The need for collaborative research - NEX
Massive data from multiple missions

Scalable and diverse computing architectures

Challenges:
• Data Volume
• Data Access
• Scalable Processing and Analytics
• Scalable Engagement
• Most data are not “analysis-ready” and processes are not well established
Need for an Earth Science Collaborative

- **Earth science at NASA is a community effort** (179 institutions in the US published 5 or more papers (2000-2012) using MODIS data alone).
- 100s of investigators spend a majority of their time dealing with data.
- Redundant storage and processing facilities result in larger overall computing budgets.
- Moving data sets that are getting larger each year is expensive & time-consuming.
- Sharing knowledge (codes, intermediate results, workflows) is difficult.
VISION
To provide “science as a service” to the Earth science community addressing global environmental challenges

GOAL
To improve efficiency and expand the scope of NASA Earth science technology, research and applications programs

+ NEX is virtual collaborative that brings scientists together in a knowledge-based social network and provides the necessary tools, computing power, and access to big data to accelerate research, innovation and provide transparency.
NEX Provides a Complete Work Environment - “Science As A Service”

**COLLABORATION**
Over 400 Members

**CENTRALIZED DATA REPOSITORY**
Over 2.3 PB of Data

**COMPUTING**
Scalable
Diverse
Secure/Reliable

**KNOWLEDGE**
- Workflows
- Model codes
- Re-useable software
- Machine Images
NEX Collaboration Tools

Network and Share

Outreach

Collaborate

Visualize
Secure enclave

NEX Architecture at NASA Supercomputing Center

NASA Users

NEX Sandbox

Limited computing, NEX science data access (1.9PB), local storage (300TB), NEX Science Platform

Non-NASA Users

Authentication

NEX Core Science Datasets and storage

Read-only data access

MODIS, Sentinel, Landsat, NAIP, Climate,... 2.5 PB

Read-write data access

NEX Computing

NASA Pleiades Supercomputer 200,000+ CPUs
**Portal**
- Web server
- Database server
- 503 registered members

**Sandbox**
- 96-core server, 264GB memory, with 320 TB storage
- 48-core server, 128 GB, 163 TB storage

**HPC**
- 720-core dedicated queue + access to rest of Pleiades
- 181 users/44 active
- 2.3 PB storage

**Models/Tools/Workflows used by NEX User Community**
- GEOS-5
- CESM
- WRF
- RegCM
- VIC
- BGC
- LPJ
- TOPS
- BEAMS
- Fmask
- LEDAPS
- METRIC
- ...

**Data (>2 PB)**
- Landsat
- MODIS
- TRMM
- GRACE
- ICESAT
- CMIP5
- NCEP
- MERRA
- NARR
- PRISM
- DAYMET
- NAIP
- Digital Globe
- NEX-DCP30
- NEX-GDDP
- WELD
- NAFD-NEX
ACCESS TO NEX

COMPONENT ARCHITECTURE

**EARTH SCIENCE COMMUNITY USERS**
- Access to Portal

**NEX SCIENCE USERS**
- Access to compute resources

**NEX HPC USERS**
- Access to NAS supercomputing resources

**Portal**
- Point of entry to NEX collaborative environment
  - Project information
  - Collaboration and Social Networking
  - Document Publication
  - Resource Requests
  - Data Discovery

**Sandbox**
- Virtualized NEX compute environment
  - Workflow Management
  - Provenance
  - Rich semantic search
  - Data/Model/Tool access (API)

**Domain Platform**
- Virtualization Support
- Model and Analytic Tool Execution

**Infrastructure Platform**
- Data Management
  - Data acquisition and pre-processing
  - Data storage

**NAS HPC**
- Environment for Computing at scale
  - Execution of Jobs migrated from Sandbox
  - Storage of results in NEX Data Management environment

*Runs on NAS web servers*
*Runs on NAS supercomputers and storage*
*Runs on dedicated NEX servers and storage*
OpenNEX – A private public partnership

NEX

- Restricted access
- NASA funded
- Focus on large-scale modeling and analysis
- Ideal for producing research quality/reproducible results

OpenNEX

- Builds on NEX
- Open access
- User-funded
- Ideal for prototyping, exploration and community engagement
Classes on NEX Big Data Projects

- Fully distributed data processing with no inter-process dependencies
  - Data sizes: 100TB – 25PB

- Machine-learning and data-mining with some inter-process data dependencies
  - Data sizes: 50TB – 2PB

- Analytics and science applications
  - Database query systems: 1 – 20TB

- Provenance and other graph queries
  - 100 million up to 1 billion triples in 2017

- Climate and ecosystem modeling
  - Computationally intensive, time and space data dependent processing
  - 2 – 20TB
Science at NEX

Global Vegetation Biomass at 100m resolution by blending data from different satellites

High resolution climate projections for climate impact studies

High resolution monthly global data for monitoring forests, crops and water resources

Mapping fallow area in California during drought

Machine learning and Data mining – moving towards more data-driven approaches
Regional multi-sensor analysis and anomaly detection

Monitoring global drought July 2012 MODIS 1km NDVI anomaly

Landsat Processing WELD, GIBS, NAFD, Landsat 8

Climate downscaling (NEX-DCP30)

TRMM Precipitation Anomaly, 2012

NEX enables National Climate Assessment
Global Landsat processing

For the first time in Landsat history of nearly 40 years, we can now process and create quantitative information about changes in global landscapes, in a matter of DAYS.

Near real-time *quantitative* updates of information about global landscapes

*(Global WELD project with D. Roy @ South Dakota State University)*

For monitoring crop growth, deforestation, and impacts of natural disasters

Data Volumes: 1800TB

For the first time in Landsat history of nearly 40 years, we can now process and create quantitative information about changes in global landscapes, in a matter of DAYS.
High Resolution Tree Cover Classification

**NAIP Processing Architecture**

- **INPUT IMAGE**
- **HIGH GRAIN PARALLELISM**
  - Each image fed in parallel to a separate core in HPC
- **FINE GRAIN PARALLELISM**
  - Each input image sliced and fed in parallel to HPC modules
- **HPC Module 1**
- **HPC Module 2**
- **HPC Module 3**
- **NEX**
- **NEX HPC**
- **SEGMENTATION** (using statistical region merging)
  - HPC M3
- **CLASSIFICATION** (Deep Belief Network)
  - HPC M2
- **VOTING**
- **OUTPUT IMAGE**
- **UPDATE TRAINING DATASET**

**NASA Earth Exchange High Performance Computing (HPC)**

**NASA Earth Exchange Storage**
California Tree Cover Mosaic

San Francisco Bay Area
Open NASA Earth Exchange (OpenNEX)
Virtual Workshop & Challenge 2014

July 1st - November 15th, 2014

President Obama has announced a series of executive actions to reduce carbon pollution and promote sound science to understand and manage climate impacts for the U.S.

Following the President’s call for developing tools for climate resilience, NEX is hosting a workshop that will feature:
1) Climate science through lectures by experts
2) Computational tools through virtual labs, and
3) A challenge inviting participants to compete for prizes by designing and implementing solutions for climate resilience.

For more information on virtual labs and lectures, visit
https://nex.nasa.gov/opennex/
NEX Summary

- Lowers the barrier of locating data, reuse models, reuse codes, and compute resources.
  - Why create your own if a similar analytics was done before?
- Allows knowledge sharing.
- Provides a framework for reproducible/verifiable results.
- Platform for prototyping or extending applications.
- Enabling broader community access through public-private partnerships.
HPCwire
2014 Readers Choice Awards

Best Data-Intensive System
(End User focused)

NASA Earth Exchange (NEX) Platform supports dozens of data-intensive projects in Earth sciences
(a) Linear mixing, (b) nonlinear mixing.
**Linear unmixing model**

\[
y(x, y) = \sum_{n=1}^{N} e_n \alpha_n (x, y) + \eta
\]

\[
y = E\alpha + \eta
\]

\(\alpha_n (x, y)\) is a scalar value representing the functional coverage of endmember vector \(e_n\) at pixel \(y(x, y)\).

**Constraints:**

1.) \(\alpha_n \geq 0, \ \forall n: 1 \leq n \leq N\)  
Abundance nonnegativity constraint (ANC)

2.) \(\sum_{n=1}^{N} \alpha_n = 1\)  
Abundance sum-to-one constraint (ASC)
Objective

1. To perform a comparative analysis of the
   • least squares
   • sparse regression
   • signal-subspace
   • geometrical and
   • statistical methods
   for subpixel classification on different data sets.

2. To develop a method to utilize the subpixel maps in obtaining fractional land cover classes.
Computer simulated data – (a) band 3 and (b) band 4 ($\sigma^2 = 32$).
Global mixing spaces were sampled* using a spectrally diverse land cover and diversity of biomes with 100 Landsat ETM+ scenes.

This defined a standardized spectral endmember of substrate \((S)\), vegetation \((V)\), and dark objects \((D)\).

- **Substrate** - soils, sediments, rocks, and non-photosynthetic vegetation.
- **Vegetation** - green photosynthetic plants.
- **Dark objects** - clear waters, deep shadows, absorptive substrate materials, etc.,

The SVD endmember coefficient, in addition to dates and locations of each subscene are available at [http://www.LDEO.columbia.edu/~small/GlobalLandsat/](http://www.LDEO.columbia.edu/~small/GlobalLandsat/).

Global endmembers
Validation

- Range - minimum and maximum abundance values
- correlation coefficient (cc)
- RMSE
- signal-to-reconstruction error (SRE)
- probability of success \(p_s\) and
- bivariate distribution function (BDF)
- Error matrix (TP, TN, FP, FN)
- Producer’s, User’s, Overall accuracy, and Kappa coefficient

were used for validation.
Results

CC and RMSE for endmember 1, 2 and 3.

Plot of probability of success ($p_s$) and SRE(dB).
Global WELD (~0.35 million scenes)

FCLS

Global endmembers

Abundance maps

NPP-VIIRS

Substrate

Vegetation

Dark objects

Thresholding

Urban impervious surface

RF classifier

Forest, farmland, water signature

NLCD

Urban built-up map

Forest cover map

Farmland map

Water bodies map

NASA Earth Exchange (NEX) Storage

NAS Pleiades

NEX High Performance Computing (HPC)
Global WELD classification

- Mosaic of 8003 scenes generated from the 2011 annual WELD.
- Each scene is composed of 5295 rows and 5295 columns, therefore, a single snapshot global data consists of 224.3 billion pixels with 6 spectral dimensions, which was processed in 29 minutes.
LC classification from Abundance Maps

Figure.
(a) Forest fractional map
(b) farmland / grassland fractional map for the state of California.
(c) FCC from Landsat (bands 5-4-3 as RGB) showing forest patch, grassland and water bodies;
(d) forest fractional map for the corresponding area in (c);
(e) grassland / farmland fractional map;
(f) water fractional map,
(g) classification of original WELD data by RF; and
(h) forest cover map from NAFD product.
(a) Spatial distribution of the fractional water bodies in California

(b) RGB composite of the Landsat bands (5, 4, 3) showing San Luis Reservoir in central California

(c) fractional map of San Luis reservoir

(d) classification of original WELD data by RF classifier and

(e) San Luis reservoir from NAFD product.
Urban area classification

San Francisco Bay Area

Los Angeles Area
The builtup urban areas extracted from substrate fractional map and NPP-VIIRS data of SF Bay Area and LA area.

Note:
(a), (b) are Landsat RGB composite (bands 5, 4, and 3 as RGB); 
(c), (d) are NPP-VIIRS data; 
(e), (f) are the extracted urban areas from the substrate fractional cover and NPP-VIIRS data; 
(g), (h) are the urban settlement obtained from classification of original WELD data by RF classifier.
Validation

Confusion matrix

<table>
<thead>
<tr>
<th></th>
<th>Forest</th>
<th>Farmland</th>
<th>Water</th>
<th>Urban (SF)</th>
<th>Urban (LA)</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total samples</td>
<td>200000</td>
<td>200000</td>
<td>20000</td>
<td>100000</td>
<td>100000</td>
<td>800000</td>
</tr>
<tr>
<td>Class samples</td>
<td>100000</td>
<td>100000</td>
<td>10000</td>
<td>50000</td>
<td>50000</td>
<td>400000</td>
</tr>
<tr>
<td>Non-class samples</td>
<td>100000</td>
<td>100000</td>
<td>10000</td>
<td>50000</td>
<td>50000</td>
<td>400000</td>
</tr>
<tr>
<td>Unmixing TPR*</td>
<td>93.87</td>
<td>90.29</td>
<td>92.65</td>
<td>89.81</td>
<td>90.83</td>
<td>91.49</td>
</tr>
<tr>
<td>Unmixing FPR*</td>
<td>3.46</td>
<td>4.12</td>
<td>0.5</td>
<td>4.99</td>
<td>8.17</td>
<td>4.25</td>
</tr>
<tr>
<td>RF TPR*</td>
<td>88.86</td>
<td>85.19</td>
<td>88.27</td>
<td>83.33</td>
<td>79.91</td>
<td>85.11</td>
</tr>
<tr>
<td>RF FPR*</td>
<td>5.12</td>
<td>8.21</td>
<td>2.17</td>
<td>12.71</td>
<td>15.13</td>
<td>8.67</td>
</tr>
</tbody>
</table>

* TPR – True positive rate, FPR – False positive rate, RF – Random Forest

Classification accuracy assessment

<table>
<thead>
<tr>
<th></th>
<th>RF</th>
<th>Unmixing based classification</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PA**</td>
<td>UA**</td>
</tr>
<tr>
<td></td>
<td>PA**</td>
<td>UA**</td>
</tr>
<tr>
<td>Class</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Forest</td>
<td>87.97</td>
<td>88.69</td>
</tr>
<tr>
<td></td>
<td>91.17</td>
<td>93.29</td>
</tr>
<tr>
<td>Farmland</td>
<td>85.11</td>
<td>86.75</td>
</tr>
<tr>
<td></td>
<td>94.87</td>
<td>85.63</td>
</tr>
<tr>
<td>Water</td>
<td>89.87</td>
<td>87.43</td>
</tr>
<tr>
<td></td>
<td>89.65</td>
<td>95.12</td>
</tr>
<tr>
<td>Urban (SF)</td>
<td>84.24</td>
<td>83.19</td>
</tr>
<tr>
<td></td>
<td>85.33</td>
<td>94.31</td>
</tr>
<tr>
<td>Urban (LA)</td>
<td>76.52</td>
<td>83.39</td>
</tr>
<tr>
<td></td>
<td>88.67</td>
<td>92.96</td>
</tr>
<tr>
<td>Average</td>
<td>84.74</td>
<td>85.89</td>
</tr>
<tr>
<td></td>
<td>89.94</td>
<td>92.66</td>
</tr>
<tr>
<td>OA</td>
<td>85.31</td>
<td>91.30</td>
</tr>
<tr>
<td>Kappa</td>
<td>0.83</td>
<td>0.89</td>
</tr>
</tbody>
</table>

** PA – Producer’s Accuracy, UA – User’s Accuracy
Proportion of forest cover (top), farmland (middle), and water bodies (bottom) obtained from unmixing, RF, NLCD and NAFD for 100 sample patches. The straight lines indicate mean of the corresponding classes.

Proportion of urban built-up in SF Bay Area (top) and LA area (bottom) obtained from unmixing, RF, and NLCD for 100 sample patches. The straight lines indicate mean of the corresponding classes.
Conclusions

• S-V-D fractional maps obtained from global WELD using unmixing were shown to be useful in deriving forest, farmland, water and urban areas.

• An overall classification accuracy of over 91% was achieved.

• Validation of these land cover maps with NLCD 2011 products and NAFD maps revealed a 6% improvement in unmixing based classification and their advantages over perpixel classifier such as Random Forest.

• Unmixing based approach is a feasible method for land cover mapping on a global scale.

• The current study may be a pathfinder, recognizing that the mapping of land cover classes from fractional estimates at both the local and global scales will continue to improve over time.
Thank you.