



## **An Extensible Geospatial Data Framework (GeoEDF) for FAIR Science**

**Dr. Carol Song**

Sr. Research Scientist, Purdue Univ.

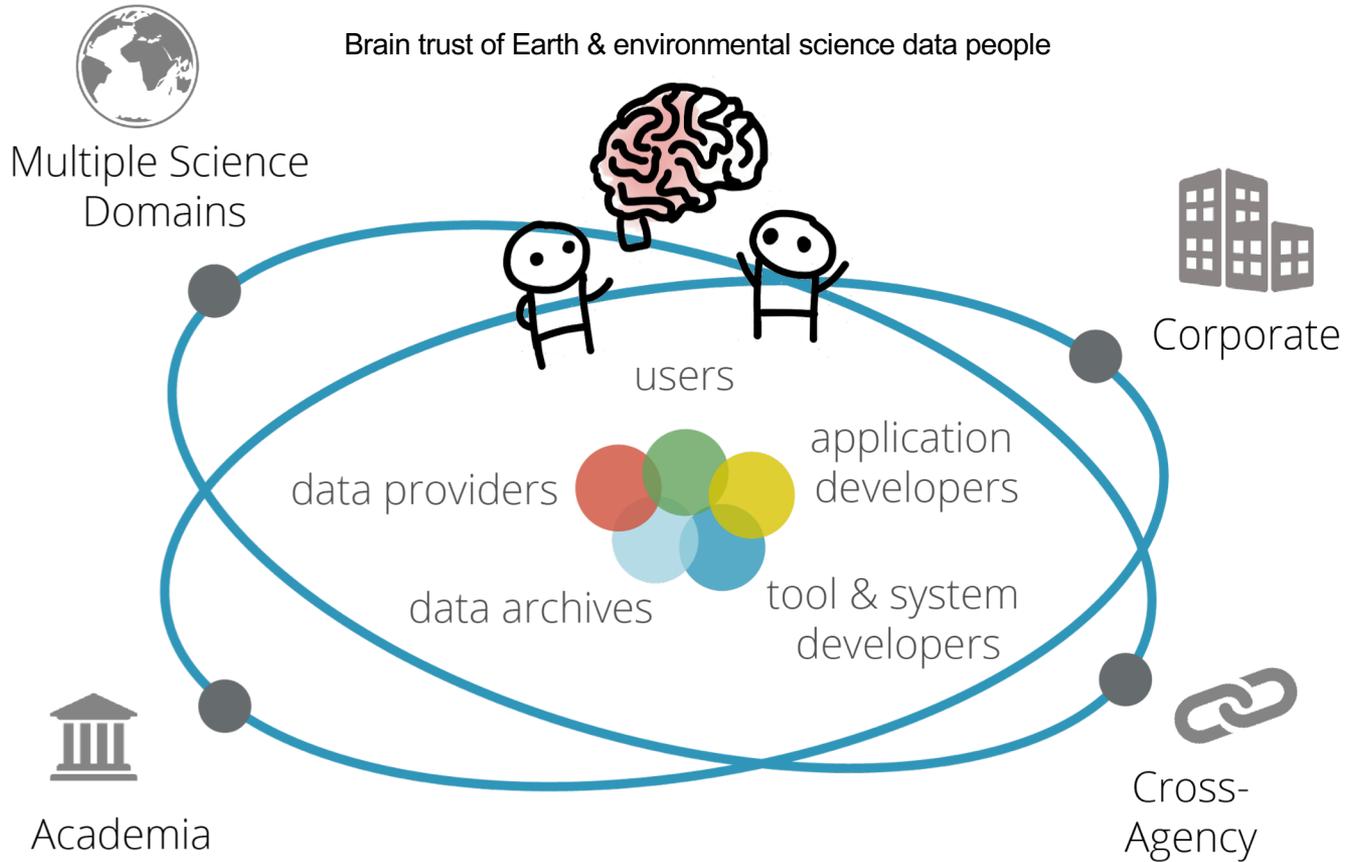
**Assisted by: Rajesh Kalyanam**

Research Scientist, Purdue Univ.

**Data to Action Webinar: Increasing the Use and Value of Earth  
Science Data and Information**

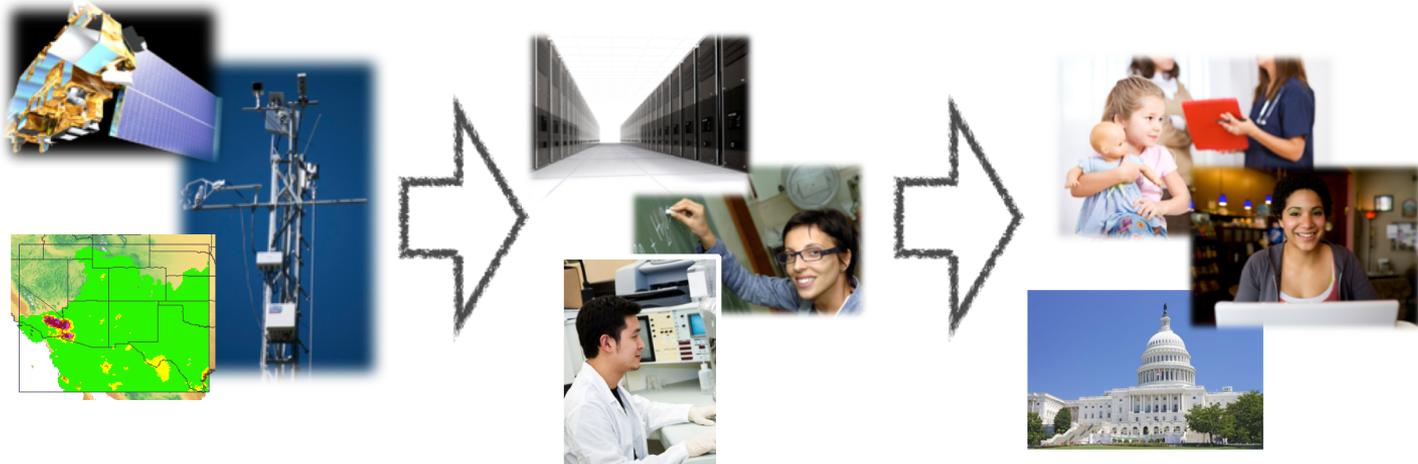
October 25th, 2019 | 1:00 pm ET

# ESIP COMMUNITY



# ESIP Vision

*To be a leader in promoting  
the **collection, stewardship and (re)use**  
Of Earth science data, information and knowledge  
that is responsive to societal needs.*





## **An Extensible Geospatial Data Framework (GeoEDF) for FAIR Science**

**Dr. Carol Song**

Sr. Research Scientist, Purdue Univ.

**Assisted by: Rajesh Kalyanam**

Research Scientist, Purdue Univ.

**Data to Action Webinar: Increasing the Use and Value of Earth  
Science Data and Information**

October 25th, 2019 | 1:00 pm ET



NSF AWARD #1835822

# An Extensible Geospatial Data Framework (GeoEDF) for FAIR Science

Carol Song, Rajesh Kalyanam, Purdue University

ESIP “Data to Action” Webinar  
OCTOBER 25, 2019



# The GeoEDF Project

---

An Extensible Geospatial Data Framework Towards FAIR Science

To help data-driven sciences to be more  
Findable, Accessible, Interoperable, Reusable

funded by NSF CSSI program (Cyberinfrastructure for Sustained  
Scientific Innovation), Data Framework track, \$4.5M

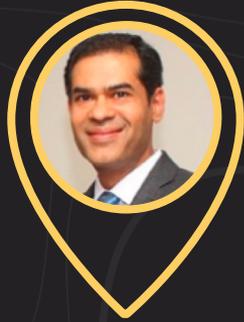
October 2018 - September 2023

# Project Leadership



Jian  
Jin

Plant phenotyping  
& sensors



Venkatesh  
Merwade

Flood modeling  
& visualization



Carol  
Song

PI  
Cyberinfrastructure



Jack  
Smith

Water Quality  
& resource  
management



Uris  
Baldos

Sustainable  
development

OVERVIEW

# Extensible, Geospatial Data Framework Towards FAIR Science (GeoEDF)

# Rapid growth of geospatial, geo-referenced data

Hazard data provided by government

Data made public by cities

Search to find a specific dataset...

BROWSE THE DATA CATALOG BY THE FOLLOWING CATEGORIES

- Administration & Finance
- Buildings
- Community
- Education
- Environment
- Ethics
- Events
- FOIA
- Facilities & Geo. Boundaries
- Health & Human Services
- Historic Preservation
- Parks & Recreation
- Public Safety
- Sanitation
- Service Requests
- Transportation

USGS science for a changing world

Earthquake Hazards Program

Data & Products

Data

Products

Earthquakes

Hazards

Data & Products

Learn

Monitoring

Research

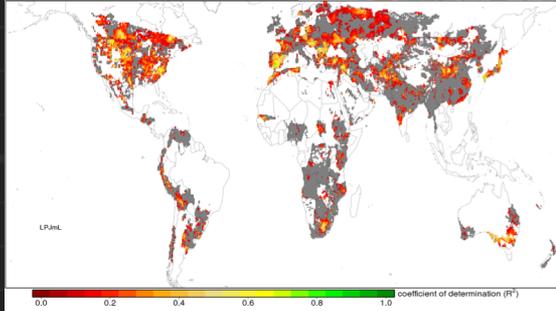
Web Services

[Web Services](#) for real-time and catalog earthquakes, hazard maps, slabs

Earthquakes

- [Real-time Data Feeds](#)
- [ANSS Comprehensive Earthquake Catalog \(ComCat\) Documentation](#)
- [Preliminary Determination of Epicenters \(PDE\)](#)
- [Regional Seismic Networks](#)

Specialized Earthquake Catalogs

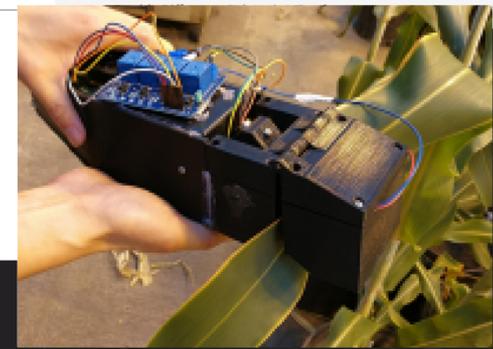


Modeling output data



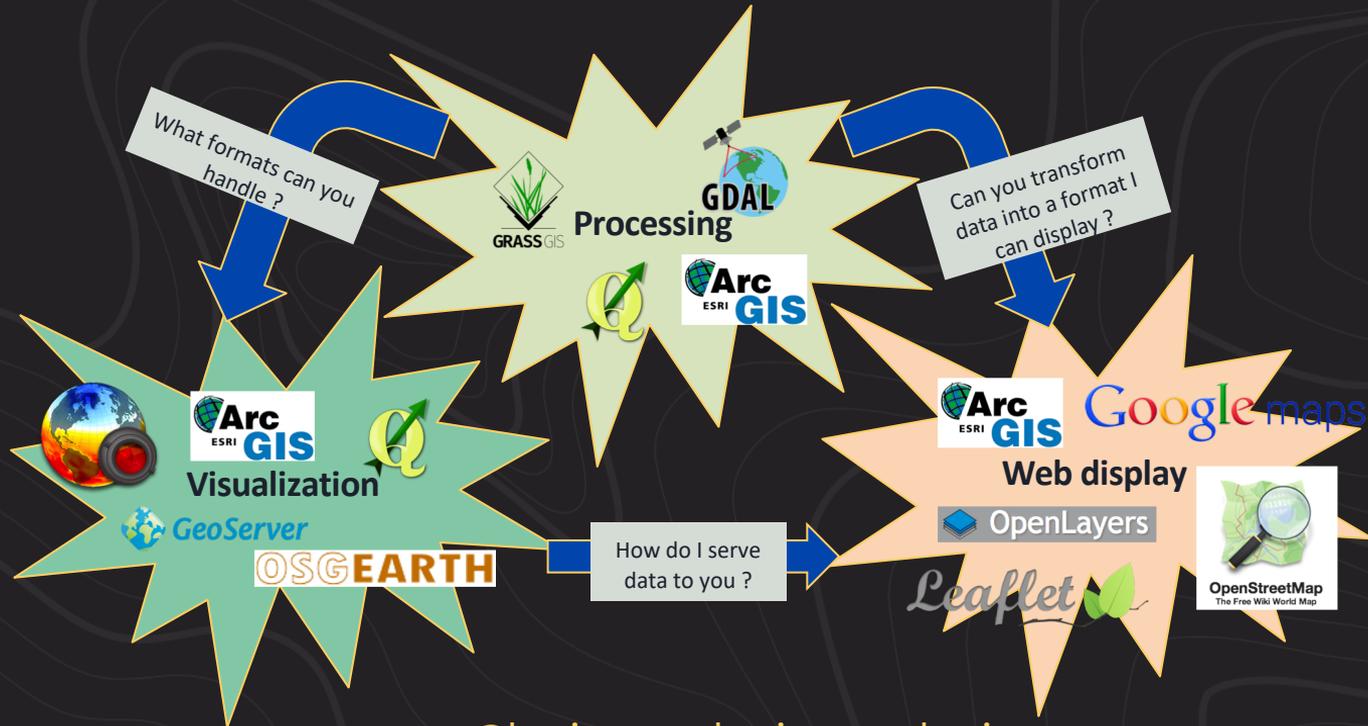
Remote sensing data

Data collected in the field



# Software stack for spatial data

It is definitely not trivial to deal with geospatial data  
(processing, displaying, exchange/sharing, etc)



Choices, choices, choices

# Take 1: GABBs -- Geospatial Data Building Blocks

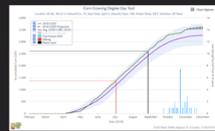
---



**G · A · B · B · S**

geospatial data analysis building blocks

# Geospatial Data Building Blocks (GABBs)



**Integrated** data management environment with **built-in** geospatial data support

Data visualization builders and tools that require **no programming**

**Publication** of data and tools (DOI)



Toolkits for rapid application development, **no GIS programming expertise** required

Data service **API**, interoperability

**Easy** to use and replicate

hubzero+



# Integrated Geospatial Data Platform

Preview and explore geospatial files right in the browser

Familiar file interface

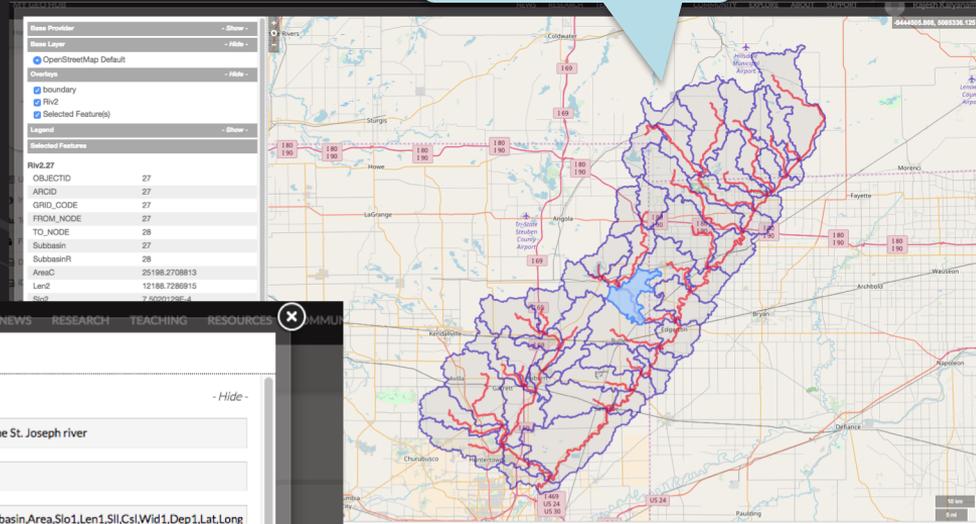
## Crop Yield Modeling Project

Connections » iData Storage

Upload

- Name
- Riv2.dbf
- Riv2.fix
- Riv2.prj
- Riv2.sbn
- Riv2.sbx
- Riv2.shp
- Riv2.shx
- avg100.csv

Storage backed by iRODS data management



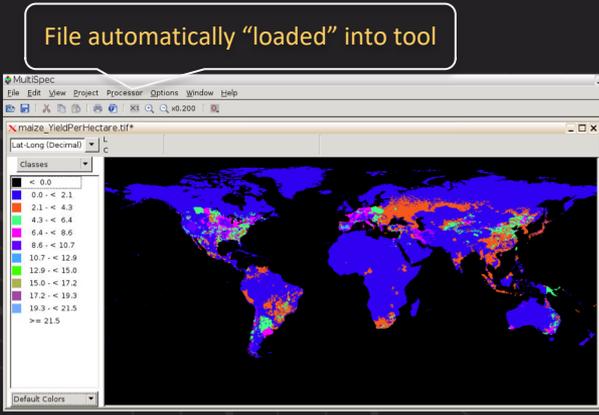
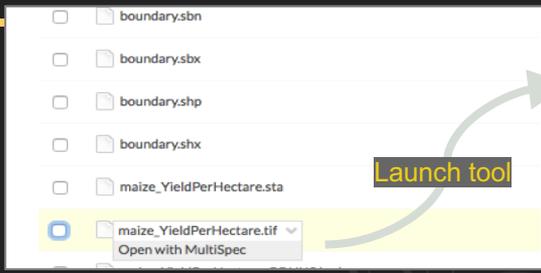
### Annotate file boundary.shp

Core Metadata

description	: watershed boundaries for the St. Joseph river
title	:
subject	: OBJECTID,GRIDCODE,Subbasin.Area,Slp1,Len1,Slp,Csl,Wid1,Dep1,Lat,Long
contributor	: rajkalya
publisher	: iData@myGeoHub
date	: 2017-Oct-20 02:15:42.191604
identifier	: boundary.shp
format	: ESRI Shapefile
type	: geospatial
source	:

Automatically extracted metadata on upload, user extensible

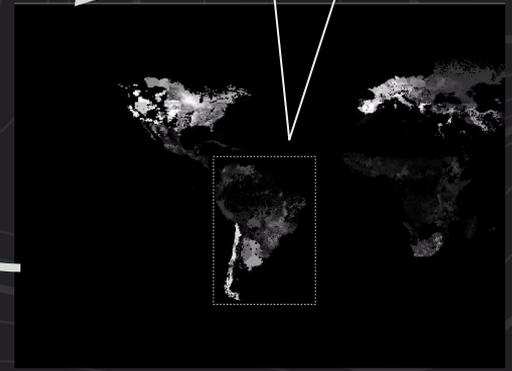
# Example workflow in one place



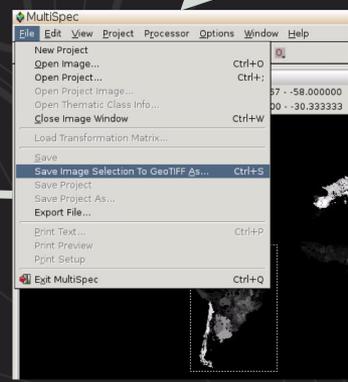
Do arbitrary processing of geospatial files (split into classes in this case)

Processing

Crop to region of interest



Save result

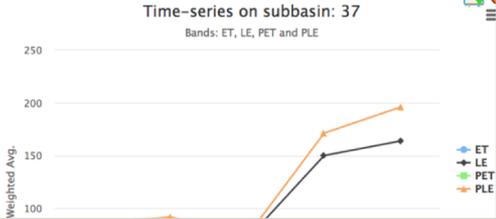
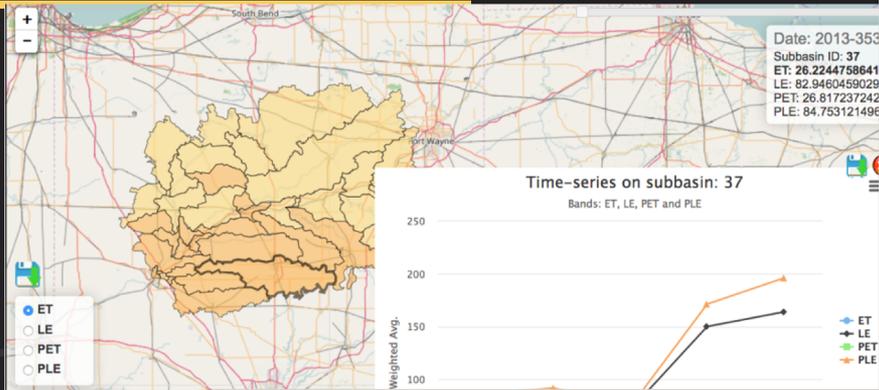


iData filespace visible to any tool



New file now accessible by others

# GABBs-enabled Tools



MultiSpec  
 File Edit View Project Processor Options Window Help

Text Output: LCB\_20130524\_Lafayette\_Area.tif (chs: 5, 4, 3)

Set Cluster Specifications  
 Algorithm:  Single Pass...  ISODATA...  
 Channel: All  
 Symbols: Default set  
 Cluster Stats: Do Not Save  
 Cluster Classification Map Area(s):  No classification map  Write Cluster Report/ Map To:

Set ISODATA Cluster Specifications  
 Initialization Options:  
 Along first eigenvector  
 Within eigenvector volume  
 Use single-pass clusters  
 Determine Clusters from:  
 Training Area(s)

Other Options:  
 Number Clusters: 10  
 Convergence (%): 99.0  
 Minimum Cluster size: 11

AgMIP Tool 1.2.4 @ GEOSHARE  
<http://www.pnas.org/content/111/9/3268-3273>

Download Aggregate Visualize

Crop Data  
 For definitions, descriptions, and limitations of these data please refer to Rosenzweig et al (2014) PNAS 111(9): 3268-3273.

Crop Model	GCM	RCP	SSP	Crop
<input checked="" type="radio"/> EPIC	<input checked="" type="radio"/> HadGEM2-ES	<input checked="" type="radio"/> Historical	<input checked="" type="radio"/> SSP2	<input checked="" type="radio"/> maize
<input type="radio"/> GEPIC	<input type="radio"/> IPSL-CM5A-LR	<input type="radio"/> RCP8.5	<input type="radio"/> CO2 fertilization	<input type="radio"/> soy
<input type="radio"/> pDSSAT	<input type="radio"/> MIROC-ESM-CHEM	<input type="radio"/> RCP6.0	<input type="radio"/> No CO2 fertilization	<input type="radio"/> wheat
<input type="radio"/> LPJmL	<input type="radio"/> GFDL-ESM2M	<input type="radio"/> RCP4.5	<input type="radio"/> IRR	<input type="radio"/> rice
<input type="radio"/> IMAGE-LEITAP	<input type="radio"/> NorESM1-M	<input type="radio"/> RCP2.6	<input type="radio"/> No Irrigation	<input type="radio"/> managed_grass
<input type="radio"/> PEGASUS			<input checked="" type="radio"/> Full Irrigation	<input type="radio"/> others
<input type="radio"/> LPJ-GUESS				<input type="radio"/> rapeseed

Fetch Data 0% Download Raw Data Suggested Citation

Log

Download Outputs

AgMIP Tool 1.2.4 @ GEOSHARE  
<http://www.pnas.org/content/111/9/3268-3273>

Download Aggregate Visualize

Crop Data  
 For definitions, descriptions, and limitations of these data please refer to Rosenzweig et al (2014) PNAS 111(9): 3268-3273.

Crop Model	GCM	RCP	SSP	Crop
<input checked="" type="radio"/> EPIC	<input checked="" type="radio"/> HadGEM2-ES	<input checked="" type="radio"/> Historical	<input checked="" type="radio"/> SSP2	<input checked="" type="radio"/> maize
<input type="radio"/> GEPIC	<input type="radio"/> IPSL-CM5A-LR	<input type="radio"/> RCP8.5	<input type="radio"/> CO2 fertilization	<input type="radio"/> soy
<input type="radio"/> pDSSAT	<input type="radio"/> MIROC-ESM-CHEM	<input type="radio"/> RCP6.0	<input type="radio"/> No CO2 fertilization	<input type="radio"/> wheat
<input type="radio"/> LPJmL	<input type="radio"/> GFDL-ESM2M	<input type="radio"/> RCP4.5	<input type="radio"/> IRR	<input type="radio"/> rice
<input type="radio"/> IMAGE-LEITAP	<input type="radio"/> NorESM1-M	<input type="radio"/> RCP2.6	<input type="radio"/> No Irrigation	<input type="radio"/> managed_grass
<input type="radio"/> PEGASUS			<input checked="" type="radio"/> Full Irrigation	<input type="radio"/> others
<input type="radio"/> LPJ-GUESS				<input type="radio"/> rapeseed

Fetch Data 0% Download Raw Data Suggested Citation

Log

AgMIP Tool 1.2.4 @ GEOSHARE

# Data Publication (with DOI)

Season-wise irrigated and rainfed crop areas for India around year 2005

By Gang Zhao<sup>1</sup>, Stefan Siebert<sup>1</sup>  
University of Bonn, Germany

Crop growing area and irrigated fraction for 21 crops in Kharif, Rabi and Zaid seasons for 2005 in 500 m spatial resolution.

Listed in [Datasets](#) | publication by group [Geoshare](#)

Download Bundle

Additional materials available

Version 1.0 - published on 15 Jan 2015  
doi:10.13019/M2CC71 - cite this

licensed under [CC0](#) - [Creative Commons](#)

Version 1.0 - published on 15 Jan 2015  
doi:10.13019/M2CC71 - cite this

0.0 RANKING

0 citation(s)

Share: [f](#) [t](#) [g+](#) [v](#) [p](#) [e](#) [m](#)

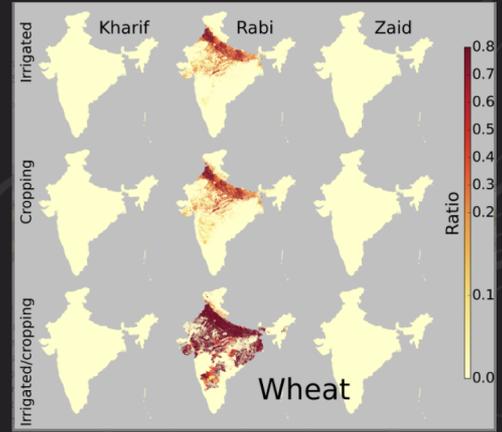
0 review(s) (Review this)

0 question(s) (Ask a question)

0 wish(es) (Add a new wish)

3398 total view(s), 719 download(s)

3398 total view(s), 719 download(s)

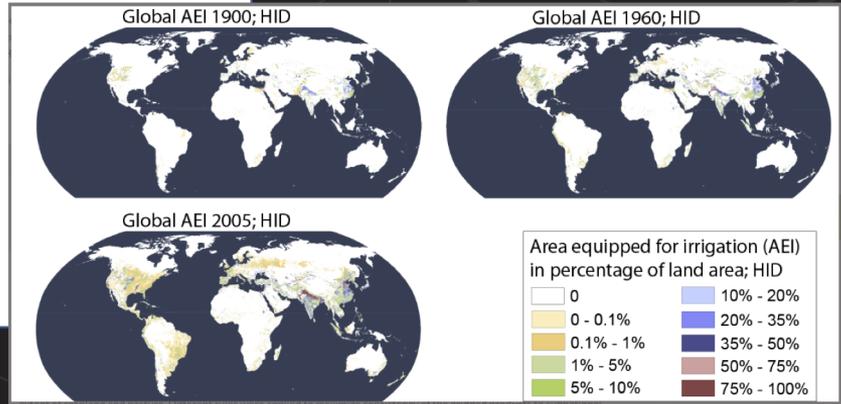


Gridded, spatial-temporal data for India

About Citations Versions Supporting Docs Reviews Questions Wishlist Usage

Description

The spatio-temporal pattern of cultivated crops varies considerably in India and is strongly determined by varying balances of water supply and demand across regions and seasons. The positive balance between precipitation and potential evaporation in the Kharif (monsoon) season lasting from June to October allows to grow one rainfed crop in most parts of the country, while there is a high water deficit in the Rabi season (November to February) and in the pre-monsoon season (Zaid, March to May). Therefore, irrigation is essential for crops in the Rabi and Zaid seasons to reduce drought stress, while it is only supplementary in the Kharif season across most of the country. This kind of knowledge and information are vital for natural resources management and climate change impact assessments. This study aims to map the rainfed and irrigated cropping area for 21 crops by integrating remotely sensed land use and land cover (LULC) data, crop calendars and survey based census data.



Another example: Global gridded, spatial-temporal data, 1900-2005

# The GeoHub Geospatial Science Gateway

mygeohub.org

myGeoHub

This hub supports the geospatial modeling, data analysis and visualization needs of the broad research and education communities through hosting of groups, datasets, tools, training materials, and educational contents. Sign up for a free account and start accessing the resources here. Please contact us if you are interested in hosting your group on mygeohub.org.

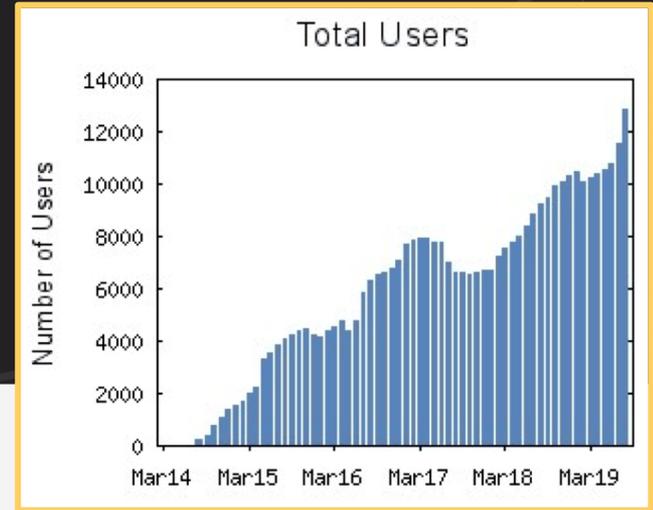
<p>Cyber Training for FAIR Science...</p> <p>Create a new generation of scientists capable of managing data-rich and computationally intensive tasks.</p> 	<p>MultiSpec</p> <p>MultiSpec is an easy to learn and use, image processing tool for interactively analyzing a broad spectrum of geospatial image data</p> 	<p>The US-China Food-Energy-Water Systems Hub</p> <p>Build a community to improve and extend a SIMPLE-based modeling and data synthesis framework to conduct multi-factorial assessment of tradeoffs needed to achieve sustainability within US and China settings</p> 	<p>Plant Phenotyping Sensor Research</p> <p>Develop innovative handheld plant phenotyping sensors and a scalable geospatial streaming data infrastructure for data management, analysis, and decision making</p> 
--	---	---	---

# The GeoHub Geospatial Science Gateway

More than 10,000 users each year

Hosting > 10 funded projects

A testbed for new GABBs capabilities



## Usage

Usage for prior 12 months



12861

Users



354

Resources



54

Tools



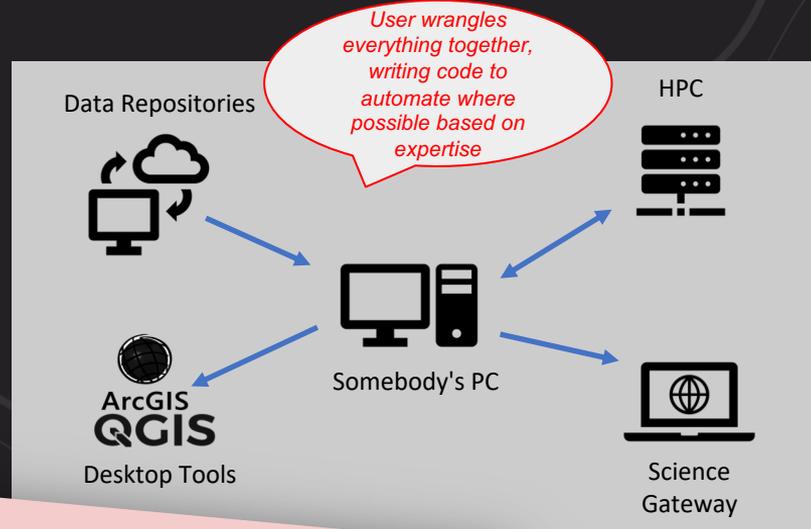
47022

Simulations

# Wrangling of data, computation, software, ...

## OUR DATA WORKFLOW - Ver. 1 2 3

1. Make sure date is just after 1st or 15th!
2. Go to: ~~usgs.gov~~ `prd-tnm.s3.amazonaws.com/index.html?prefix=StagedProducts`
3. Browse: Hydrography...NHDPlusHR...Beta...GDB.
4. Download NHDPLUS\_H\_01###\_HU4\_GDB.zip where ### is 02 to 14.
5. Unzip it - **WARNING: Have enough space!!!**
6. Run our tool. **WARNING: Takes a loooong time!!! DO NOT TURN OFF PC!!!**
7. Upload output files to cluster - ~~Note: wait until all successful!~~
8. Kick off our standard jobs.
9. ~~Occasionally check 'em.~~ Wait for email(s)???
10. Download new images.
11. Ask ~~Fred~~ to upload to website. **Mary?**
12. Tell everybody there's new stuff.



Process your files before running our tool

- get the latest code from **Nicole Jaewoo** for filtering the input data
- Note: that's Windows code - use the lab desktop
- you need to get the maps from the group folder  
`/depot/lyllegroup/project1/maps/exp1/`  
... for aggregation

# Data Challenges in Being FAIR

---

Even in the age of large computational resources, research is faced with:

- Manual workflows, or custom-made automated processing
- Coding data acquisition and processing requires diverse software knowledge
- Code requires frequent modification on data provider technology changes
- Cross-domain research requires domain-specific data format knowledge
- Workflows seldom reproducible

# Capturing complex data pipelines (example)

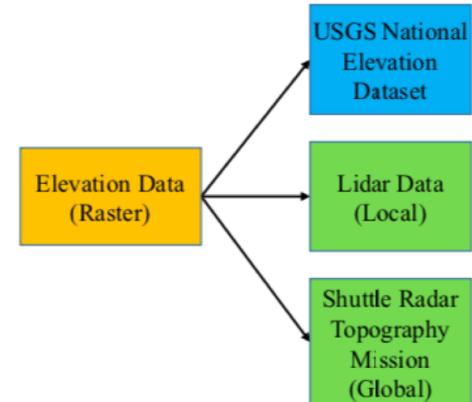
## Potential pitfalls with other elevation data sources

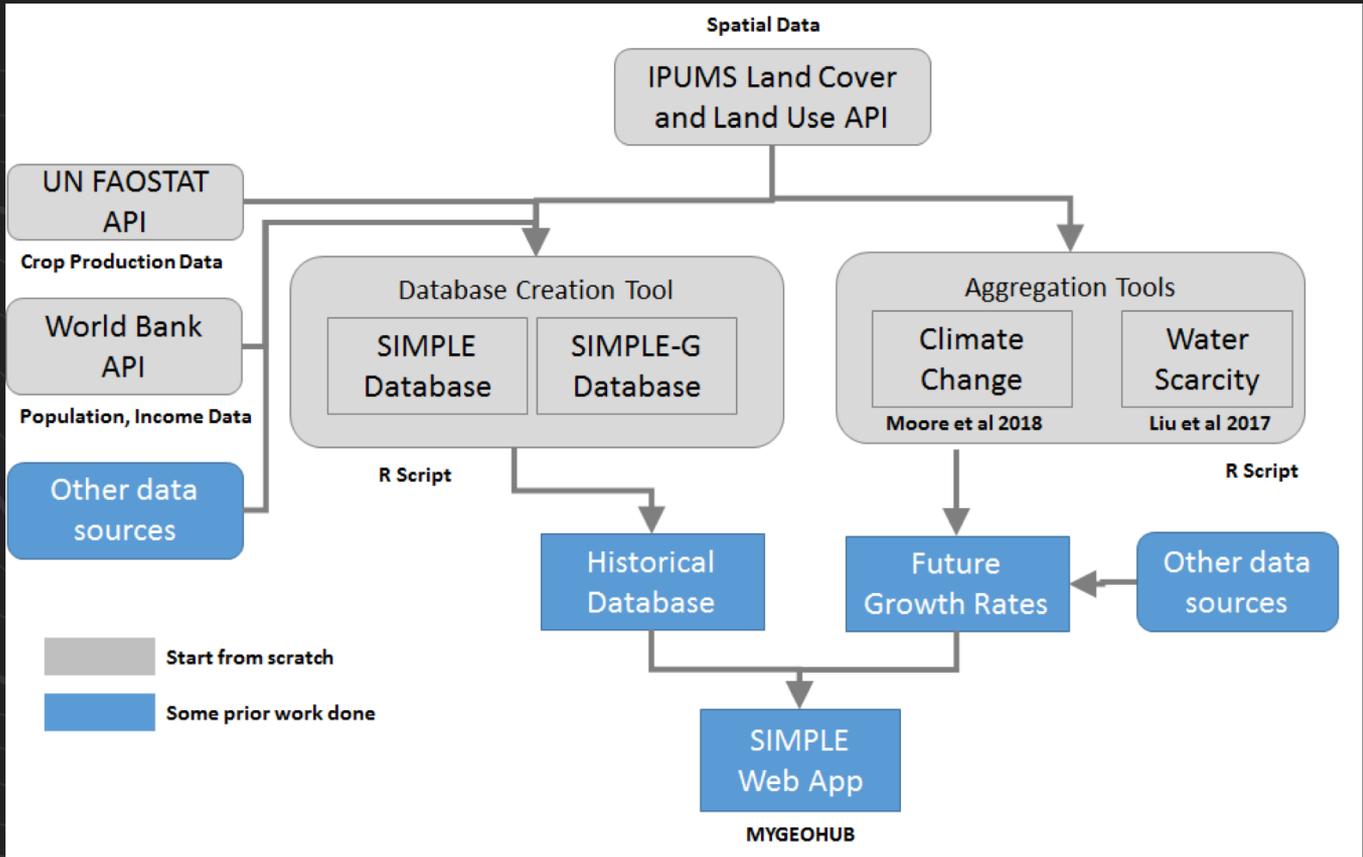
- 1/9<sup>th</sup> arc-second or LiDAR or 1-m DEM not available nationally
- Pre-processed Lidar available in some states (IN, OH, MN, NC)
- Create Lidar acquisition tool where available or not?
- Conversion of Lidar point cloud to bare-earth DEM is an issue
- SRTM → available globally at different resolutions
- Probably need to re-project user shapefile to USGS coordinates first



1-m DEM availability

Lidar Point cloud availability





# Example Workflow



# GABBs 2.0: GeoEDF -- Vision

---

Create an **extensible geospatial data framework** that will address the challenges by providing **seamless connections** among platforms, data and tools, hence making valuable, large scientific and social datasets **usable directly** in scientific models and tools.

The ultimate goal is to put **easy-to-use tools and platforms** into the hands of researchers and students to conduct scientific investigations following **FAIR science principles**.

**FAIR = Findable, Accessible, Interoperable, Reusable**

# Vision: After GeoEDF

## OUR DATA WORKFLOW - Final

1. Go to the science gateway
2. Define "my\_workflow.yml" (or use tool GUI if needed)
3. Ask GeoEDF to execute!
4. Data and workflow automatically published to science gateway

GeoEDF abstracts away complexities of data access, transfer, and HPC execution; user only need define a logical workflow

Data Repositories



HPC

GeoEDF Framework

Web tools



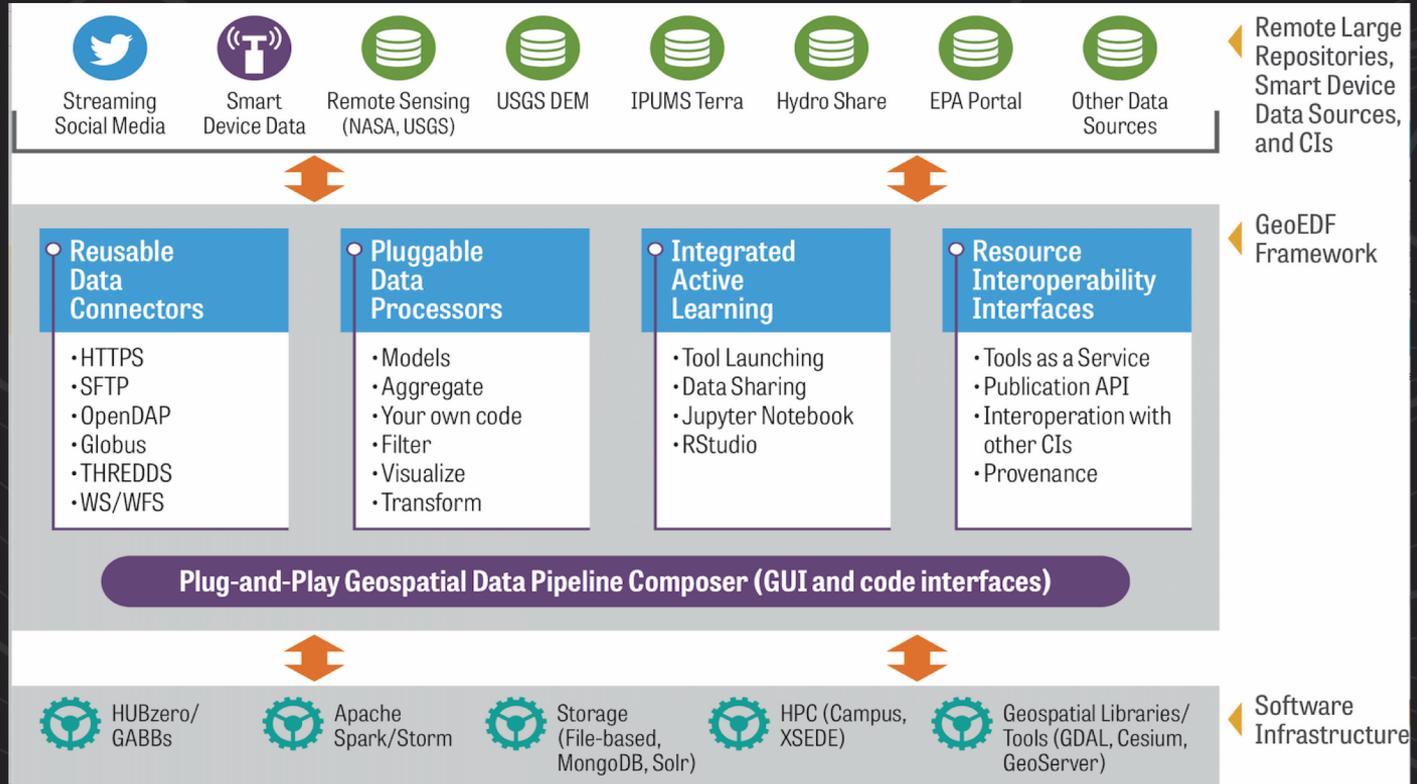
Science Gateway



User's PC

- Automated, secure, logged process running on dedicated infrastructure - You can log off!
- Leverage building blocks from existing workflows
- Data transfer and HPC execution abstracted away
- Automatic provenance capture and data annotation for future discoverability, reproducibility

# GeoEDF High-Level View



# Deliverables

---

## CI

Plug-and-play geospatial data framework, open-source packages installable on CI platforms

## Science

Scientific workflows composed of reusable building blocks



## FAIR

Enhanced HUBzero publication, course, tool/data linkage for FAIR science

## Outreach

Training materials, CI interoperability, workshops, support

# Project Timeline

## Requirements & Design

Evolve Science Requirements  
Design system  
Prototype software  
Collect feedback

Year 1

Year 2

Incorporate feedback  
Integrate science use cases  
Develop connectors and processors  
Develop plug and play framework

## Implementation & Alpha Release

## Beta Release & Interoperability

Develop reusable science workflows  
Improve interoperability  
Extend to other Hubs

Year 3

Year 4

Enhance GeoEDF  
Integrate new science use cases  
Improve interoperability  
Share research findings  
Training workshops

## Release V1 & Outreach

## Release V2 & Outreach

Disseminate results  
Support adoption  
Packaging and release  
Training Workshops

Year 5



# Design and Cyberinfrastructure

# Workflow Example I

Select an Earth Observation product type

- ✓ MODIS-ET/PET/LE/PLE
- MODIS-LAI/FPAR
- SMAP
- AMSR-E
- GPM
- NLDAS

...a (maximum of 1 year)

...01/05/2014

Enter name for this data request ⓘ

**Mask with shapefile, compute weighted aggregate for each polygon**

# Workflow Example I - Opportunity

Select an Earth Observation product type

✓ MODIS-ET/PET/LE/PLE

MODIS-LAI/FPAR

SMAP

AMSR-E

GPM

NLDAS

Enter name for this data request 

[http://files.ntsg.umt.edu/data/NTSG\\_Products/MOD16/MOD16A2.105\\_MERRAGMAO/Y2001/D001/MOD16A2.A2000001.h00v08.105.2013121200130.hdf](http://files.ntsg.umt.edu/data/NTSG_Products/MOD16/MOD16A2.105_MERRAGMAO/Y2001/D001/MOD16A2.A2000001.h00v08.105.2013121200130.hdf)

<https://e4ftl01.cr.usgs.gov/MOTA/MCD15A3H.006/2002.03.19/MCD15A3H.A2002193.h07v07.006.2015149100709.hdf>

[https://n5eil01u.ecs.nsidc.org/SMAP/SPL4SMGP.003/2015.03.31/SMAP\\_L4\\_SM\\_gph\\_20150331T013000\\_Vv4030\\_001.h5](https://n5eil01u.ecs.nsidc.org/SMAP/SPL4SMGP.003/2015.03.31/SMAP_L4_SM_gph_20150331T013000_Vv4030_001.h5)

# Workflow Example I - Opportunity

---

Year

Day of  
year

[http://files.ntsg.umt.edu/data/NTSG\\_Products/MOD16/MOD16A2.105\\_MERRAGMAO/Y2001/D001/MOD16A2.A2000001.h00v08.105.2013121200130.hdf](http://files.ntsg.umt.edu/data/NTSG_Products/MOD16/MOD16A2.105_MERRAGMAO/Y2001/D001/MOD16A2.A2000001.h00v08.105.2013121200130.hdf)

MODIS  
grid

# Workflow Example II

```
def GetNED(NL, WL):
    name1 = "n"+NL+"w"+WL
    address = "ftp://rockyftp.cr.usgs.gov/vdelivery/Datasets/Staged/Elevation/1/ArcGrid/USGS_NED_1_"
    url_final = address + name1 + "_ArcGrid.zip"
    print(url_final)
    work_folder_name = os.path.join(input_folder_name, "WorkFolder")
    if os.path.exists(work_folder_name) == False:
        os.mkdir(work_folder_name)
    boundary_path = os.path.join(input_folder_name, boundary_file)
    input_crs = QgsVectorLayer(boundary_path, '', 'ogr' ).crs().authid()
    #processing.run('qgis:reprojectlayer',{'INPUT': full_input_path, 'TARGET_CRS':'EPSG:10267'})
    processing.run('native:reprojectlayer',{'INPUT': boundary_path, 'TARGET_CRS':'EPSG:4326',
    input_list = [os.path.join(work_folder_name, cur_raster) for cur_raster in raster_names]
    print("Merging Raster...")
    processing.run("gdal:merge", {'INPUT':input_list, 'OUTPUT':work_folder_name + "/merged_rast.tif"})
    print("Projecting Raster...")
    processing.run('gdal:warp', {'INPUT': work_folder_name + "/merged_rast.tif", 'TARGET_CRS':
    print("Clipping Raster...")
    processing.run('gdal:cliprasterbymasklayer',{'INPUT': work_folder_name + "/proj_rast.tif", 'MASK':
    print("DEM prepared successfully!!!")
```

# Workflow Example II - Opportunity

Get NED from USGS

```
def GetNED(NL, WL):
    name1 = "n"+NL+"w"+WL
    address = "ftp://rockyftp.cr.usgs.gov/vdelivery/Datasets/Staged/Elevation/1/ArcGrid/USGS_NED_1_"
    url_final = address + name1 + "_ArcGrid.zip"
    print(url_final)
```

Reproject watershed  
shapefile

```
work_folder_name = os.path.join(input_folder_name, "WorkFolder")
if os.path.exists(work_folder_name) == False:
    os.mkdir(work_folder_name)
boundary_path = os.path.join(input_folder_name, boundary_file)
input_crs = QgsVectorLayer(boundary_path, '', 'ogr' ).crs().authid()
#processing.run('qgis:reprojectlayer',{'INPUT': full_input_path, 'TARGET_CRS':'EPSG:10267'})
processing.run('native:reprojectlayer',{'INPUT': boundary_path, 'TARGET_CRS':'EPSG:4326',
```

Mosaic -> reproject ->  
clip raster(s)

```
input_list = [os.path.join(work_folder_name, cur_raster) for cur_raster in raster_
print("Merging Raster...")
processing.run("gdal:merge", {'INPUT':input_list, 'OUTPUT':work_folder_name + "/merged_rast.tif"})
print("Projecting Raster...")
processing.run('gdal:warp', {'INPUT': work_folder_name + "/merged_rast.tif", 'TARGET_CRS':
print("Clipping Raster...")
processing.run('gdal:cliprasterbymasklayer',{'INPUT': work_folder_name + "/proj_rast.tif", 'MASK':
print("DEM prepared successfully!!!")
```

# Data Connectors

---

## ❖ What are they?

- Help retrieve remote data (NASA, USGS, field sensors, etc.) and make available in scientific workflows
- Abstract away specifics of implementation
- Data sources, sinks in a workflow

### Reusable Data Connectors

- HTTPS
- SFTP
- OpenDAP
- Globus
- THREDDS
- WS/WFS

# Remote Data Sources (under consideration)

NASA	MODIS, SMAP, other Earthdata DAACs
USGS	Elevation, land use, hydrography, Gage, NLDI
USDA	Soil, land cover, land use
CUASHI	Rainfall, Hydroshare resources
EarthStat	Crop data
FAO	Arable land, harvest data
CIESIN	Population data
EPA	Water quality
Others (no API yet)	Open Data Cubes, Google Earth Engine, ESS-Dive

# Data Processors

---

## ❖ What are they?

- Data transformers that can be plugged into a workflow
- Range from simple geospatial data transformation to scientific simulation models
- Pre and post processing

### Pluggable Data Processor

- Models
- Aggregate
- Your own code
- Filter
- Visualize
- Transform

# Processing Operations (under consideration)

---

Domain Independent	Reproject, resample, format transformation, filter, mosaic, clip/mask, aggregate (spatial & temporal), visualization, reclassification
Hydrology	Terrain analysis, flood models
Digital Ag	Query, spatial/temporal filter, ML training, decision support
Sustainability	Downsample, (weighted) aggregate, FEWS models

# Plug-and-play Workflow Composer

---

Plug-and-Play Geospatial Data Pipeline Composer (GUI and code interfaces)

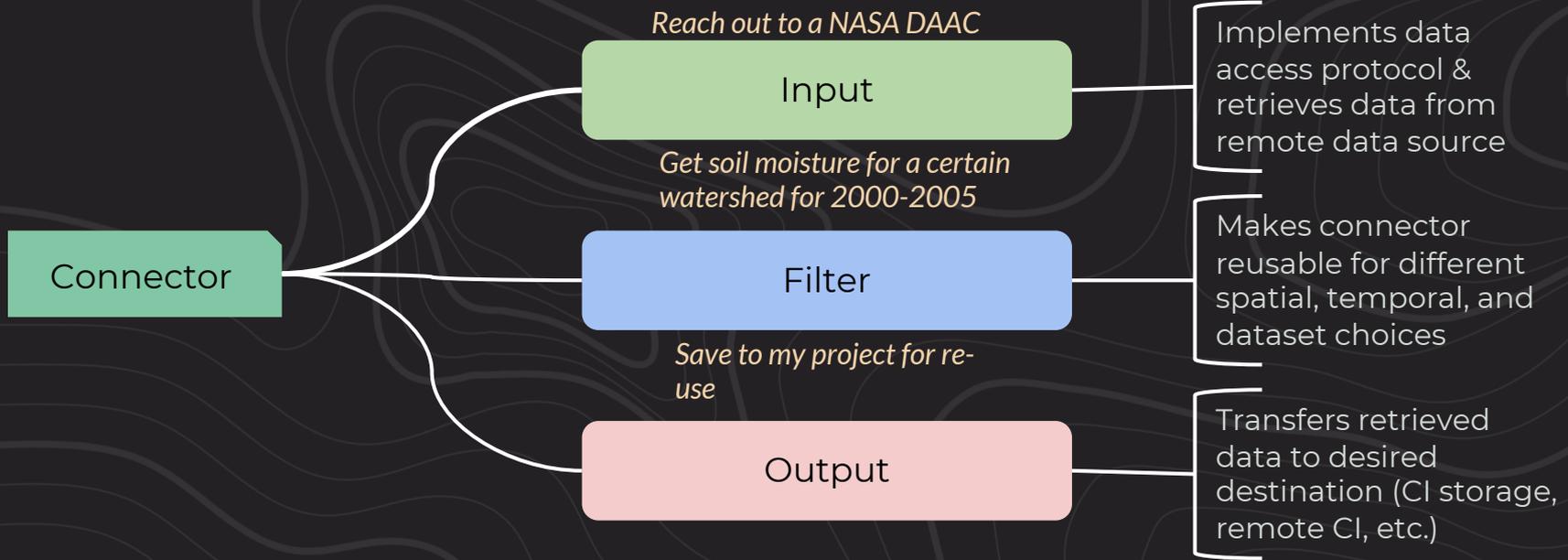
## ❖ What is this?

- Framework for composing data connectors and processors into scientific workflows
- Transforms abstract workflow into actual workflow executing on heterogeneous compute

01

# Implementation

# Data Connectors - Design



# Data Connectors - Example Definition

Python  
Input class

Variable to  
be bound  
by filter

Input:

NASAInput:  
url:

`http://files.ntsg.umd.edu/data/NTSG_Products/MOD16/MOD16A2.105_MERRAGMAO/{file}`  
user: `rkalyana`  
password:

Filter:  
file:

Filter returns  
string bindings  
for variable

PathFilter:

pattern: `'Y{year}/D001/*.h00v08*.hdf'`

year:

DateTimeFilter:

pattern: `'%Y'`  
start: `01/01/2000`  
end: `12/31/2005`  
period: `1Y`

Filter params  
can also have  
variables

# Data Processors - Example Definition

---

Python processor  
class implementing  
masking operation

```
HDFShapefileEOSMask:
```

```
  hdffile: /data/workflow263/mod16Y2001D1T1200.h00v08.hdf
```

```
  shapefile: /home/rkalyana/subs1.shp
```

Processor specific  
params; validated  
during instantiation

# GeoEDF Workflow - Example Definition

```
$1:
  Input:
    NASAInput:
      url:
        http://files.ntsg.umd.edu/data/NTSG_Products/MOD16/MOD16A2.105_MERRAGMAO/{file}
      user: rkalyana
      password:

  Filter:
    file:
      PathFilter:
        pattern: 'Y{year}/D001/*.h00v08*.hdf'
      year:
        DateTimeFilter:
          pattern: '%Y'
          start: 01/01/2000
          end: 12/31/2005
          period: 1Y

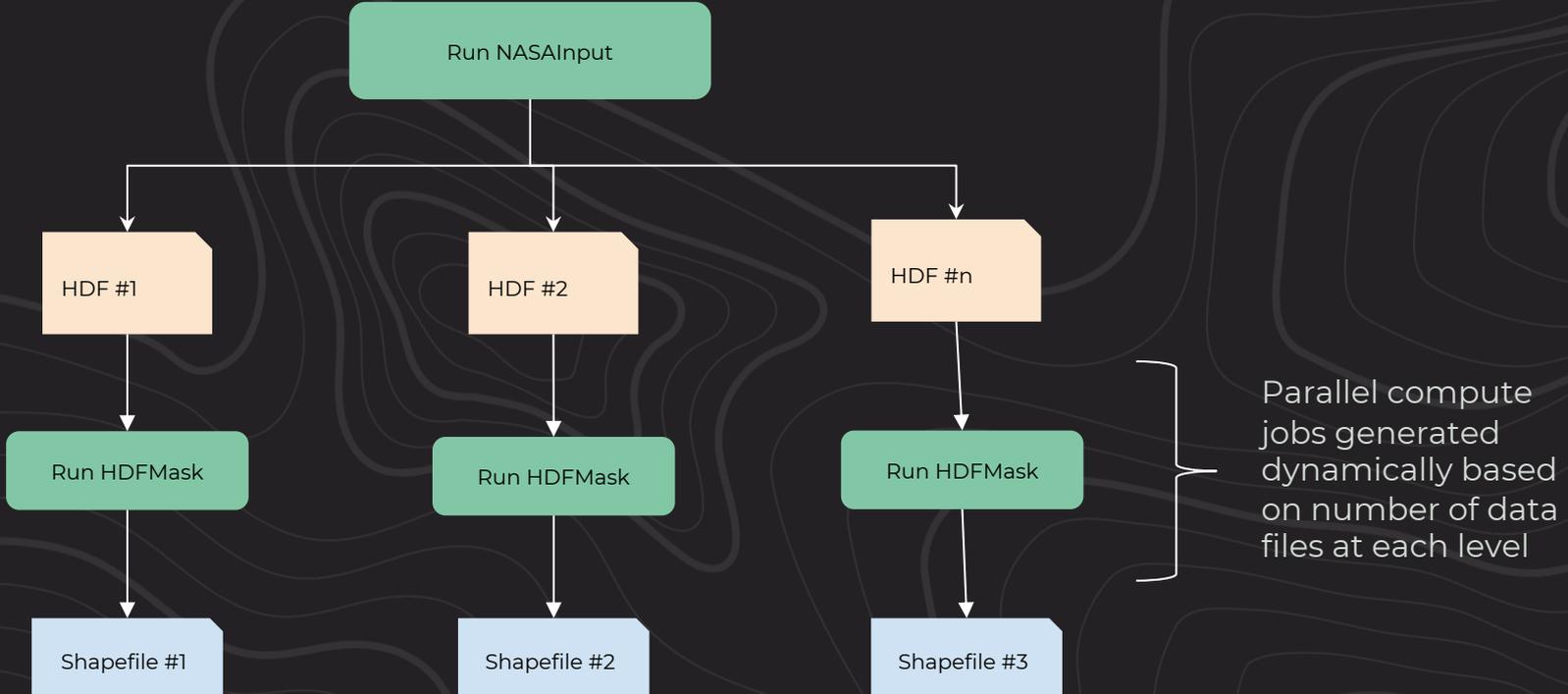
$2:
  HDFShapefileEOSMask:
    hdfFile: $1
    shapefile: /home/rkalyana/subs1.shp
```

Workflow stage

Reference output of prior stage

- ★ NASAInput can be used to access any Earthdata-associated repository
- ★ HDFShapefileEOSMask can be applied to any HDF4 or HDF5 file

# Actual Scientific Workflow



# Putting It All Together

---

## Workflow Definitions

Users pick and choose different connector & processor classes to define a workflow

(as YAML file/via GUI/through API)

2



## Workflow Execution

Workflow engine transforms declarative specification into concrete Pegasus scientific workflow and executes on heterogeneous compute

3

## GeoEDF Building Blocks

Users contribute various connector (Input, Filter, Output), and processor classes

1

# Interoperability & FAIR Science

---

## ❖ How do we do FAIR?

- Data publications can be searched using their content metadata, accessed via APIs & used in workflows
- Automatically track metadata, provenance in workflows
- Launch tools, workflows seamlessly from a remote CI (with remote data inputs)

### Resource Interoperability Interfaces

- Tools as a Service
- Publication API
- Interoperation with other CIs
- Provenance

# Collaboration and Interoperability

# Broader Impacts - Community

---

Usable by especially **interdisciplinary** research domains

- a. Critically important to research supporting the SDGs
- b. Data synthesis, multi-scale analysis
- c. Lower technology barrier (e.g., seamlessness, extensibility)

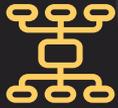
Help **domain** researchers meet new **FAIR** data/software requirements (e.g., journal, funding agencies)

Help domain science with broader **dissemination** (e.g., decision makers, public)

Future workforce **training** (learn and start with good practices)

# Opportunities

---



Interoperate  
with other  
CIs



Users  
contribute  
connectors  
and  
processors



Engage data  
producers to  
expose data  
through  
connectors

# Questions





# Questions?



## **Data to Action Webinar: Increasing the Use and Value of Earth Science Data and Information**

October 25th, 2019 | 1:00 pm ET





**Putting Data to Work: Building Public-Private  
Partnerships to Increase Resilience & Enhance the  
Socioeconomic Value of Data**  
**[2020esipwintermeeting.sched.com](https://2020esipwintermeeting.sched.com)**

**2020 Winter Meeting**

January 7-9, 2020

Bethesda North Marriott, Bethesda, MD



# Engagement Ops.



## DISCOVER

Find people and tools to make your data findable, accessible, interoperable, and reusable.



## COLLABORATE

Join-in or create a new collaboration area around your Earth science data challenges.



## INNOVATE

Utilize small-grant funding to build or expand Earth data technologies.



## NETWORK

Extend your network. Build connections across federal agencies, the private sector, and academia.

## JOIN

Encourage your organization to join ESIP's 110+ member organizations. Unlock membership benefits: start new collaborations, apply for funding, and more.

Stay up-to-date on all things ESIP by signing up to receive Monday Updates: <http://eepurl.com/rJQYn>.

---

# Thank you!

## Upcoming Webinars

- November 13th at 4 pm ET:
  - Data for our Planet: Increasing the Use and Value of Global Information Infrastructures to Support Resilient Cities, Disaster Risk Reduction and Infectious Diseases (Lesley Wyborn & Erin Robinson)
- Check the webinar homepage: <https://www.esipfed.org/webinars>.
- Webinar recordings are shared on the ESIP YouTube Channel.