CDI Monthly Meeting 20200812

August 12, 2020: A restoration management support tool and CDI Pop-up Lab

The Community for Data Integration (CDI) meetings are held the 2nd Wednesday of each month from 11:00 a.m. to 12:30 p.m. Eastern Time.

Meeting Recording and Slides

Recordings and slides are available to CDI Members approximately 24 hours after the completion of the meeting.

These are the publicly available materials. Log in to view ALL meeting resources. If you need a CDI log-in to this wiki, join at https://listserv.usgs.gov/mailman/listinfo/cdi-all.

200812-c...ides.pdf

Agenda (in Eastern time)

11:00 am Welcome and Opening Announcements

11:15 am Working Group Announcements

11:25 am **Integrating short-term climate forecast into a restoration management support tool** - Caitlin Andrews, USGS

11:45 am **CDI Pop-up Lab: Q&A with the CDI community**

- Cloud optimized files and new transfer options - Theo Barnhart and Drew Ignizio

- Irregular meshes for data operations - quadtrees - Thomas Rapstine

- Streamstats - Kitty Kolb

- Speaking Git and code.usgs.gov - available resources - Discussion

- Who is in the CDI? Staff Profiles - Leslie Hsu

12:30 pm *Adjourn*

Highlights

1. **Finding CDI Resources**
    a. MS Team - for more frequent announcements and Q&A: https://my.usgs.gov/confluence/x/AI2pJg
    b. Wiki - Hub of information: Home
2. **Machine Learning Mondays**
    a. Learn more and register your interest: https://dbuscombe-usgs.github.io/MLMONDAYS/
3. **Speaking Git, some resources**
    a. 18F: How do I speak Git(hub)
    b. Git(hub)Glossary
    c. GS-Software MicrosoftTeam(and wiki tab)
    d. USGS SoftwareManagementWebsite
4. **Staff Profile Project**
    a. See the data: Staff Profiles
    b. Select or check terms using the USGS Thesaurus: https://apps.usgs.gov/thesaurus/

Notes

1. **Welcome and Opening Announcements**
    a. Preview of Machine Learning Mondays: a weekly course on image analysis using machine learning. Dan Buscombe will be offering this course covering image recognition, object recognition, image segmentation, and semi-supervised image classification. The course is targeted at USGS employees and contractors in the fields of satellite aerial imaging, image analysis, geospatial analysis, machine learning software development. The course is only available to those with a USGS email, with no charge. Experience in Python and the command line interface are recommended as a pre-requisite. Apply here: https://dbuscombe-usgs.github.io/MLMONDAYS/
2. **Working Group Announcements**
    a. For more info on any collaboration area, see https://my.usgs.gov/confluence/x/yhv1I
    b. See slides for more detail on events and upcoming and prior dates.
    c. eDNA
        i. Site is now up and running.
    d. Usability
        i. This month, there will be a town hall meeting on August 19th.
        ii. Next month, there will be another resource review.
    e. Risk
        i. Currently holding the annual meeting online. They are currently on the 2nd out of 3 days of the meeting.
    f. DevOps
        i. Last event was August 4; next event is October 6.
    g. Data Visualization
        i. Going through survey feedback to determine focus of the group.
        ii. Next event is in October
    h. Data Management Working Group
        i. Last event was August 10; Next event is September 14.
    i. Software Development
        i. Next meeting is August 26, topic TBD. Please email Cassandra Ladino or other group contacts if you have a topic for discussion.
    j. Cloud Hosting Solutions
        i. Cloud Hosting Solutions is Soliciting Artificial Intelligence/Machine Learning Use Cases: https://my.usgs.gov/confluence/x/XIFbK
    k. Tech Stack
        i. The IT&I / Tech Stach topic at the Summer Meeting was "Structured Data on the Web - Putting Best Practices into Practice" It was great to see some USGS faces at that session! The recording is here: https://www.youtube.com/watch?v=BLx8y73q4PA
        ii. This month, the TI&I/tech stack webinar is being taken over by the ESIP Summer Meeting Recap Webinar. https://www.esipfed.org/get-involved/telecon-calendar
3. **Integrating short-term climate forecast into a restoration management support tool** - Caitlin Andrews, USGS
    a. The goal of this project is to create a link between data and how it can be used in a management context.
    b. Climate forecasts are typically spatially or temporally coarse data, while managers need more temporally fine and site-specific data.
    c. For example, the success of seeding and planting rely on short-term monthly and seasonal climate that occurs immediately after seeding and planting. There is a 90% failure rate for seeding and planting in the western U.S.
    d. The project will facilitate the link between climate data/climate knowledge and management need by creating a short term moisture forecaster app.
    e. In the western U.S., water is limiting factor and drought is natural part of ecosystem. Drought will be exacerbated in years to come.
        i. For managers, seeding/planting and drought are connected
        ii. Water availability is based on soil moisture instead of precipitation.
        iii. SOILWAT2 model is essentially a translation tool. Give the model info on a specific site (climate, vegetation, soil), where water moves on daily basis and the model measures soil moisture at different depths.
    f. Management need: managers need more info on climate forecast for after they seed or plant. Climate knowledge for this use case is probabilities on whether conditions will be hotter/colder and dryer/wetter. This is coarse information that needs translation so managers can use it.
        i. The app develops code to synthesize short term climate predictions to a finer temporal and spatial scale in order to derive the soil moisture model
    g. Generating multi-month forecasts of climate & soil moisture
        i. This multi-month forecast data is very coarse
        ii. The National Weather Service provides one prediction for each of 102 regions for a time period
        iii. Monthly or multi-month scale
        iv. Data is conveyed through graphics
    h. Spatially and temporally refining this data was a challenge
        i. A Jupyter Notebook detailing the steps the project team took is available here for USGS employees: https://code.chs.usgs.gov/candrews/shorttermdroughtforecaster
        ii. Steps were:
        iii. Gather a historical record of site-specific data from GridMET (1980-yesterday)
        iv. Generate samples of what the future will look like - 30 future realizations
        v. Apply future realization to the years in the historical record. This is how future anomalies are integrated with historical patterns.
        vi. Produces 900 climate futures
    i. Example output
        i. See recording for graphs and explanation of graphs
    j. Ecologically relevant metrics
        i. Prediction of probability of establishment of sagebrush seeding
        ii. Mean temperature in 250 days following seeding and soil volumetric water content 70 days after
    k. Integrating this application in the Land Treatment Exploration Tool (LTET), a Bureau of Land Management and USGS collaboration. LTET is a tool for managers planning restoration projects.
        i. Technical information on integrating this project's new application with the LTET on slides and in recording
    l. Takeaways
        i. Producing a management relevant ecological forecasting tool
        ii. Targeting a major restoration challenge
        iii. 10s of millions spent on restoration with a 90% establishment failure is bad news bears
        iv. App is useful as is - as new ecologically relevant metrics are wanted, or if better forecasting comes around, we can include those.
4. **CDI Pop-up Lab: Q&A with the CDI community**
    a. *Cloud optimized files and new transfer options - Theo Barnhart and Drew Ignizio*

i. Theo Barnhart
   1. CDI project this year is to generate a set of continuous basin characteristics for all of contiguous U.S. Leaves us with lots of very large GeoTIFFs and a need to disseminate in an efficient manner.
   2. Need: want something that was a geospatial format, would be easy to generate, good compression, something that could stand alone, don't want to be stuck maintaining a server needed to access data.
   3. Solution: Cogeo.org and Rasterio!
   4. How: Trial and error process by working through examples from Cogeo website. Used Jupyter Notebooks to start generating files and uploading them to a friendly S3 bucket. It only took a few hours to get a working example.

ii. Drew Ignizio
   1. Working on approach for handling files from the ScienceBase side.
   2. Cloud-optimized GeoTIFF (COG): why is it useful?
      a. ScienceBase has been integrating with CHS and AWS resources to solve some large file issues.
      b. In one approach, a user can download a 240 gig file from where it is stored in an S3 bucket.
      c. After downloading, the user can then work with data locally.
   3. Is this still the best way?
      a. Users can avoid downloading, instead just accessing the file in place.
      b. COG enables users to publish to a public S3 bucket and connect to the COG through a Jupyter Notebook. They can also be read directly from a viewer.
      c. For display and analysis where you need the value, COGs provide a nice way to get at the data without downloading the whole resource. Can be retrieved from an http request.

b. ***Irregular meshes for data operations - quadtrees*** - *Thomas Rapstine*
   i. Need stemmed from a project mapping out ground failure for project in Alaska
   ii. Real-time earthquake products can map out very quickly in places there may not be sensors by social media (Did You Feel It).
   iii. Issue: Too much variety. The inputs to models are grids, points, polygons, lines (a lot of geometric variety). Data can be categorical, physical, or temporal. Inputs come with their own notion of uncertainty, or not. Could be pulled from a global raster or local.
   iv. Challenge and approach: diverse datasets: how can we structure them in a way that enables robust, calculable integration and evaluation?
      1. Use multi-scale, hierarchical data structure to represent data on varying scales; representation that allows for multiple resolution grids to be put together (a quadtree). Quadtree is dividing regions into squares, can subdivide to be more refined.
   v. Solution:
      1. Quadtree mesh was built using points. The result is finer representation in the mesh areas.
      2. Using python package discretize
      3. **Questions for CDI**
         a. How are others solving these data integration issues?
         b. Any other solution recommendations other than quadtrees?
         c. Thoughts on using quadtrees for solving these challenges?
         d. Are you using quadtrees? What packages would you recommend?

c. **Streamstats - Kitty Kolb**
   i. StreamStats is a USGS website that allows users to delineate watersheds for an area of interest. Built to be used by civil engineers to design highway bridges and culverts.
   ii. Broad question: What's the biggest flood I can expect in a given year? How do we get information on un-gaged areas?
   iii. Technical need: A GIS system to calculate things more quickly than the old fashioned planimeter method for un-gaged streams.
   iv. Solution to the need: StreamStats.usgs.gov built on ArcHydrol, SSHydro, and Leaflet. REduces planimeter method from hours to minutes. Provides an image of your watershed and a report. Can download watershed outline and table. Can use StreamStats API to incorporate delineations into your own mapping tools. Consistent, repeatable results in a timely manner.
   v. How I learned it: StreamStats Training docs and webinars, classes on ArcHydro

d. **Speaking Git and code.usgs.gov - available resources - Discussion**
   i. Speaking Git Links are posted on the meeting page: https://my.usgs.gov/confluence/display/cdi/CDI+Monthly+Meeting+20200812
   ii. Discussion on mysterious Git terms.

e. **Who is in the CDI? Staff Profiles** - Leslie Hsu
   i. 47 profiles/193 keywords for the CDI community
   ii. To view staff profiles that were used in the analysis: https://my.usgs.gov/confluence/display/cdi/Staff+Profiles
   iii. There are a surprisingly good number of expertise keywords that are in the USGS Thesaurus - this will help to align and identify people with relevant expertise https://apps.usgs.gov/thesaurus/
   iv. There is a look-up feature in the Thesaurus that allows you to find out what terms are preferred or not in from the Thesaurus: https://apps.usgs.gov/thesaurus/term-check.php

Questions & Comments

1. **Integrating short-term climate forecast into a restoration management support tool** - Caitlin Andrews, USGS
   a. Is there a particular reason you'd expect the temperature and precipitation anomalies to be normal? My experience is they're generally weibull or otherwise fat-tailed.

      i. Caitlin: They're not expected to be normal. Temperature is expected to be normal, but forecast is taken to a power function. When we predict, we back transform it so it's normal, then transform it to the power again. Details are in Jupyter Notebook for exactly what we did.

   b. When did you start using Notebooks to document your workflows?
      i. Only recently. Heard about Jupyter Notebooks at a CDI meeting. Felt fatigued by sharing R studio screen with team; Notebooks is better for sharing code. It was an iterative process to get forecasts where we wanted them. Wanted to make sure not to forget decisions made through the process.

   c. It's a bit unclear how you converted the multi-month (qualitative?) forecast to a daily quantitative forecast.... can you briefly restate this...?

      i. See recording.

   d. Do you have any metrics/stats on BLM use of the tool for decision support?

i. No. There are metrics for BLM support of the Land Treatment Exploration Tool.

e. Jake Weltzin: All, speaking of Ecological Forecasting, the 2019 Ecological Forecasting workshop report was just released as an open-file report, here: https://pubs.er.usgs.gov/publication/ofr20201073

f. Can you describe how you've worked with managers to create the tool? For example - what types of visualizations are most helpful for them? How will they be using the tool?

i. We hope for managers that they will be going to the LTET, already querying management area. Made drafts and figures shown in the pres. Hope to put those in the tool. Want to be concise. Thought is to come up with first draft of figures, build app around them, test user groups, then alter figures form there.

g. LTET: Justin Welty

i. LTET - this is our first year as an official tool. Working with BLM on ESR wildfire restoration projects. Every fire that burns in the West has to go through an ESR treatment process, going through LTET. LTET will attach a report to restoration plans with metrics on ownership, precipitation, temperature, and sagebrush environment. This gives federal/state stakeholders.

ii. If interested further in LTET, contact mjeffries@usgs.gov or LTDL_Project@usgs.gov

2. **Cloud optimized files and new transfer options** - Theo Barnhart and Drew Ignizio

a. I don't follow how the 'author' or data creater can move large files about. I don't know how to move large files outside of using ftp. Are there other bureau services?

i. Drew Ignizio: transferring files about is still a challenge yes. We have some new functionality we'll be rolling out soon that will help users upload large files directly to ScienceBase cloud storage. but yes, moving things around still requires some flexibility and creativity. If you have a case, please contact me and we can discuss options (dignizio@usgs.gov / sciencebase@usgs.gov)

b. What is the difference between this and a WCS?

i. good question. Traditionally a WCS has offered the same type of functionality (view data / consume in an app / pull values into analysis). However, a WCS requires an actual GIS server, either something like Geoserver or ArcGIS Server. So you have to store the data where that server can reach it, then spin up the service. That requires more resources than just storing the data and accessing it directly there, 'in place' so to speak. Additionally, I've found some limits to WCS being used effectively in analysis (say, as an input to a tool). COG seems to be able to provide much more efficient subsetting to pull values in one area /extent, and is also gaining support in things like gdal and rasterio.

c. I'd love to hear much more about S3/COG and Sciencebase...

i. Drew Ignizio: let's chat. I'd be happy to share notes, discuss options, etc.

d. Is there a repository with example code using COG?

i. Theo Barnhart: Check out make_fcpg here: https://code.usgs.gov/StreamStats/FCPGtools/-/blob/master/FCPGtools/tools.py

e. I'm interested to hear more about the limits you've encountered with WCS. As you know we've been using WCS regularly to access inputs to ARIES but are exploring/open to the idea of COGs too. Are there other problems you've seen with WCS aside from efficiency and greater support from the packages you mention?

i. Dave Blodgett: it's a little further out, but I'm really excited to see where this will go in relation to some of the limits of WCS: http://docs.opengeospatial.org/DRAFTS/19-086.html

ii. Drew Ignizio: In the past, I've found varying success with just providing a WCS uri or layer as a direct input to say a tool in ESRI. It can work well but sometimes is finicky. Additionally, as someone who works with a team of hosting these services, having the option to to tell users: 'Just get the data here, in this way' instead of maintaining all those services, that is advantageous. Also I find myself working just in Python with GIS libraries more, and the COG functionality has gained a lot of traction there. I also think COG supports better subsetting of values than WCS, but to be fair, there may be better strategies for WCS that I've overlooked before.

f. Perhaps I have missed previous talks, but would generally love to see more involved discussions of: zarr, netCDF, COG, S3, and Sciencebase. Basically I am drowning under "large" (> 1GB) netCDFs, and have stuggled to find a way to publish and share these files, particularly any solution that requires one to be logged into Pulse Secure.

i. Drew Ignizio: this is all pretty new but we should probably have a chat, including with Rich Signell, who had been involved with our conversations and attempts here. This is an example Rich put together demonstrating a similar type of access (just read straight from S3) for a netCDF that had been converted to zarr: https://gist.github.com/rsignell-usgs/3cbe15670bc2be05980dec7c5947b540

g. Drew Ignizio: anyone else who's interested in hammering on some of these COGs to see if they live up to the hype with a real workflow, here's the example I showed in ScienceBase: https://www.sciencebase.gov/catalog/item/5e693773e4b01d50925da94b

h. Drew Ignizio: Here is the link to the COG viewer I showed. You can provide a URL to any publicly accessible COG in the box there, and view the data too. https://www.cogeo.org/map/#/url/https%3A%2F%2Fprod-is-usgs-sb-prod-publish.s3.amazonaws.com%2F5e7d36c1e4b01d5092751e09%2FWhiskeytown_2019-06-03_DSM_25cm_hll.tif/center/-122.5737,40.6379/zoom/13

3. **Irregular meshes for data operations - quadtrees - Thomas Rapstine**

a. How do you store the quadtree?

i. it depends on the implementation of the quadtree data structure. It's a tree-like data structure, so you can envision storage being a connection of nodes with parents and children. The specific quadtree object I'm using is here: https://discretize.simpeg.xyz/en/master/api/generated/discretize.TreeMesh.html#discretize.TreeMesh.

b. Response to Thomas' questions:

i. We have done some testing of large datasets with sparse coverage using the Meta Raster Format, which uses an index to subset data on the fly, and understands NoData values. https://github.com/nasa-gibs/mrf/blob/master/doc/MUG.md. Our use case: https://planetarygis.blogspot.com/2019/01/uncontrolled-global-hirise-mosaic.html

ii. NRCS has been dealing with this issue for a while - it is their Soil Database Join and Reconciliation project. but, not sure how they did it.

4. **Speaking Git and code.usgs.gov - available resources - Discussion**