CDI Monthly Meeting 20200311

March 11, 2020: CDI Projects - Subsidence, Biosurveillance, Invasive Species

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 public slides. Log in as a CDI member to view ALL of the meeting resources, including recording.
If you would like to become a member of CDI, join at https://listserv.usgs.gov/mailman/listinfo/cdi-all.

CDI_2020...ides.pdf

200311-C...ents.pdf

Agenda (in Eastern time)

11:00 am Welcome and Opening Announcements

11:15 am Working Group Announcements

11:30 am **Subsidence Susceptibility Map for the Conterminous U.S.** - Jeanne Jones, USGS

11:45 am **High-Resolution, Interagency Biosurveillance of Threatened Surface Waters in the United States** - Sara Eldridge, USGS (Elliott Barnhart, USGS presented)

12:00 pm **National Public Screening Tool for Invasive and Non-native Aquatic Species Data** - Wesley Daniel, USGS

12:30 pm *Adjourn*

Highlights and Links

1. Sign up for Group Learning with the CDI for the Spring
2. 2020 CDI Funded Projects Announced
3. New EarthMAP information links (for USGS employees) | recent blog post | Intranet page | MS Team
4. Send feedback about elements of the EarthMAP conceptual model to help the EarthMAP project team
5. What's happening around the CDI - Join a Collaboration Area here
6. Dave Blodgett (dblodgett@usgs.gov) wants your suggestions for presentations to the Tech Stack group on the theme "Putting Data to Work"
7. Jeanne Jones presented on the subsidence susceptibility map - the team built a national-scale map of sinkhole subsidence susceptibility. This dataset is being incorporated with other hazard and risk layers to inform Dept of Interior agencies. The output is also a dataset ready for machine learning.
    a. **Jeanne's question to the CDI: How do different methods for flow accumulation processing with DEMs compare in terms of speed, consistency of results, max size of raster for high performance computing? (for example, Arcpy, TauDem, RichDem) Contact Jeanne at jmjones@usgs.gov**
8. Elliott Barnhart presented about a project to incorporate Environmental Sample Processors (ESPs) at USGS stream gauging stations and collect near real-time eDNA surveillance of invasive species or pathogens. The project developed a data science pipeline and highlighted the benefits of combining information and methods from MBARI, KBase, and NGWOS (USGS Next-generation Water Observing System).
9. Wes Daniel presented about SEINed - a tool for Screening and Evaluating Invasive and Non-native Data. This tool will be launched in April on the Nonindigenous Aquatic Species site, and help to get non-native species occurrence data from groups not focused on invasive species.

Notes

1. **Opening**
    a. Group learning opportunities
        i. Sign up for topics on usability, intro to netCDF, Unit Testing, Microsoft Power Automat and Power Apps here

2. **Tim Quinn chat**
    a. Announcement of the 2020 CDI Funded Project Teams
        i. Fourteen projects funded - see full list here
        ii. Many projects address broader USGS goals of expanding predictive capabilities and actionable intelligence
        iii. We thank all applicants and appreciate their attention and hard work!
    b. CDI is a discussion venue
        i. CDI is a diverse group with different interests and areas of expertise, one of its strengths
        ii. Any CDI member should feel comfortable to ask a question, offer points of view to help others, and build connections
        iii. All feedback is used in planning future CDI activities
        iv. Opportunity to provide feedback today for the EarthMAP project
3. **EarthMAP update**
    a. New communications venues
        i. Recent blog post
            1. Covered the why, how and what of EarthMAP (see slides or blog post)
        ii. Intranet page
        iii. Microsoft Team
    b. EarthMAP conceptual model
        i. Form on which portion of the model interests you here
        ii. Venn diagram: data & information integration, integrated predictive science, and actionable intelligence, with EarthMAP int the middle!
            1. Data & information integration
                a. Improved framework for science data and information (collecting, assessing, analyzing, integrating)
                    i. Readily available and accessible
                    ii. Embrace relevant data
                    iii. Recognize changing definition of data
            2. Integrated predictive science
                a. a system of integrated, scalable models that simulate and predict changes in connected human and natural systems
                    i. Advanced modeling
                    ii. Integrated across boundaries, disciplines, geographies, sectors
                    iii. Developed in collaborative partnership with stakeholders
            3. Actionable intelligence
                a. Observation and predictions developed with partners to provide information at the speed and scales needed to inform their decision-making
                    i. Decision support tools and processes
                    ii. Operational capacity
                    iii. Iterative improvements
            4. How do your projects fall into EarthMAP's model? One or two areas? In the middle?
4. **Collaboration Area Announcements**
    a. Interagency Collaborative for Environmental Modeling and Monitoring
        i. Next event: Annual ICEMM meeting scheduled for March 17-18, 2020 at USGS in Reston, VA: "Integrated Modeling, Monitoring, and Working with Nature"
        ii. Contact: pglynn@usgs.gov
    b. Semantic Web Working Group
        i. Thursday, March 12th, 12pm MT, discussion with Sky Bristol: a practical example of semantic technology in action
        ii. Contact: Fran Lightsom, flightsom@usgs.gov
    c. Metadata Reviewers Community of Practice
        i. Next meeting: April 6th, 12pm MT: What about metadata for software and code?
        ii. Contact: Fran Lightsom, flightsom@usgs.gov
    d. Tech Stack Working Group
        i. Next meeting: March 12th, 3-4pm ET: "Discrete Global Grid Systems in action: Provision of rapid response during Australian bushfires and other applications" by Shane Crossman and Irina Bastrakova
        ii. Looking for ideas for future tech dive talks on this years ESIP theme: Putting Data to Work. Please email dblodgett@usgs.gov wi th ideas
        iii. Contacts: Dave Blodgett, dblodgett@usgs.gov; Rich Signell, rsignell@usgs.gov
    e. Data Management Working Group
        i. Next event: April 13th: "Upcoming changes to the Science Data Catalog", Lisa Zolly
        ii. In past meetings, created draft value propositions
        iii. Contact: Madison Langseth, mlangseth@usgs.gov
    f. Risk Community of Practice
        i. Next meeting: March 19th, 1PM ET: "Human Centered Design and Inclusive Problem Solving Training with Impact 360, part 2" Register here
        ii. Funding 7 projects of 29 received for the FY20 Risk funding
            1. See 3/6/20 Leader's Blog / recent NTK for full list of awards
        iii. Contact: riskyworld@usgs.gov
    g. Open Innovation Community
        i. Next meeting: March 12, 10:30am PT: "Innovation Center Talk: Automatic satellite-based flood mapping for disaster response" (More details here)
        ii. Next meeting: March 16th, 3PM ET / 9AM HT: Using Volcanic Hazards in Hawai'i as a STEM platform for problem-based learning with raspberry shakes
        iii. Received Risk funding for Open Innovation Playbook for Risk
        iv. Working on open innovation community newsletter
        v. Contact: Sophia Liu, sophialiu@usgs.gov; openinnovation@usgs.gov
    h. Software development
        i. Next event: March 26th, 3:30pm ET: "Cloud Efforts - Automated deployment for scientific processing with AWS cloud formation" by Kirstie Haynie
        ii. Contacts: mguy@usgs.gov; jknewson@usgs.gov; ccladino@usgs.gov
    i. Usability
        i. Next event: March 18th, resource review: "Using analytics to inform how our web pages/tools are being used"

ii. Next town hall meeting: April 15th, 3pm ET/1pm MT: "How to select test users and how many to test?"
iii. Contact: Sophie Hou, chungyihou@usgs.gov

5. **Subsidence Susceptibility Map for the Conterminous U.S.** - Jeanne Jones, USGS
   a. Subsidence susceptibility - sinkholes and areas susceptible to developing sinkholes
      i. Focused on karst regions
   b. Why is this important?
      i. Sinkholes are hazardous; focus contaminated/polluted surface water into groundwater
      ii. Create instability in the foundations of buildings roads, etc.
   c. The U.S. lacks a consistent national map
   d. Working to incorporate this dataset into the SHIRA (CDI) Risk map for use by DOI emergency agencies
   e. Used the National Map, karst research, and the Yeti supercomputer
   f. Five step process
      i. Hydrological conditioning of DEM
      ii. Identification of closed depressions
      iii. Screening and morphometric statistics
      iv. Validation against state maps
      v. Creation of heat map
   g. See slides for diagram on processing steps for conditioning DEMs and finding closed depressions
   h. See slides for map of sinkhole hotspots
   i. Challenges
      i. Data collection and screening
         1. Screening data visually
            a. Patching in other data to close gaps
         2. Screening data in Python
      ii. Processing issues, edge effects
         1. DEM was too large to process well with ArcGIS
         2. Had to do each individual DEM at a time
      iii. Open source software
         1. Existing data used different software that defined some terms and statistics differently
      iv. Closed depression screening
         1. Screened out wetlands, open water, urban areas; soils with a flood signature; quarries or strip mining sites; to shallow, too small, wrong shape, close to roads (drainage ditches), etc.
      v. Post-processing
         1. Used geologic information and expert knowledge to remove depressions that may have formed through non-karst processes
   j. Project data, tools, products to share
      i. Closed depression polygons
      ii. Sink density and hot spot raster datasets
      iii. 10-meter DEMs, NHD streams and roads on Yeti
      iv. Technique - pubs by Dan Doctor and others
      v. Code on code.usgs.gov
   k. Follow-up collaboration
      i. Have a great training data set for machine learning
   l. Follow-up question for CDI
      i. Flow accumulation for processing with DEMs
         1. Arcpy, TauDem, RichDem: How do these compare?
6. **High-Resolution, Interagency Biosurveillance of Threatened Surface Waters in the United States** - Sara Eldridge, USGS (presented by Elliott Barnhart)
   a. Project to incorporate ESPs at USGS stream gauging stations; to provide near real-time DNA surveillance of invasive species or pathogens
   b. With high frequency data collection, we need rapid analysis
      i. Need to give resource managers time to respond, so time is of the essence
   c. ESPs installed in stream gauges along the Yellowstone River
      i. Tested for non-native species
   d. Needed a way to combine stream gauge data and weather data
   e. Data Science Pipeline
      i. Created a cloud-hosted digital ocean database that combines all the collected data
      ii. Can easily incorporate eDNA and other data streams into models that can indicate presence/absence of organisms
   f. See slides for figure on processing steps on creating the Digital Ocean PostgreSQL database
   g. Challenges and lessons learned
      i. Quality control filters from multiple data sources
      ii. Linking the benefits and capabilities of:
         1. MBARI ESP in situ sample collection and analysis at stream gauges
         2. Dept of Energy Systems Biology Knowledgebase (Kbase), open environment for computational systems biology
         3. USGS Next-generation Water Observing System (NGWOS)
7. **National Public Screening Tool for Invasive and Non-native Aquatic Species Data** - Wesley Daniel, USGS
   a. Central repository for spatially referenced accounts of introduced aquatic species
   b. Tracks over 1,290 aquatic species, with over 600k observations
   c. Data is national, dating back to 1800's
   d. Constantly updating data sources and adding new information
   e. Data aggregated from museum collections, researchers, state and federal agencies, scientific literature, and public sighting reports
   f. The problem
      i. How does the NAS database get non-native occurrence data from groups not focused on invasive species?
   g. Biosurveillance tool
      i. SEINeD tool allows stakeholders to upload any biological dataset; these datasets will be screened for invasive occurrences
   h. See slides for diagram of SEINeD tool process
      i. Automated process
         1. Checks the spatial accuracy
         2. Checks for taxonomic errors (misspelled names, old or non-specific nomenclature)

3. Native status filter
    a. Flags non-native species that are exotic (from other countries/continents), AND non-native species from within the U.S. (rainbow trout native to the west coast on the east coast)
4. SEINeD Tool does not store any of the information provided
    a. User receives two CSVs back: the original, and one that only contains the data that SEINeD would like to use.
5. Benefits
    a. Automated
    b. Easy way to link sampling data to multiple spatial GIS layers
    c. Early detection screening tool for potential new invasions
    d. Incentivize stakeholders that utilize the tool to contribute their data to the NAS program
    e. Increase the visibility of the NAS program
6. SEINeD goes live May 4th: https://nas.er.usgs.gov

Questions

1. Jeanne: when you found "strange" data (weird min/max, no data), did you have a path to communicate that to someone? Were there contacts for these datasets?
    a. Talked to National Map people about the gaps in road data
    b. Put a call out on the GIS talk listserv and asked about missing points
    c. Some incorrect grids may have been due to proximity to water
2. PRISM delivers historical data (lagged by a year, right?); but, one could 'get' real-time or forecasted met data from NCEP... What is the latency on the eDNA data? If there was an emergency application, what could you reduce it to (at least pragmatically for now)?
    a. Historically, it takes a long time to analyze DNA data. However, it can be provided a lot faster now. With MBARI's 2nd generation robot, it only needs a half hour to analyze. QPCR and others are still a bit of a challenge. Right now, a half hour is the time it needs, but could become faster with new technology, or slower with more data.
3. Thank you for the great presentation Jeanne! Suggestion: clarify distinction between sinkhole subsidence vs GW-pumping subsidence (i.e. CA, Houston)
4. Jake - can you expand NCEP?
    a. National Center for Environmental Prediction...essentially the source of National Weather Service data
5. How did you find the right contacts to work on this project?
    a. For this project, the right contacts were already in place. Going to conferences is a great place to meet potential collaborators. First ran into MBARI robots at a conference.
6. Wes, can you give an example of users you are targeting that have this info but are not usually concerned with non-native species? (what orgs or professions?)
    a. NGOs, university that are conservation-focused, state employees who don't have the time to screen data themselves.
7. How do you validate the scientific names?...assuming you compare them to an index?
    a. ITIS index for scientific names. Also using an internal index for more recent changes.
8. As a non-specialist, can you suggest some news sources to get current news on invasive and non-native species?
    a. NAS database has an alert email and Twitter account that notifies users of new non-native species. Depending on the region, you can contact your regional representative (email Wesley, wdaniel@usgs.gov for info on this)
9. What type of outreach will you do to let people who don't know about NAS know about this new tool?
    a. Canvas as many resources as possible: professional society newsletters on fish/water, tapping botanists to look at other professional societies, social media, internal news link, advertising through all state contacts
10. Seined wont automatically harvest the 3rd party dataset after tagging; but, what are the incentives to get users to come back and share their dataset with "NAS... Could you build a checkbox so that their analyzed set could be shared with NAS "automatically?"
    a. Initial thought is that the second CSV comes with an email talking about the importance of sharing the data. Many people are not getting back to NAS with the second CSV.
11. iNaturalist might be great early collaborator
    a. Work closely with iNaturalist, will notify them of this new tool.