Alternative Earth Science Datasets For Identifying Patterns and Events

Kaylin Bugbee¹, Robert Griffin¹, Brian Freitag¹, Jeffrey Miller¹, Rahul Ramachandran², and Jia Zhang³

(1) University of Alabama in Huntsville  (2) NASA MSFC (3) Carnegie Mellon Universityv
Earth Observation Big Data

• Earth observation data volumes are growing exponentially
• NOAA collects about 7 terabytes of data per day\(^1\)
  • Adds to existing 25 PB archive
  • Upcoming missions will generate another 5 TB per day
• NASA’s Earth observation data is expected to grow to 131 TB of data per day by 2022\(^2\)
  • NISAR and other large data volume missions\(^3\)
• Other agencies like ESA expect data volumes to continue to grow\(^4\)
• How do we effectively explore and search through these large amounts of data?

Over the next five years, the daily ingest of data into the EOSDIS archive is expected to grow significantly, to more than 131 terabytes (TB) of forward processing. NASA EOSDIS image.
Alternative Data

• Data which are extracted or generated from non-traditional sources
  • Social media data
  • Point of sale transactions
  • Product reviews
  • Logistics

• Idea originates in investment world
  • Include alternative data sources in investment decision making process

• Earth observation data is a growing alternative data source for investing
  • DMSP and VIIRS nightlight data

Image Credit: NASA
Alternative Data for Earth Science

• Are there alternative data sources in the Earth sciences that can be used in a similar manner?
  • Yes
    • Social media
    • Flight reports for airborne field campaigns
    • Agricultural reports
    • Weather forecasts
  • Alternative Earth science data can be analyzed to
    • Identify interesting events or trends
    • Look for spatial, temporal or climatological patterns
    • Assist in efficiently identifying events or use cases in large volume datasets
Area Forecast Discussions

- Weather Forecast Offices
  - National Weather Service operates 122 WFOs
  - Responsible for issuing forecasts and severe weather warnings

- Area Forecast Discussions
  - Written every 6 hours
  - Covers most significant weather issues facing a WFO including a forecast, summary of outlooks, watches, warnings, etc

- How do we identify important information within these reports?

Image Credit: Gus Polly
https://commons.wikimedia.org/wiki/File:NWS_Weather_Forecast_Offices.svg
Can use the American Meteorological Society Glossary of Meteorology to identify important information within the AFDs

- Over 12,000 important meteorological terms
- Curated and domain specific

Includes broad terms
- Hurricane, flooding, and snow

More specific meteorological terms
- Vorticity, gap wind, etc

http://glossary.ametsoc.org/wiki/Main_Page
How Did We Create an EO Alternative Dataset?

- Created an alternative Earth observation dataset using the following method:
- Used the Iowa State University, Iowa Environmental Mesonet website to obtain AFDs
  - NWS only stores last 50 version of AFDs
  - Scraped each page for text
- Used the AMS Glossary of Meteorology to extract terms from the AFDs
- Followed a heuristic, rule-based technique to extract terms
- Data includes word count, time of forecast, location
Exploratory Use Case: Hard Freeze

• Subsetted list of glossary terms to 20 for an exploratory analysis

• This exploratory use case will focus on the term ‘hard freeze’
  
  “A freeze in which seasonal vegetation is destroyed, the ground surface is frozen solid underfoot, and heavy ice is formed on small water surfaces such as puddles and water containers” \(^6\)

• Identifying hard freeze events is important to agricultural community
  
  • Need to understand past events
  • Early detection of these events as they occur

Created by IconTrack from Noun Project
Exploratory Use Case: Methods

• Approach: Look for temporal and geospatial statistical trends in AFD extraction data
• Temporal analysis to identify interesting events

1. Explore yearly counts
2. Explore monthly counts for a year of interest
3. Identify days of interest within a given month
Exploratory Use Case: Event Identification

- Yearly analysis shows a peak in mentions in 2010
- Subsetting down to years 2009 – 2011
- Peak in January 2010
- Coincides with expected increase in usage in winter months

Hard Freeze Mentions 2009 - 2011
Exploratory Use Case: Event Identification

December 2010

Record cold temperatures in Florida including Jacksonville, Tallahassee

Baton Rouge

Other parts of the south
• Approach: Look for geospatial statistical trends in AFD extraction data yearly and over a decade
  - **Spatial analysis to identify interesting events**
  - Based on the idea that observations are non-independent, nearby units in some way are associated
  - Can identify spatially significant patterns in extractions from year to year and over a decade
• **Hot spot analysis**
• To be a statistically significant hot spot, a feature will have a high value and be surrounded by other features with high values as well.
• Results indicate positive Z-scores are inversely related to winter temperature trends.
Exploratory Use Case: Combined Methods

- Can leverage known spatial region of interest in combination with temporal analysis
- Florida and south Florida peak in mid January coinciding with coldest day
Lessons Learned

• **Large volume of data**
  - Challenging to scrape a large number of web pages – easy to miss pages
  - Difficult to check for quality
  - Broad exploratory check did not always find data gaps

• **Ambiguities of human communication**
  - Writing styles and human perception affect analysis results
  - Assumptions of relevance are made for each WFO
  - Thresholds of concern for a WFO
    - Hard freeze example
    - Most of the U.S. experiences hard freeze conditions
    - Offices which are concerned about direct impacts of a hard freeze use the term more frequently
  - These uncertainties make using the AFD data impractical for certain scientific applications
  - Still helpful for identifying events and trends
Future Work and Conclusions

Possible Future Work
• Explore other terms for interesting events and trends
• Monitor for events in AFDs in near real time
• Investigate automated techniques for identifying events

Conclusions
• As data volumes grow, alternative Earth science datasets offer one solution to help users more efficiently search for relevant data
Questions?

Contact:
Kaylin.m.Bugbee@nasa.gov
References

2. https://earthdata.nasa.gov/about/eosdis-cloud-evolution
5. https://mesonet.agron.iastate.edu/