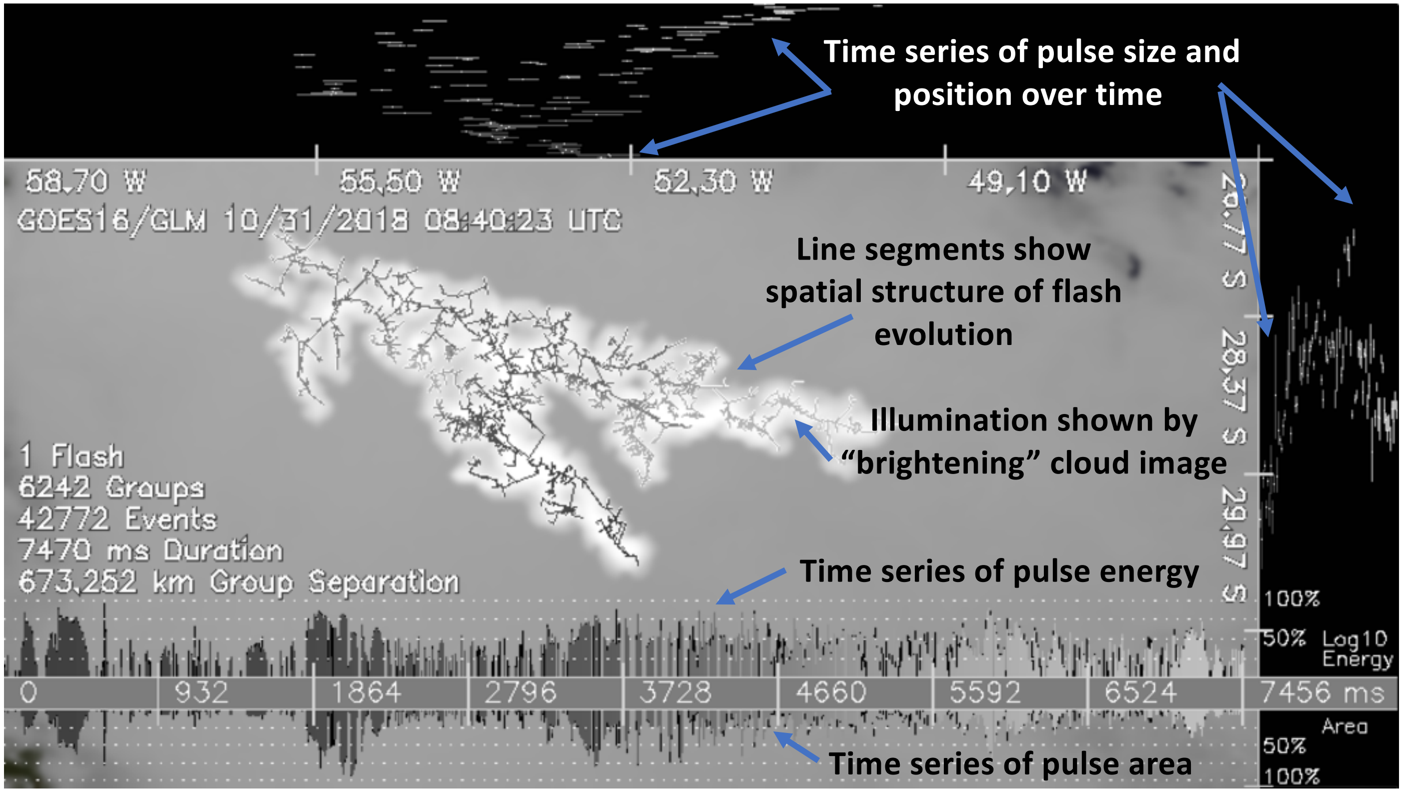


So, How do we Find Solar Artifacts?

- The instrument records each flash at 500 frames per second and can describe how each flash develops over time.
- This raw pixel/frame-level data is used to construct a number of parameters that are reported to describe each flash.
- Examining cases by hand shows that solar artifacts are easily identifiable on an individual basis (some examples in the following slides).
- The challenge is finding a good combination of these summary parameters that can reliably separate natural lightning from solar artifacts.
  - The good news is that we have a LOT of data to throw at this problem

What does Lightning Look Like?

- Example lightning "megaflash"
- Flashes have complicated structures that develop into one or more branches over time



1 Flash
6242 Groups
42772 Events
7470 ms Duration
673,252 km Group Separation

Time series of pulse size and position over time

Line segments show spatial structure of flash evolution

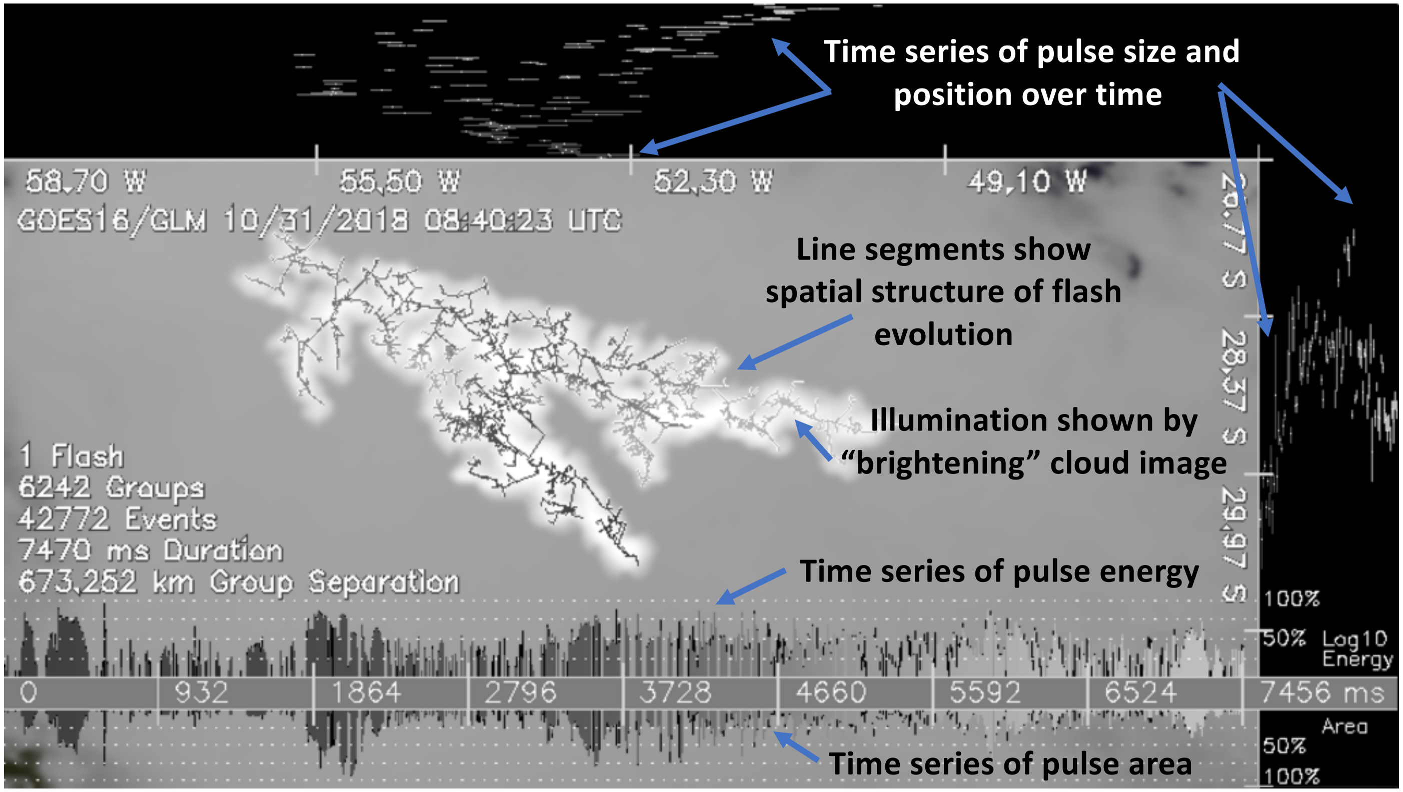
Illumination shown by "brightening" cloud image

Time series of pulse energy

Time series of pulse area

What does Lightning Look Like?

- The area/energy timeseries are also chaotic with many dim pulses interspersed with isolated very bright pulses



1 Flash
6242 Groups
42772 Events
7470 ms Duration
673,252 km Group Separation

Time series of pulse size and position over time

Line segments show spatial structure of flash evolution

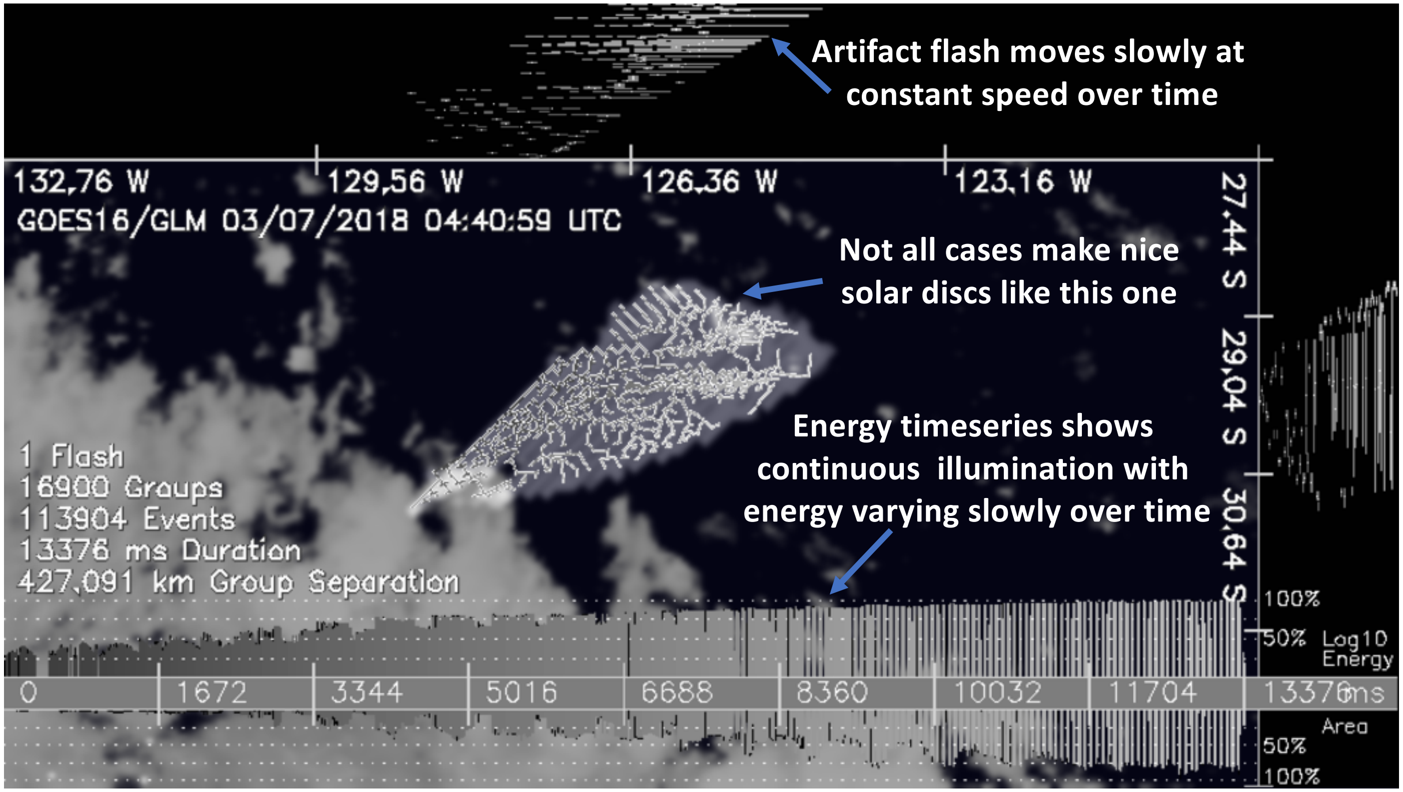
Illumination shown by “brightening” cloud image

Time series of pulse energy

Time series of pulse area

What does Solar Contamination Look Like?

- Example dawn/dusk glint
- Note that the gaps you see near the end of the timeseries are due to instrument overflow and are not physical



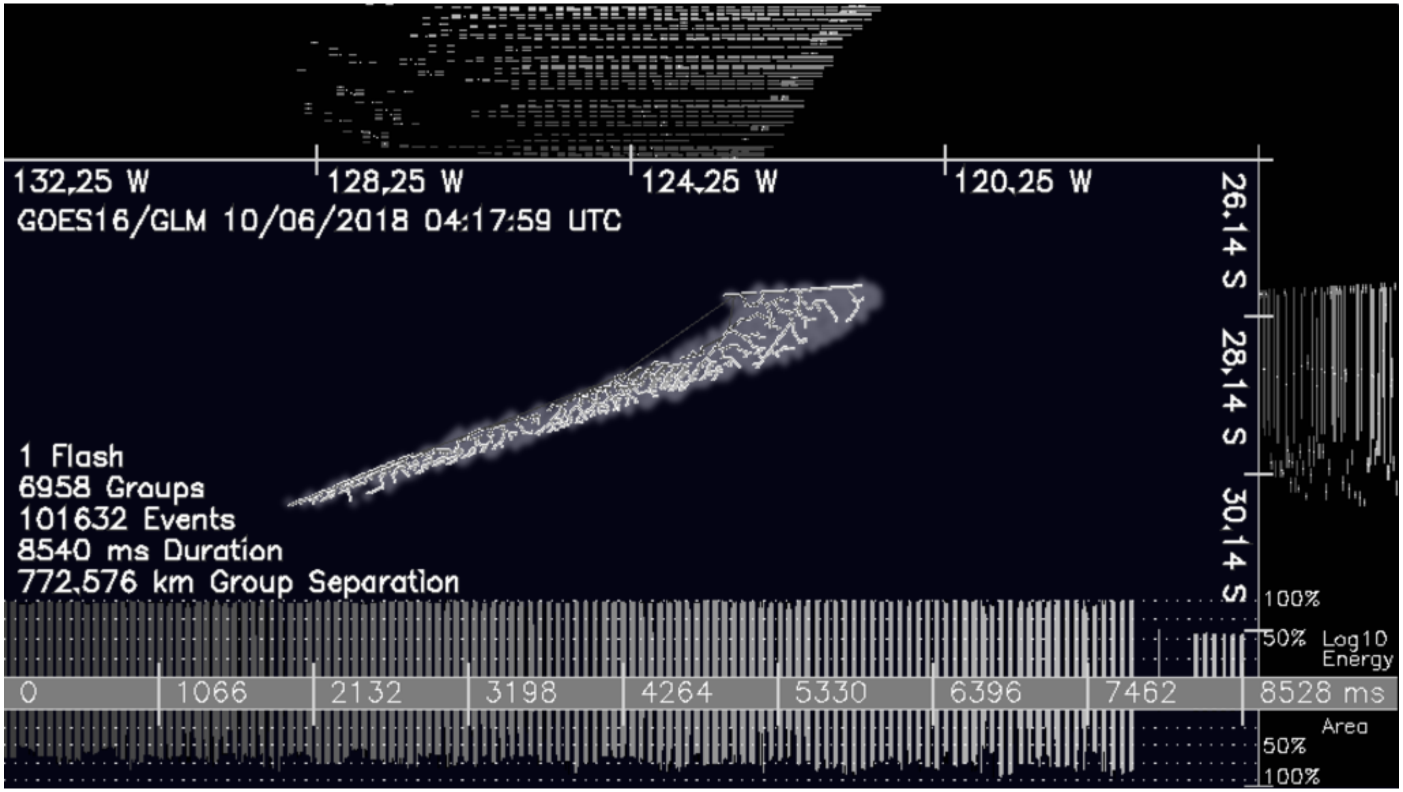
Artifact flash moves slowly at constant speed over time

Not all cases make nice solar discs like this one

Energy timeseries shows continuous illumination with energy varying slowly over time

What does Solar Contamination Look Like?

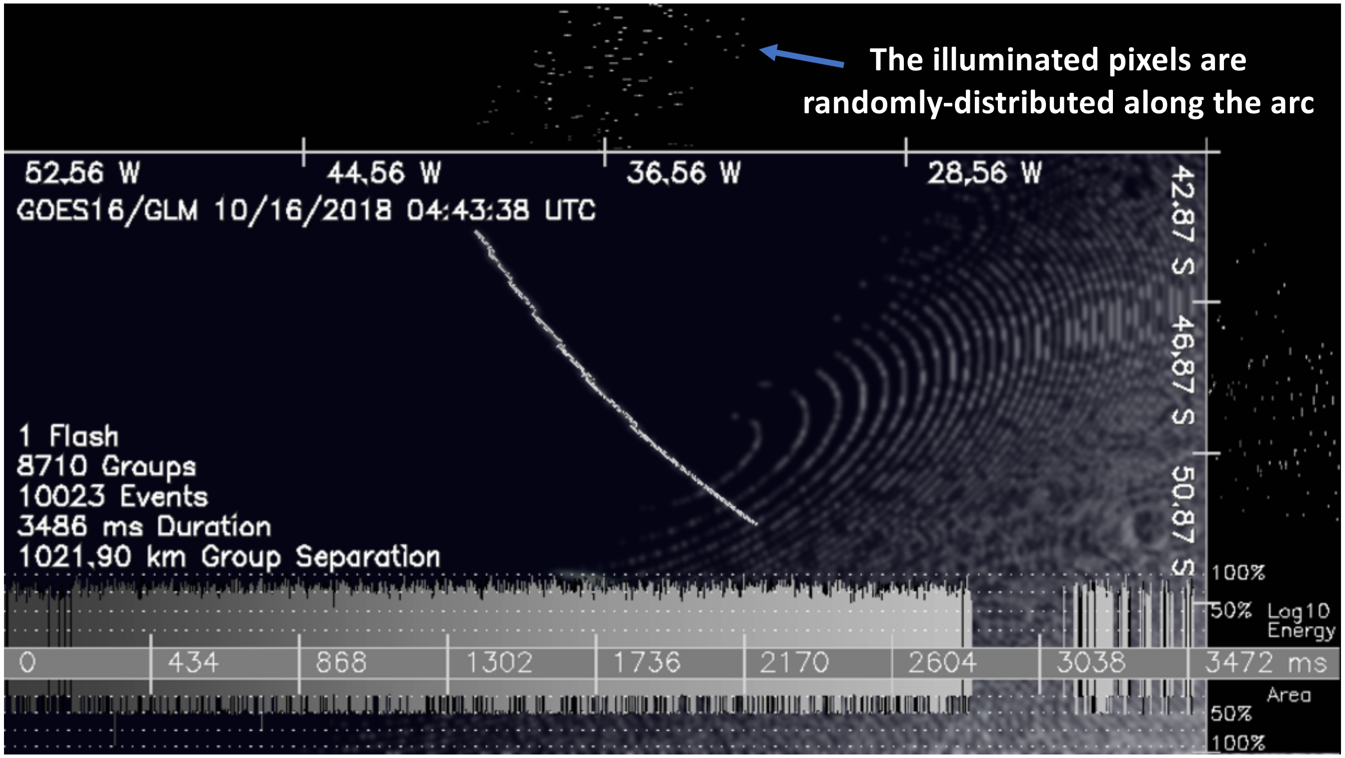
- Another example dawn/dusk glint
- This one is a half-disc case with a flat light curve and lots of instrument overflow gaps


GOES16/GLM 10/06/2018 04:17:59 UTC

1 Flash
6958 Groups
101632 Events
8540 ms Duration
772.576 km Group Separation

What does Solar Contamination Look Like?

- Example solar “arc” artifact
- Similar to previous example, but only a line of pixels is illuminated (probably refraction with sun near horizon)
- Same overall behavior – continuous illumination, but not always in same pixels



The illuminated pixels are randomly-distributed along the arc

Machine Learning Approach

- GLM data from the paper is organized into 3 types of files:
  - **glm_lightning_db_final.nc** -> list of natural lightning cases. Days are chosen where solar contamination is not observed, so should be free of solar artifacts
  - **glm_glint_db_final.nc** -> list of confirmed solar contamination cases. Identified by continuous illumination in a single pixel. This filter is very good at detecting all types of glint, but has an unreasonably-high missed event rate
    - NOTE: Files for daytime glint only and nighttime glint only also exist.
  - **glm_both_20180901.nc** -> a random sample of GLM data that contains both lightning and glint cases. Can be used for testing an independent sample of data
    - NOTE: 2 other days – 9/30/2018 and 10/20/2018 – also exist.
  - These are NetCDF files that can be loaded via the Python NetCDF4 module (should be available in either pip or conda)
- All 3 files contain lists of parameters describing each flash, which are explained on the following slides…

Contents of glm_lightning_db_final.nc

FLASH_ID	LONG	Array[29000001]	Unique identifier of lightning flash
FLASH_LCFA_CDATE	LONG	Array[29000001]	GLM data packet date stamp YYYYDOY (DOY is day of year)
FLASH_LCFA_TSTAMP	LONG	Array[29000001]	GLM data packet time stamp in UTC HHMMSS
FLASH_TIME_OFFSET_OF_FIRST_EVENT	FLOAT	Array[29000001]	Starting time of flash in seconds after LCFA_TSTAMP
FLASH_TIME_OFFSET_OF_LAST_EVENT	FLOAT	Array[29000001]	Ending time of flash in seconds after LCFA_TSTAMP
FLASH_LAT	FLOAT	Array[29000001]	Flash latitude
FLASH_LON	FLOAT	Array[29000001]	Flash longitude
FLASH_AREA	FLOAT	Array[29000001]	Flash illuminated area in m^2
FLASH_ENERGY	FLOAT	Array[29000001]	Flash total optical energy in J
FLASH_GROUP_COUNT	LONG	Array[29000001]	Number of optical pulses (termed “groups”) in the lash
FLASH_SERIES_COUNT	LONG	Array[29000001]	Number of distinct periods of illumination (termed “series”) in the flash
FLASH_EVENT_COUNT	LONG	Array[29000001]	Number of unique instrument illuminated piels (termed “events”) in the flash
FLASH_DURATION	FLOAT	Array[29000001]	Flash duration in seconds
FLASH_GROUP_MAX_SEPARATION	FLOAT	Array[29000001]	Maximum separation of groups in the flash in km
FLASH_GROUP_TOTAL_SEPARATION	FLOAT	Array[29000001]	Total separation of all line segments connecting groups in the flash in km
FLASH_EVENT_MAX_SEPARATION	FLOAT	Array[29000001]	Maximum separation of events in the flash in km
FLASH_1SIG_GROUP_COUNT	LONG	Array[29000001]	Number of bright groups in the flash at the mean+1*sigma (standard deviation) energy level
FLASH_2SIG_GROUP_COUNT	LONG	Array[29000001]	Number of bright groups in the flash at the 2-sigma energy level
FLASH_3SIG_GROUP_COUNT	LONG	Array[29000001]	Number of bright groups in the flash at the 3-sigma energy level
FLASH_1SIG_SERIES_COUNT	LONG	Array[29000001]	Number of series in the flash with 1-sigma bright groups
FLASH_2SIG_SERIES_COUNT	LONG	Array[29000001]	Number of series in the flash with 2-sigma bright groups
FLASH_3SIG_SERIES_COUNT	LONG	Array[29000001]	Number of series in the flash with 3-sigma bright groups
FLASH_EVENT_MAX_ENERGY	FLOAT	Array[29000001]	Optical energy of brightest event (illuminated pixel) in the flash in J
FLASH_EVENT_MIN_ENERGY	FLOAT	Array[29000001]	Optical energy of dimmest event (illuminated pixel) in the flash in J
FLASH_GROUP_MAX_ENERGY	FLOAT	Array[29000001]	Optical energy of brightest group (i.e., pulse) in the flash in J
FLASH_GROUP_MEAN_ENERGY	FLOAT	Array[29000001]	Mean optical energy of all groups in the flash in J
FLASH_GROUP_MIN_ENERGY	FLOAT	Array[29000001]	Optical energy of dimmest group in the flash in J

Contents of glm_dayglint_db_final.nc

NOTE: the energy / area parameters have different units in this file compared to the other two. Be careful!

FLASH_ID	LONG	Array[61028]	Unique identifier of lightning flash
FLASH_LCFA_CDATE	LONG	Array[61028]	GLM data packet date stamp YYYYDOY (DOY is day of year)
FLASH_LCFA_TSTAMP	LONG	Array[61028]	GLM data packet time stamp in UTC HHMMSS
FLASH_TIME_OFFSET_OF_FIRST_EVENT	FLOAT	Array[61028]	Starting time of flash in seconds after LCFA_TSTAMP
FLASH_TIME_OFFSET_OF_LAST_EVENT	FLOAT	Array[61028]	Ending time of flash in seconds after LCFA_TSTAMP
FLASH_LAT	FLOAT	Array[61028]	Flash latitude
FLASH_LON	FLOAT	Array[61028]	Flash longitude
FLASH_AREA	FLOAT	Array[61028]	Flash illuminated area in **km^2**
FLASH_ENERGY	FLOAT	Array[61028]	Flash total optical energy in **fJ**
FLASH_GROUP_COUNT	LONG	Array[61028]	Number of optical pulses (termed "groups") in the lash
FLASH_SERIES_COUNT	LONG	Array[61028]	Number of distinct periods of illumination (termed "series") in the flash
FLASH_EVENT_COUNT	LONG	Array[61028]	Number of unique instrument illuminated piels (termed "events") in the flash
FLASH_DURATION	FLOAT	Array[61028]	Flash duration in seconds
FLASH_GROUP_MAX_SEPARATION	FLOAT	Array[61028]	Maximum separation of groups in the flash in km
FLASH_GROUP_TOTAL_SEPARATION	FLOAT	Array[61028]	Total separation of all line segments connecting groups in the flash in km
FLASH_EVENT_MAX_SEPARATION	FLOAT	Array[61028]	Maximum separation of events in the flash in km
FLASH_1SIG_GROUP_COUNT	LONG	Array[61028]	Number of bright groups in the flash at the mean+1*sigma (standard deviation) energy level
FLASH_2SIG_GROUP_COUNT	LONG	Array[61028]	Number of bright groups in the flash at the 2-sigma energy level
FLASH_3SIG_GROUP_COUNT	LONG	Array[61028]	Number of bright groups in the flash at the 3-sigma energy level
FLASH_1SIG_SERIES_COUNT	LONG	Array[61028]	Number of series in the flash with 1-sigma bright groups
FLASH_2SIG_SERIES_COUNT	LONG	Array[61028]	Number of series in the flash with 2-sigma bright groups
FLASH_3SIG_SERIES_COUNT	LONG	Array[61028]	Number of series in the flash with 3-sigma bright groups
FLASH_EVENT_MAX_ENERGY	FLOAT	Array[61028]	Optical energy of brightest event (illuminated pixel) in the flash in **fJ**
FLASH_EVENT_MIN_ENERGY	FLOAT	Array[61028]	Optical energy of dimmest event (illuminated pixel) in the flash in **fJ**
FLASH_GROUP_MAX_ENERGY	FLOAT	Array[61028]	Optical energy of brightest group (i.e., pulse) in the flash in **fJ**
FLASH_GROUP_MEAN_ENERGY	FLOAT	Array[61028]	Mean optical energy of all groups in the flash in **fJ**
FLASH_GROUP_MIN_ENERGY	FLOAT	Array[61028]	Optical energy of dimmest group in the flash in **fJ**

Contents of glm_both_20180901.nc

FLASH_ID	LONG	Array[29999999]	Unique identifier of lightning flash
FLASH_LCFA_CDATE	LONG	Array[29999999]	GLM data packet date stamp YYYYDOY (DOY is day of year)
FLASH_LCFA_TSTAMP	LONG	Array[29999999]	GLM data packet time stamp in UTC HHMMSS
FLASH_TIME_OFFSET_OF_FIRST_EVENT	FLOAT	Array[29999999]	Starting time of flash in seconds after LCFA_TSTAMP
FLASH_TIME_OFFSET_OF_LAST_EVENT	FLOAT	Array[29999999]	Ending time of flash in seconds after LCFA_TSTAMP
FLASH_LAT	FLOAT	Array[29999999]	Flash latitude
FLASH_LON	FLOAT	Array[29999999]	Flash longitude
FLASH_AREA	FLOAT	Array[29999999]	Flash illuminated area in m^2
FLASH_ENERGY	FLOAT	Array[29999999]	Flash total optical energy in J
FLASH_GROUP_COUNT	LONG	Array[29999999]	Number of optical pulses (termed “groups”) in the lash
FLASH_SERIES_COUNT	LONG	Array[29999999]	Number of distinct periods of illumination (termed “series”) in the flash
FLASH_EVENT_COUNT	LONG	Array[29999999]	Number of unique instrument illuminated piels (termed “events”) in the flash
FLASH_DURATION	FLOAT	Array[29999999]	Flash duration in seconds
FLASH_GROUP_MAX_SEPARATION	FLOAT	Array[29999999]	Maximum separation of groups in the flash in km
FLASH_GROUP_TOTAL_SEPARATION	FLOAT	Array[29999999]	Total separation of all line segments connecting groups in the flash in km
FLASH_EVENT_MAX_SEPARATION	FLOAT	Array[29999999]	Maximum separation of events in the flash in km
FLASH_1SIG_GROUP_COUNT	LONG	Array[29999999]	Number of bright groups in the flash at the mean+1*sigma (standard deviation) energy level
FLASH_2SIG_GROUP_COUNT	LONG	Array[29999999]	Number of bright groups in the flash at the 2-sigma energy level
FLASH_3SIG_GROUP_COUNT	LONG	Array[29999999]	Number of bright groups in the flash at the 3-sigma energy level
FLASH_1SIG_SERIES_COUNT	LONG	Array[29999999]	Number of series in the flash with 1-sigma bright groups
FLASH_2SIG_SERIES_COUNT	LONG	Array[29999999]	Number of series in the flash with 2-sigma bright groups
FLASH_3SIG_SERIES_COUNT	LONG	Array[29999999]	Number of series in the flash with 3-sigma bright groups
FLASH_EVENT_MAX_ENERGY	FLOAT	Array[29999999]	Optical energy of brightest event (illuminated pixel) in the flash in J
FLASH_EVENT_MIN_ENERGY	FLOAT	Array[29999999]	Optical energy of dimmest event (illuminated pixel) in the flash in J
FLASH_GROUP_MAX_ENERGY	FLOAT	Array[29999999]	Optical energy of brightest group (i.e., pulse) in the flash in J
FLASH_GROUP_MEAN_ENERGY	FLOAT	Array[29999999]	Mean optical energy of all groups in the flash in J
FLASH_GROUP_MIN_ENERGY	FLOAT	Array[29999999]	Optical energy of dimmest group in the flash in J

Machine Learning Approach

- The ML approach I'd recommend is as follows:
  - Load in **glm_lightning_db_final.nc** and **glm_dayglint_db_final.nc**.
  - Make new "glint_flag" field for both datasets. Set 0 for the lightning data and 1 for the glint data.
  - Concat arrays in both datasets. Divide into testing / training data.
  - Run ML fit of choice to predict glint_flag from other fields. Generate performance statistics for large sample. Refine as needed / as makes sense.
    - For example, I wouldn't use all fields in the model. The three I'd definitely remove are FLASH_ID, FLASH_TIME_OFFSET_OF_FIRST_EVENT, and FLASH_TIME_OFFSET_OF_LAST_EVENT because they are just rolling counters.
    - If you want to use FLASH_LCFA_CDATE, first strip off the year first.
    - These refinements provide ample room for creativity. There isn't a clear answer of what will work (at least not to me).
  - Run the data in **glm_both_20180901.nc** through the ML model. We don't know which flashes are which here – but mapping where the solar artifacts found by the model occur and plotting their diurnal cycle will give us a good sense of how well it's doing.