Alternative Earth Science Datasets For Identifying Patterns and Events

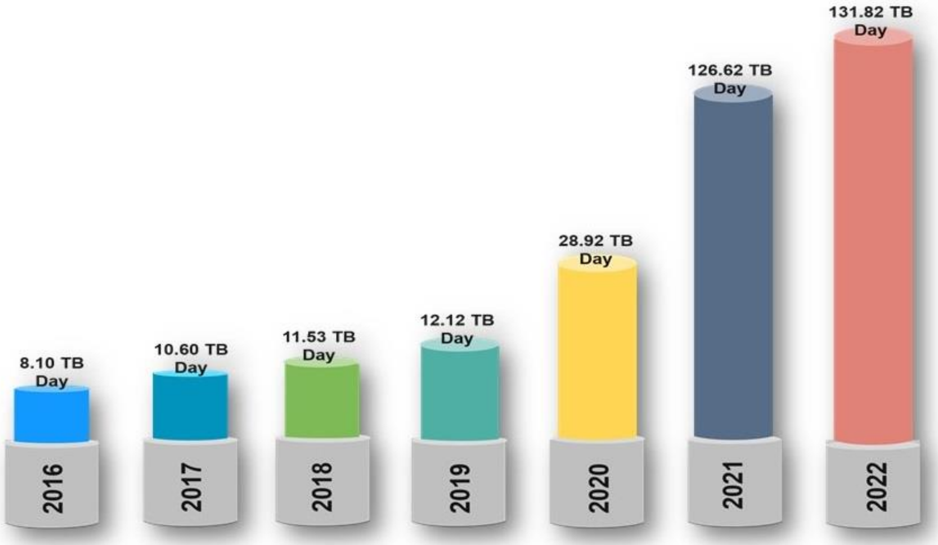
Kaylin Bugbee[1], Robert Griffin[1], Brian Freitag[1], Jeffrey Miller[1], Rahul Ramachandran[2], and Jia Zhang[3]

(1) University of Alabama in Huntsville (2) NASA MSFC (3) Carnegie Mellon Universityv

NASA Carnegie Mellon University THE UNIVERSITY OF ALABAMA IN HUNTSVILLE IMPACT Interagency Implementation and Advanced Concepts Team

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.*

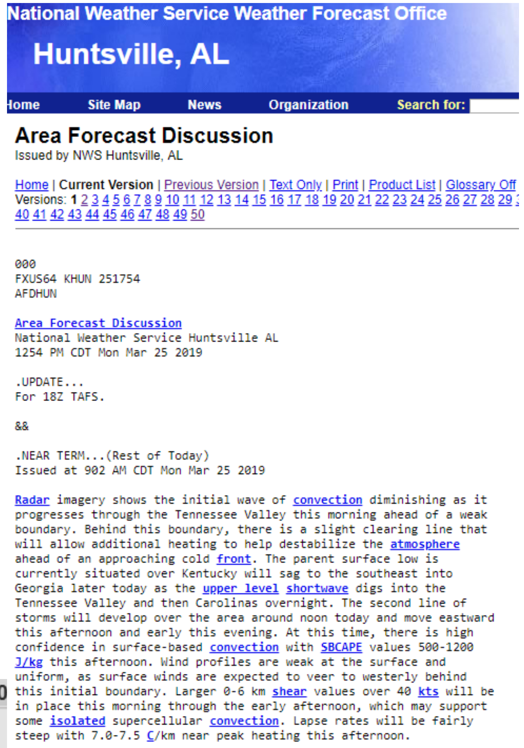
Alternative Data

- 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



National Weather Service Weather Forecast Office

Huntsville, AL

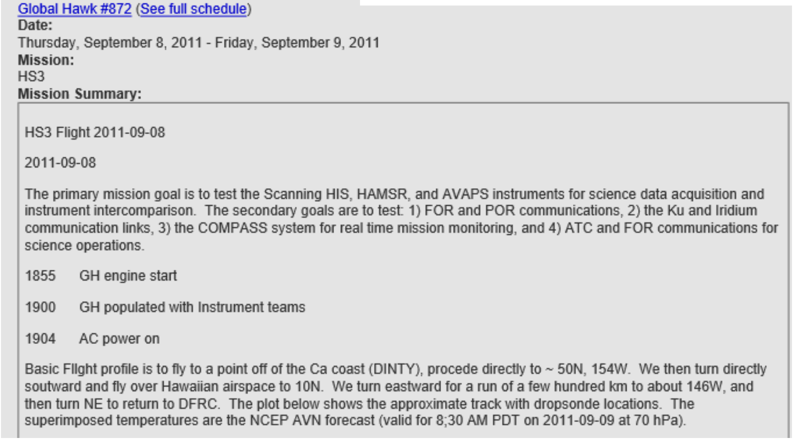
Area Forecast Discussion

Issued by NWS Huntsville, AL

Home | **Current Version** | Previous Version | Text Only | Print | Product List | Glossary Off
Versions: **1** 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 3
40 41 42 43 44 45 46 47 48 49 50

```
000
FXUS64 KHUN 251754
AFDHUN

Area Forecast Discussion
National Weather Service Huntsville AL
1254 PM CDT Mon Mar 25 2019

.UPDATE...
For 18Z TAFS.

&&

.NEAR TERM...(Rest of Today)
Issued at 902 AM CDT Mon Mar 25 2019

Radar imagery shows the initial wave of convection diminishing as it
progresses through the Tennessee Valley this morning ahead of a weak
boundary. Behind this boundary, there is a slight clearing line that
will allow additional heating to help destabilize the atmosphere
ahead of an approaching cold front. The parent surface low is
currently situated over Kentucky will sag to the southeast into
Georgia later today as the upper level shortwave digs into the
Tennessee Valley and then Carolinas overnight. The second line of
storms will develop over the area around noon today and move eastward
this afternoon and early this evening. At this time, there is high
confidence in surface-based convection with SBCAPE values 500-1200
J/kg this afternoon. Wind profiles are weak at the surface and
uniform, as surface winds are expected to veer to westerly behind
this initial boundary. Larger 0-6 km shear values over 40 kts will be
in place this morning through the early afternoon, which may support
some isolated supercellular convection. Lapse rates will be fairly
steep with 7.0-7.5 C/km near peak heating this afternoon.
```

HS3 - Global Hawk #872 09/08/11 - 0



Aircraft:
Global Hawk #872 (See full schedule)
Date:
Thursday, September 8, 2011 - Friday, September 9, 2011
Mission:
HS3
Mission Summary:

HS3 Flight 2011-09-08

2011-09-08

The primary mission goal is to test the Scanning HIS, HAMSR, and AVAPS instruments for science data acquisition and instrument intercomparison. The secondary goals are to test: 1) FOR and POR communications, 2) the Ku and Iridium communication links, 3) the COMPASS system for real time mission monitoring, and 4) ATC and FOR communications for science operations.

1855 GH engine start

1900 GH populated with Instrument teams

1904 AC power on

Basic FIlght profile is to fly to a point off of the Ca coast (DINTY), procede directly to ~ 50N, 154W. We then turn directly soutward and fly over Hawaiian airspace to 10N. We turn eastward for a run of a few hundred km to about 146W, and then turn NE to return to DFRC. The plot below shows the approximate track with dropsonde locations. The superimposed temperatures are the NCEP AVN forecast (valid for 8;30 AM PDT on 2011-09-09 at 70 hPa).

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?**

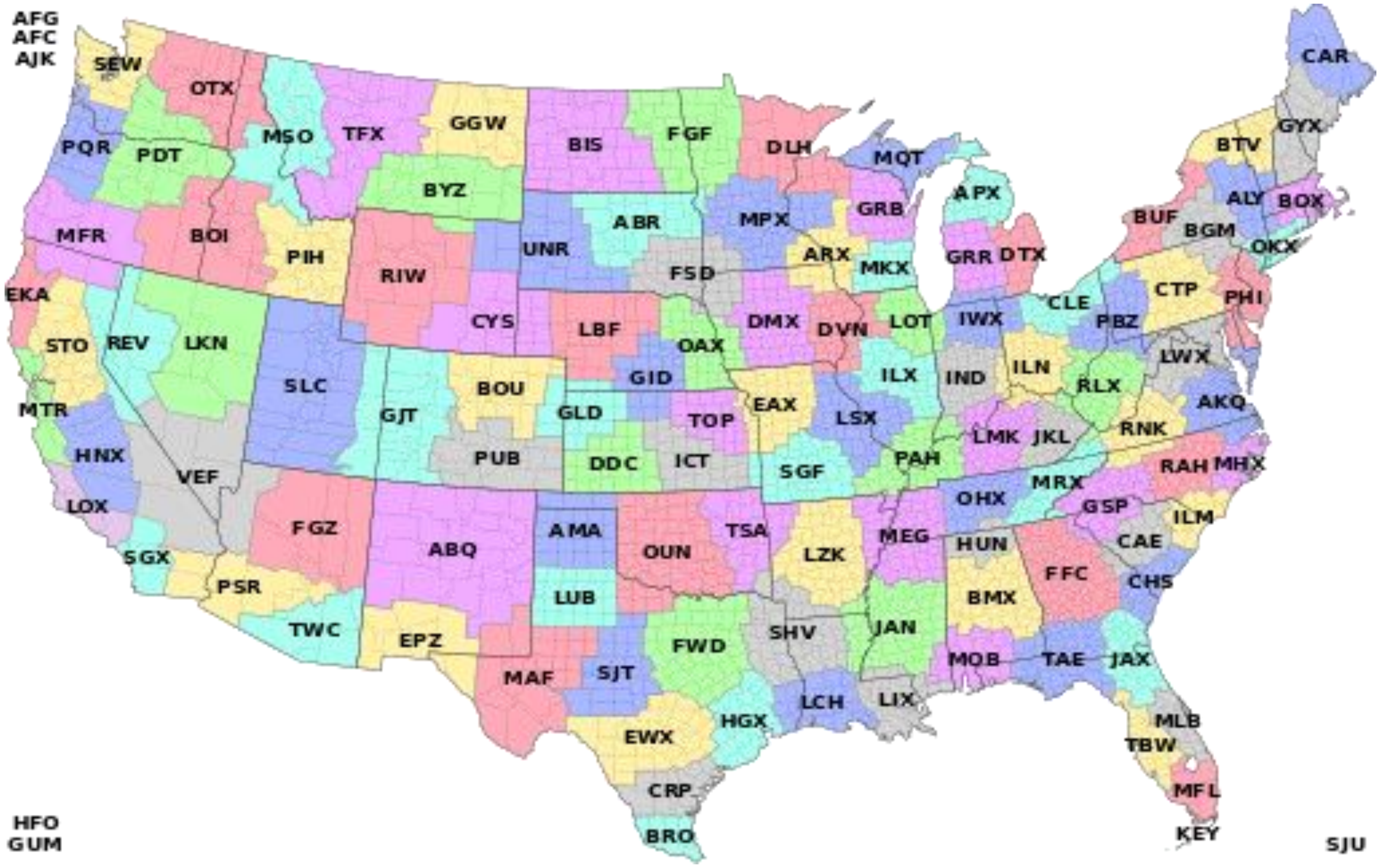
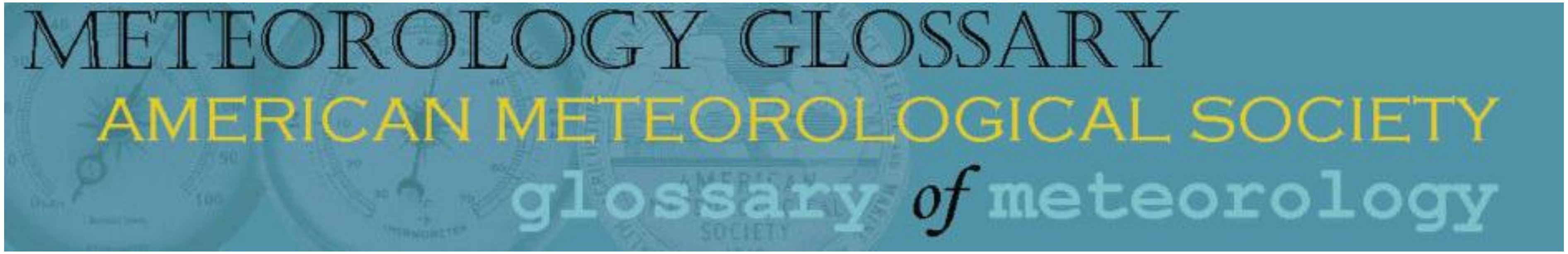


Image Credit: Gus Polly
https://commons.wikimedia.org/wiki/File:NWS_Weather_Forecast_Offices.svg

AMS Glossary of Meteorology

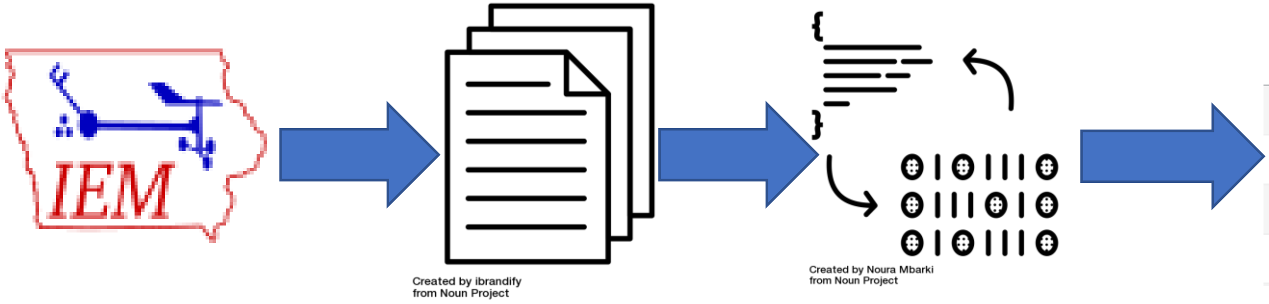
- Can use the American Meteorological Society Glossary of Meteorology to identify important information within the AFDs
  - Over 12,000 important meteorological terms
  - Curated and domain specific
- Includes broad terms
  - Hurricane, flooding, and snow
- More specific meteorological terms
  - Vorticity, gap wind, etc



http://glossary.ametsoc.org/wiki/Main_Page

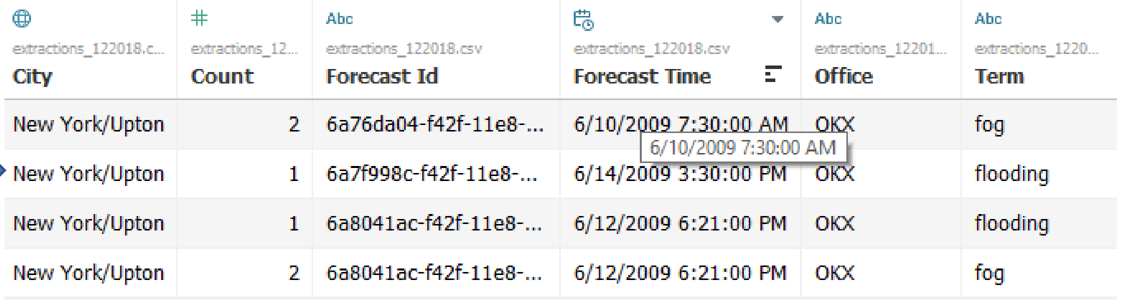
How Did We Create an EO Alternative Dataset?

- **Created an alternative Earth observation dataset using the following method:**
- Used the Iowa State University, Iowa Environmental Mesonet website[5]
- 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



Created by ibrandify from Noun Project

Created by Noura Mbarki from Noun Project

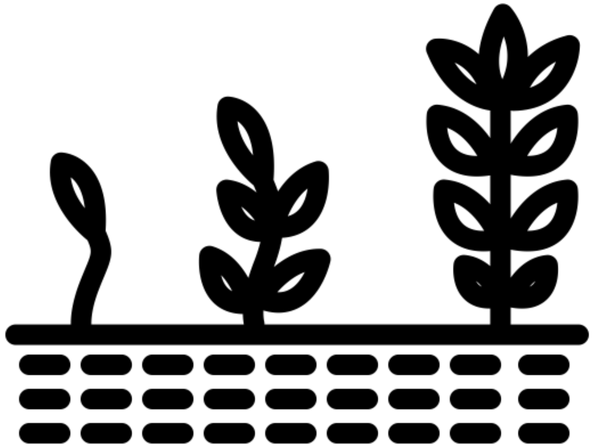


extractions_122018.c... City	extractions_12... Count	extractions_122018.csv Forecast Id	extractions_122018.csv Forecast Time	extractions_12201... Office	extractions_1220... Term
New York/Upton	2	6a76da04-f42f-11e8-...	6/10/2009 7:30:00 AM	OKX	fog
New York/Upton	1	6a7f998c-f42f-11e8-...	6/14/2009 3:30:00 PM	OKX	flooding
New York/Upton	1	6a8041ac-f42f-11e8-...	6/12/2009 6:21:00 PM	OKX	flooding
New York/Upton	2	6a8041ac-f42f-11e8-...	6/12/2009 6:21:00 PM	OKX	fog

6/10/2009 7:30:00 AM

Exploratory Use Case: Hard Freeze

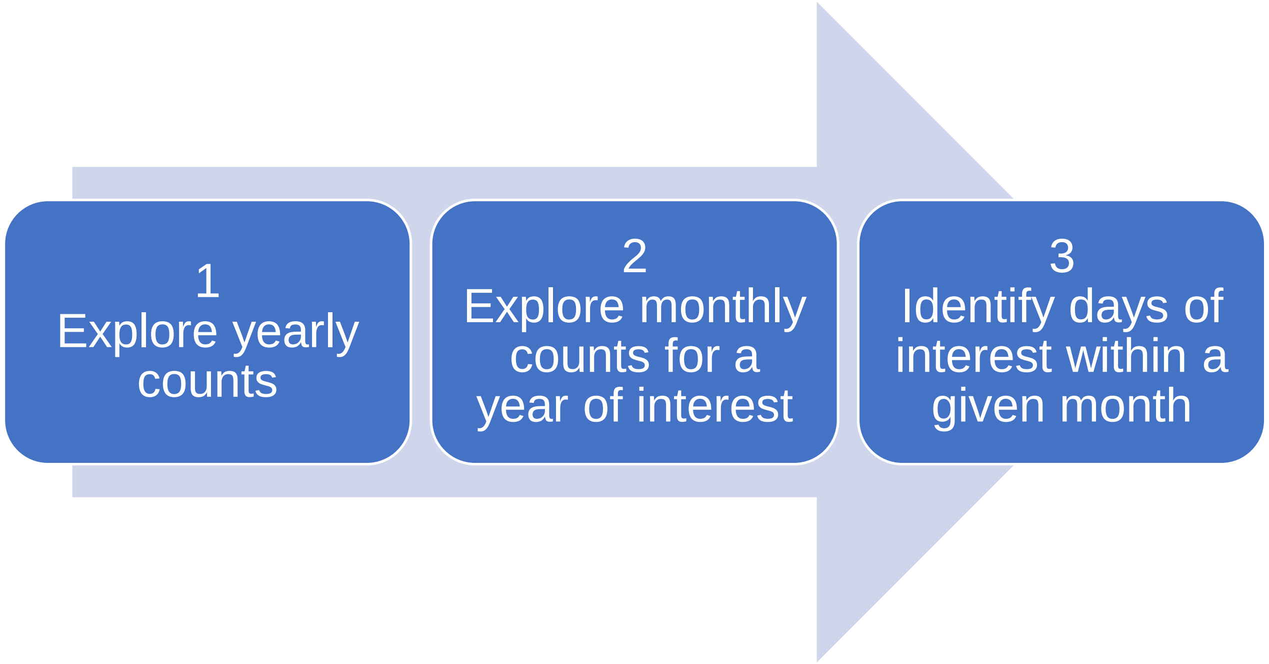
- Subsetted list of glossary terms to 20 for an exploratory analysis
- **This exploratory use case will focus on the term ‘hard freeze’**
- “A freeze in which seasonal vegetation is destroyed, the ground surface is frozen solid underfoot, and heavy ice is formed on small water surfaces such as puddles 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**



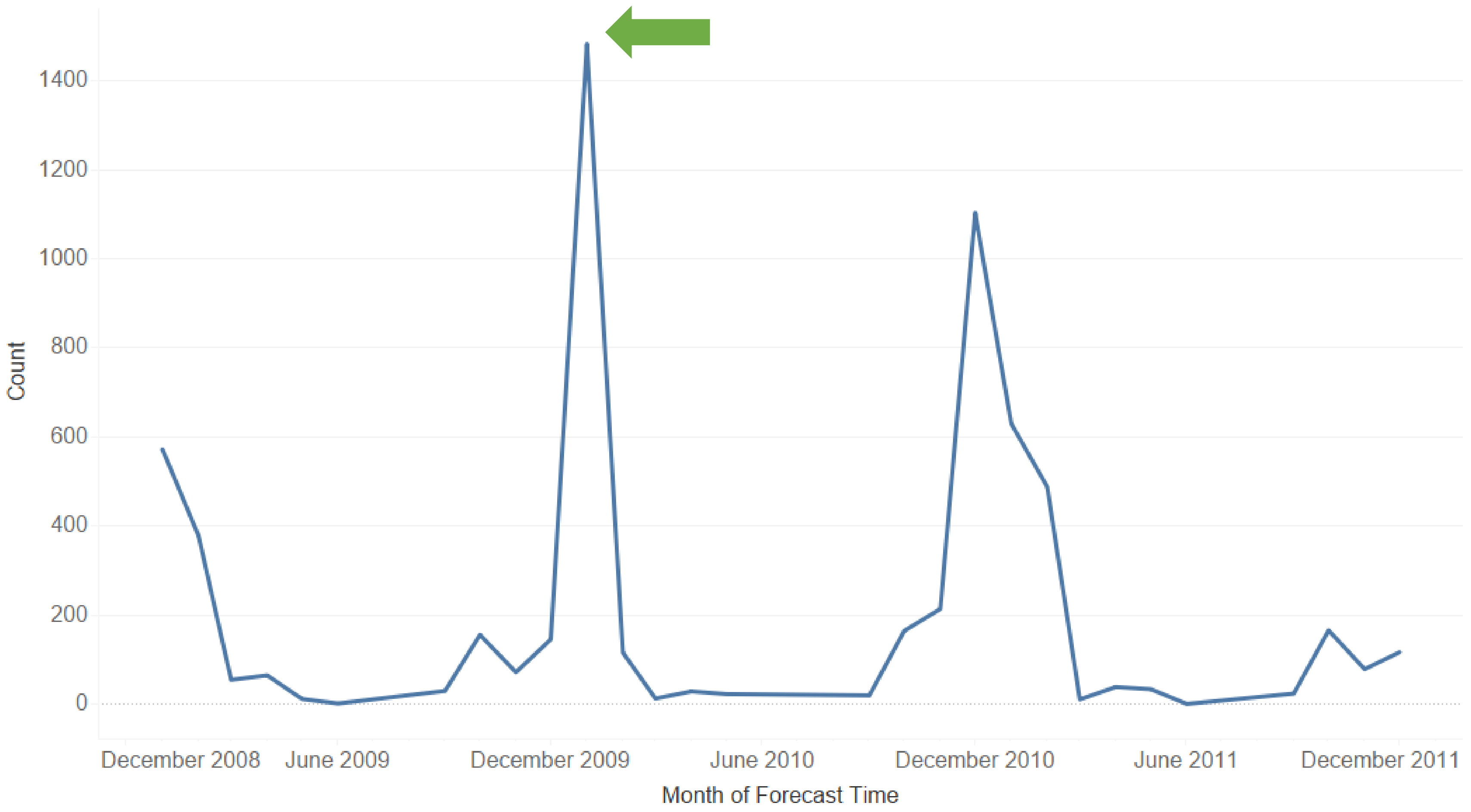
1
Explore yearly counts

2
Explore monthly counts for a year of interest

3
Identify days of interest within a given month

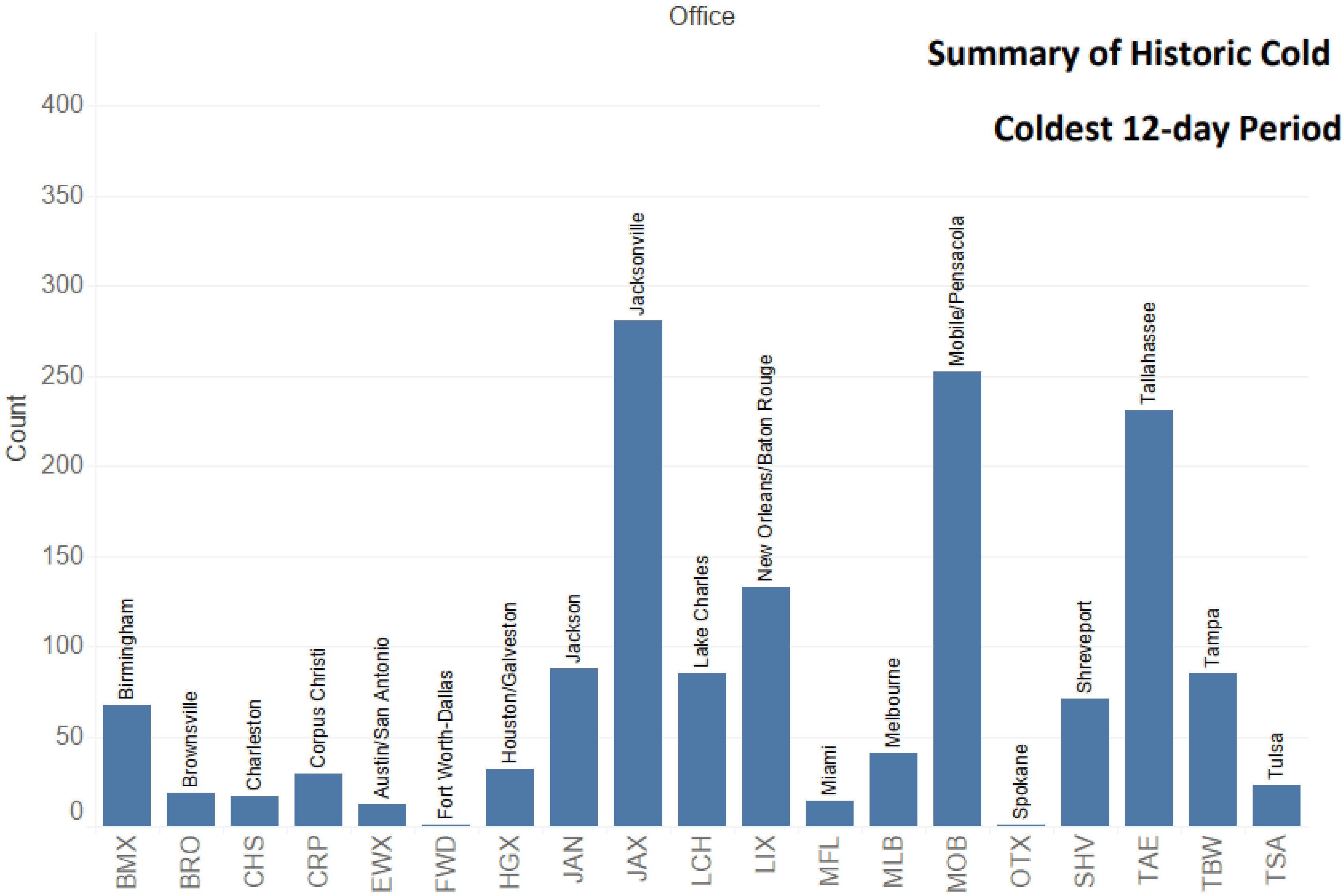
Exploratory Use Case: Event Identification

- 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



Summary of Historic Cold Episode of January 2010

Coldest 12-day Period Since At Least 1940

January 2010

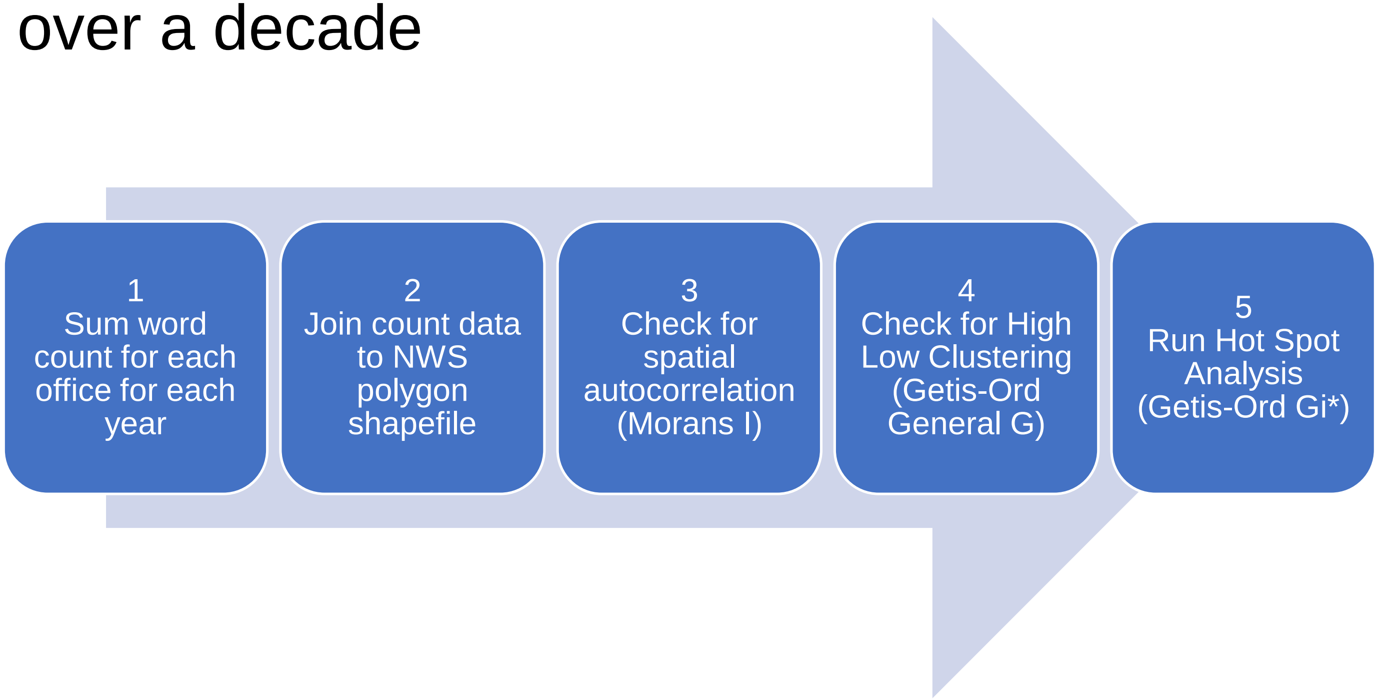
Record cold temperatures in Florida including Jacksonville, Tallahassee

Baton Rouge

Other parts of the south

Exploratory Use Case: Method

- 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



1 Sum word count for each office for each year

2 Join count data to NWS polygon shapefile

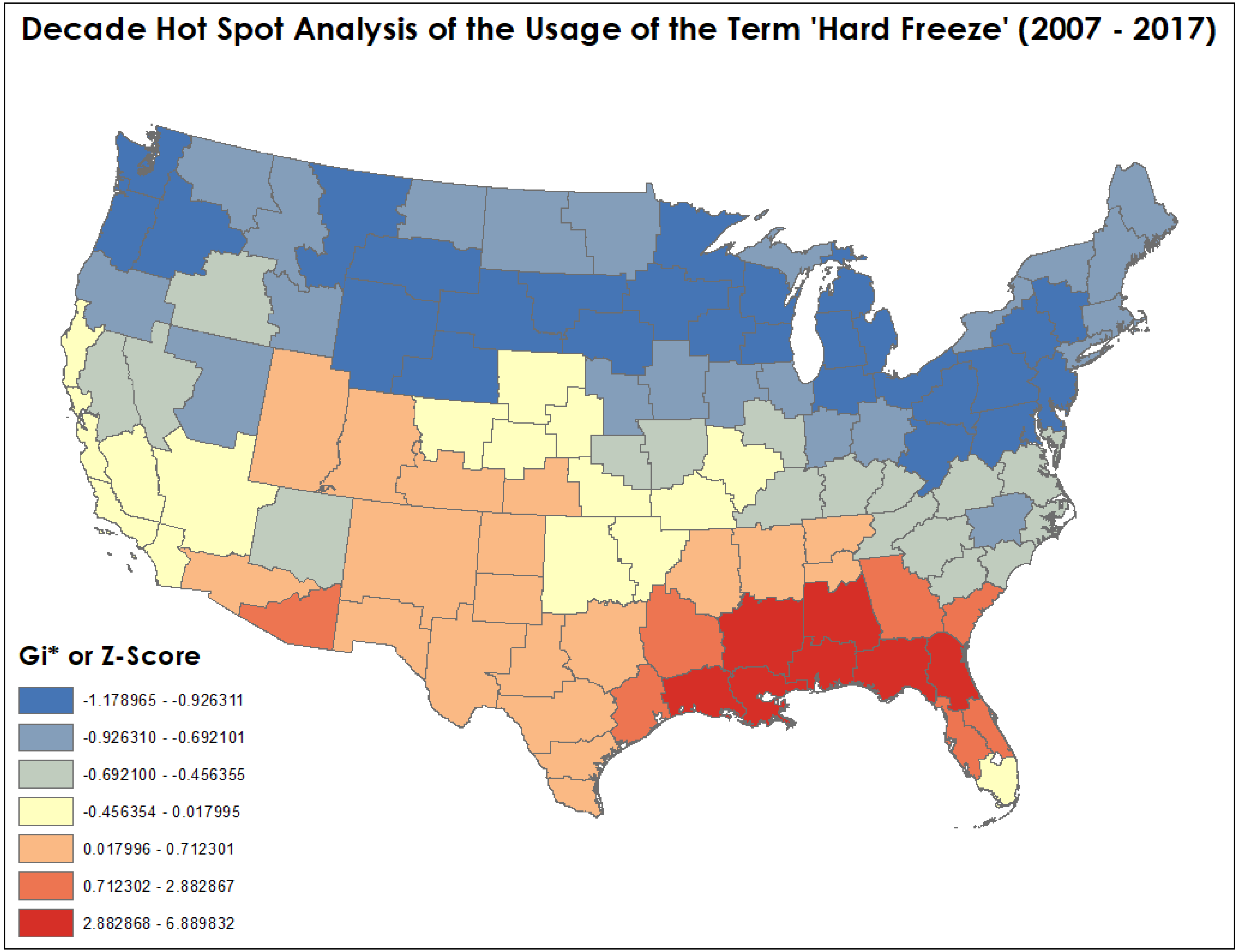
3 Check for spatial autocorrelation (Morans I)

4 Check for High Low Clustering (Getis-Ord General G)

5 Run Hot Spot Analysis (Getis-Ord Gi*)

Exploratory Use Case: Results

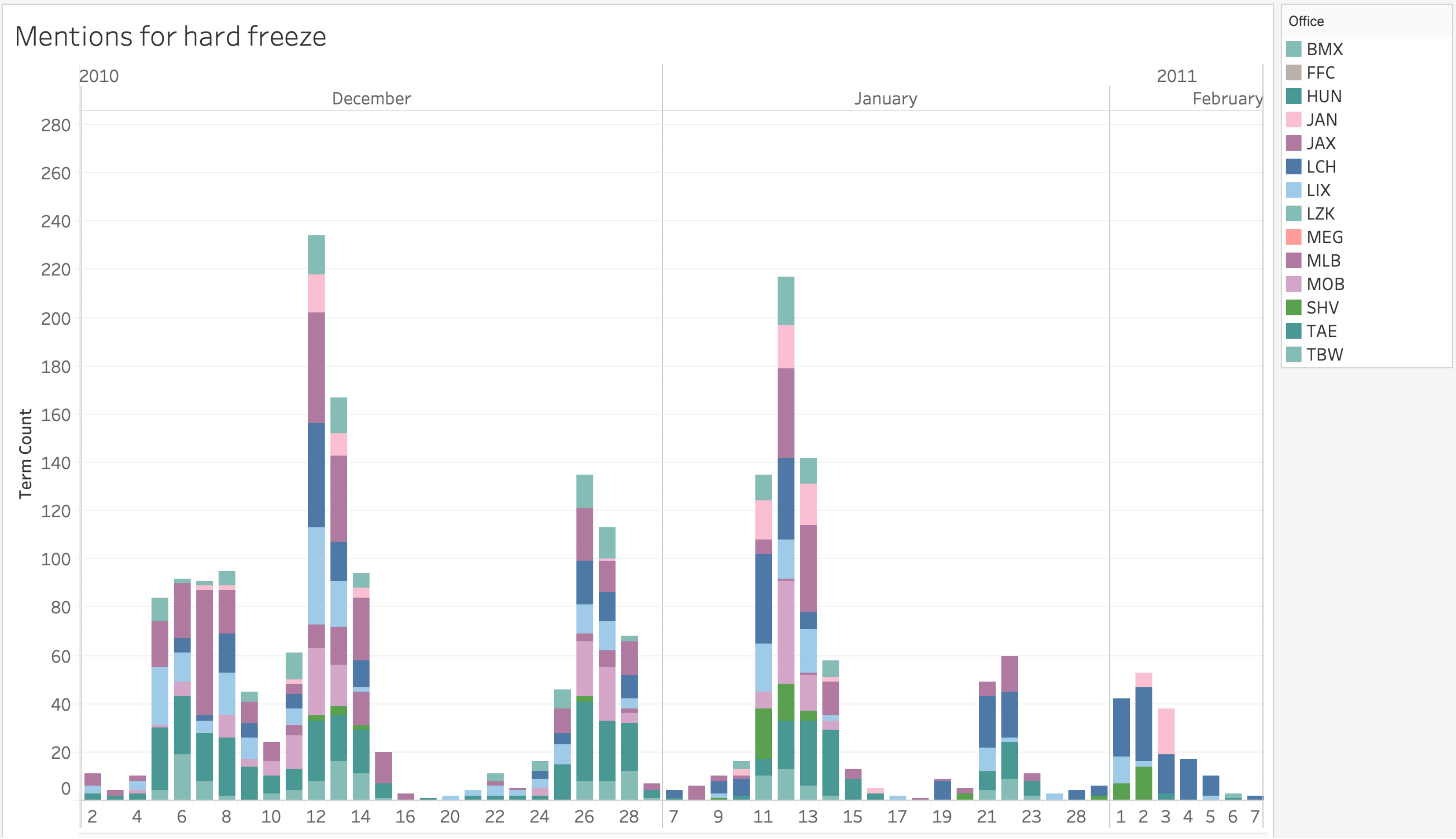
- **Hot spot analysis**
- To be a statistically significant hot spot, a feature will have a high value and be surrounded by other features with high values as well.
- Results indicate positive Z-scores are inversely related to winter temperature trends



Decade Hot Spot Analysis of the Usage of the Term 'Hard Freeze' (2007 - 2017)

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

Created by Gregor Cresnar
from Noun Project

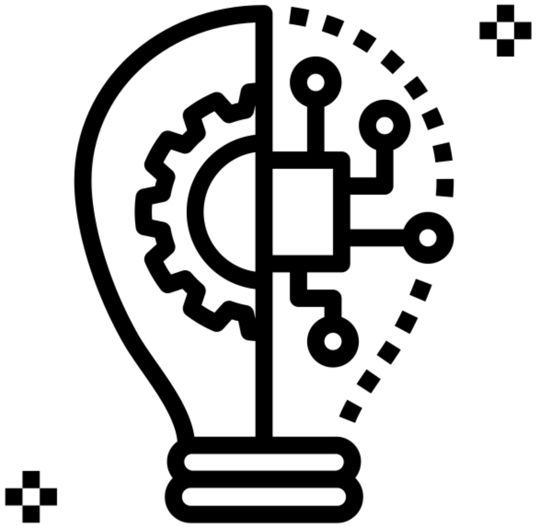
Future Work and Conclusions

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



Created by Nithinan Tatah
from Noun Project

Questions?

Contact:
Kaylin.m.Bugbee@nasa.gov

References

1. https://www.datainnovation.org/2017/04/5-qs-for-ed-kearns-chief-data-officer-at-noaa/
2. https://earthdata.nasa.gov/about/eosdis-cloud-evolution
3. https://earthdata.nasa.gov/getting-ready-for-nisar
4. https://directory.eoportal.org/web/eoportal/satellite-missions/e/edrs
5. https://mesonet.agron.iastate.edu/
6. http://glossary.ametsoc.org/wiki/Hard_freeze