Advanced Information Systems Technology (AIST)

2023 Annual Reviews

Analytic Collaborative Frameworks (ACF)

July 21, 2023
<table>
<thead>
<tr>
<th>Time</th>
<th>End</th>
<th>Duration</th>
<th>Project #</th>
<th>Short Title</th>
<th>PI / Presenting</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:00 AM</td>
<td>8:30 AM</td>
<td>0:30</td>
<td>Gathering</td>
<td>Introduction</td>
<td></td>
</tr>
<tr>
<td>8:30 AM</td>
<td>8:45 AM</td>
<td>0:15</td>
<td>AIST-21-0068</td>
<td>Coupled statistics-physics guided learning</td>
<td>Xie</td>
</tr>
<tr>
<td>8:45 AM</td>
<td>9:15 AM</td>
<td>0:30</td>
<td>AIST-21-0093</td>
<td>Detecting artifacts and transients</td>
<td>Bock</td>
</tr>
<tr>
<td>9:15 AM</td>
<td>9:45 AM</td>
<td>0:30</td>
<td>AIST-21-0051</td>
<td>Knowledge Transfer for Robust GeoAI</td>
<td>Prasad</td>
</tr>
<tr>
<td>9:45 AM</td>
<td>10:15 AM</td>
<td>0:30</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10:15 AM</td>
<td>10:30 AM</td>
<td>0:15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10:30 AM</td>
<td>11:00 AM</td>
<td>0:30</td>
<td>AIST-21-0037</td>
<td>ACF for river water quality and quantity</td>
<td>Gleason</td>
</tr>
<tr>
<td>11:00 AM</td>
<td>11:30 AM</td>
<td>0:30</td>
<td>AIST-21-0067</td>
<td>Integration of Obs and Models for Coastal Water Quality</td>
<td>Schollaert Uz / Ames</td>
</tr>
<tr>
<td>11:30 AM</td>
<td>12:00 PM</td>
<td>0:30</td>
<td>AIST-QRS-23-0002</td>
<td>ML for processing GHG and air pollutants</td>
<td>Yang</td>
</tr>
<tr>
<td>12:00 PM</td>
<td>1:20 PM</td>
<td>1:20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1:20 PM</td>
<td>1:50 PM</td>
<td>0:30</td>
<td>AIST-21-0005</td>
<td>Stochastic Parameterization of Atmospheric Model with Quantum Annealing</td>
<td>Guillaume</td>
</tr>
<tr>
<td>1:50 PM</td>
<td>2:20 PM</td>
<td>0:30</td>
<td>AIST-21-0025</td>
<td>SLICE: Semi-supervised ML for images</td>
<td>Wilson</td>
</tr>
<tr>
<td>2:20 PM</td>
<td>2:50 PM</td>
<td>0:30</td>
<td>AIST-21-0015</td>
<td>TO2SCA: Thematic Observation Search</td>
<td>Haddad</td>
</tr>
<tr>
<td>2:50 PM</td>
<td>3:20 PM</td>
<td>0:30</td>
<td>AIST-21-0020</td>
<td>Open Climate Workbench</td>
<td>H. Lee</td>
</tr>
<tr>
<td>3:20 PM</td>
<td>3:30 PM</td>
<td>0:10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>BREAK</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>LUNCH</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>Wrap-up</strong></td>
<td></td>
</tr>
</tbody>
</table>
Advanced Information Systems Technology (AIST)

Analytic Collaborative Frameworks (ACF)
Annual Grouped Technical Reviews

Jacqueline Le Moigne
July 21, 2023

Three AIST Thrusts

Observe, Target and Coordinate
Edge and on-the-ground intelligent planning, evaluating, coordinating and operating collections of diverse and distributed observing assets

Fuse, Analyze, Share and Collaborate
Simplify access to diverse and large amounts of data, analytics & modeling tools and advanced computational resources for collaborative science

Interrogate, Simulate, Trade and Visualize
Robust tools for interrogating, assessing uncertainties & causality, and for visualization, leveraging diverse data, models and products

Intervene and Assess
Running “what-if” scenarios to assess the impact of natural and human activities on the planet.

NOS = Novel Observing Strategies
ACF = Analytic Collaborative Frameworks
From Archives to Analytic Frameworks
Focus on the Science User

Data Archives
Focus on data capture, storage, and management
Each user has to find, download, integrate, and analyze

Analytic Frameworks
Focus on the science user
Integrated data analytics & tools tailored for a science discipline

?!

80% Prepare
20% Analyze

20% Prepare
80% Analyze

Facilitates collaborative science across multiple missions and data sets

Analytic Collaborative Frameworks (ACF)
Focus is on the Science User

Allow flexibility/tailor configurations for Science investigators to choose among a large variety of datasets & tools

Data
• Catalog
• NASA DAAC
• Other US Govt
• Non-US
• Local or non-public

User
• Project Definition
• Plan for Investigation

Tools
• Discovery & Catalog
• Work Management
• Data Interfaces
• Analytic Tools
• Modeling
• Collaboration
• Visualization
• Sharing/Publication
• Local/custom

Storage
• Data Containers
• Thematic model
• Metadata/Ontology
• Resulting Products
• Published data
• Provenance

Computing
• Local systems
• High End Computing
• Cloud Computing
• Quantum Computing
• Neuromorphic Computing

Project Work Environment
Computational Infrastructure
• Computing
  • Capacity
  • Capability
• Storage
• Communications

Search

Science

Applications

Decision Support
**Analytic Collaborative Frameworks (ACF)**

**Focus is on the Science User**

- **Develop ACFs for new science domains:**
  - Manage large data volumes
  - Manage wide variety of data types
  - Manage frequent data updates (high data velocity)

- **User**
  - Project Definition
  - Plan for Investigation

- **Tools**
  - Discovery & Catalog
  - Work Management
  - Data Interfaces
  - Analytic Tools
  - Modeling
  - Collaboration
  - Visualization
  - Sharing/Publication
  - Local/custom

- **Data**
  - Catalog
  - NASA DAAC
  - Other US Govt
  - Non-US
  - Local or non-public

- **Computational Infrastructure**
  - Computing
    - Capacity
    - Capability
  - Storage
  - Communications

- **Storage**
  - Data Containers
  - Thematic model
  - Metadata/Ontology
  - Resulting Products
  - Published data
  - Provenance

- **ADVANCED ANALYTICS:**
  - Data Accessibility
  - Data Fusion
  - Big Data Analytics
  - Data Mining
  - On-Demand Product Generation
  - Data Operations Workflows
  - Data Incorporation of Metadata, Provenance, Semantics, etc.

- **IMPROVED MODELING CAPABILITIES:**
  - Science Data Model Validation
  - Software Architecture Frameworks
  - Science Code Development & Reuse
  - Modeling Systems
  - Model Data Inter-Comparisons
  - Custom Tools
  - Forecasting/Prediction

- **AI CAPABILITIES:**
  - Machine Learning
  - Deep Learning
  - Data Services Discovery
  - Uncertainty Quantification Methods

**Grouped Reviews Objectives**

- **Respond to Annual ESTO AIST Reporting Requirements**
  - Technical Annual Reviews Grouped by Focus Areas
  - Individual Programmatic Reporting

- **Establish Relationship between Awardees**
  - Assess complementarity of various approaches and technologies in same AIST thrust
  - Investigate potential collaboration/coordination opportunities (potentially share algorithms, codes or ideas)
  - Investigate 3rd Optional Year teaming arrangements:
    - **If proposed, optional 3rd Years – will be selected 18 Months after project start**
    - **For one of three purposes:**
      1. Transition AIST technology to another Program or project
      2. Develop NOS-Testbed Concept and/or Demonstration
      3. ESDT Prototype
    - **Not all proposed 3rd Years might be funded**
    - **Can be different than original proposal but no budget increase**
    - **Collaborative AIST Projects will be prioritized/encouraged (i.e., several AIST projects in a system-of-systems approach**

- **Introduce AIST Projects and PIs to Broader Community**
  - Present AIST projects to NASA ESD Program Managers/Scientists and partner organizations
  - Facilitate technology infusions and knowledge transfer of AIST projects upon completion.

- **Review Needs in terms of:**
  - SMCE (NASA Science Managed Cloud Environment): AWS system access
  - ESIP: Project analysis to improve infusion and transition opportunities
Between 12 and 18 months in your project, you can request an [Assessment of Maturity by ESIP](#) ("Earth Science Information Partners")

- No cost to the PIs

**Process:**

1. **Objectives Set up and Facilitation:**
   - ESIP provides access to the Earth Science community & feedback on your technology/product/tool
   - ESIP will work with PIs to set specific objectives, taking into consideration TRL
   - ESIP will facilitate evaluator calls, development of evaluation plan, communication with PIs

2. **Technical Exchange Meeting:**
   - PI team meets evaluators.
   - Big picture to backend... evaluators should have a solid understanding of the purpose and goals of technology

3. **Evaluation Period:**
   - ESIP coordinates evaluation process.
   - Evaluators meet regularly, requesting information from PIs when necessary.

4. **Final Report:**
   - ESIP works with evaluators to create final report to be shared with PIs & AIST.
   - Reports can be public upon PI request.
Coupled Statistics-Physics Guided Learning to Harness Heterogeneous Earth Data at Large Scales

Yiqun Xie (PI, University of Maryland)
Sergii Skakun (Co-I, University of Maryland)
Xiaowei Jia (Co-I, University of Pittsburgh)

AIST-21-0068 Annual Programmatic Review
Date: 07/21/2023

Team listing: Zhihao Wang (UMD), Yiming Zhang (UMD), Erhu He (University of Pittsburgh)

Objective
Develop a novel statistically and physically guided machine learning framework to address key challenges in harnessing big data from Earth science, including spatial heterogeneity and limited availability of training data. The newly developed technology will be evaluated and validated on important Earth science tasks including Land Cover and Land Use Change (LCLUC) monitoring, and surface water monitoring. Open-source the code and tools developed during the project, as well as benchmark datasets.

Approach
- Thrust 1: Develop statistically guided learning frameworks to spatially transform a user-selected deep network architecture to add spatial heterogeneity-awareness, with extensions to cover random forest.
- Thrust 2: Develop physics-guided learning, which incorporates physics laws or knowledge into the network architecture to address the challenge on limited training data.
- Thrust 3: Develop coupled statistics-physics models to unite the strengths of both to cover various Earth science application scenarios.

Co-I/Partners: Xiaowei Jia (Co-I), University of Pittsburgh; Sergii Skakun (Co-I), University of Maryland, College Park.

Key Milestones
- Thrusts 1 & 2: design and implementation
  - GitHub repository v1 published
  - Papers accepted at AAAI 2023
- Thrust 3: design and implementation
  - Preliminary results obtained
  - Papers accepted at AAAI 2023
- Evaluation on existing validation data (TRL 4) 07/23
- Evaluation on ES-task benchmark data (TRL 5) 12/23

TRL\textsuperscript{current} = 4
Presentation Contents

- Background and Objectives
- Technical and Science Advancements
- Summary of Accomplishments and Future Plans
- Actual or Potential Infusions and Collaborations
- Publications - List of Acronyms

Background / Objectives (1)

- Earth science applications:
  - **Potential programs**: Land-Cover and Land-Use Change (LCLUC) Program, NASA Harvest Program, and the Terrestrial Hydrology Program.
  - **LCLUC and Harvest**:
    - **Mapping of major commodity crops** (e.g., wheat, maize and soybean):
      - We will use USDA's Cropland Data Layer (CDL) as reference data in the US, and ground in situ measurements from other known-challenging geographic regions (e.g., Ukraine in Europe, Kenya in Africa), to validate the effectiveness of the new frameworks.
      - We will apply the new methods for mapping land use changes occurring in urban areas (e.g. transformation from industrial land use to residential, parking to buildings).
    - **Cloud mapping**: A common critical task shared by many ES studies, including LCLUC mapping.
  - **Surface water monitoring**:
    - We will build predictive models for stream water temperature and streamflow in the Delaware River Basin, which is an ecologically diverse region and a societally important watershed along the East Coast of the US as it provides drinking water to over 15 million people.
Background/Objectives (2)

- **Thrust 1: Heterogeneity-aware learning:** To harness the spatial heterogeneity, we will develop spatial statistical frameworks to automatically recognize heterogeneous generation functions embedded in data (e.g., target predictions as functions of spectral bands), capture their complex footprints, and enable sharing and transferring of data-generated knowledge over space.

- **Thrust 2: Knowledge/physics-guided learning:** To go beyond the limit of pure data-driven method in scenarios with insufficient training data, which is common for many ES problems, we will develop new physics-guided learning frameworks and strategies to leverage the power of existing scientific knowledge.

- **Thrust 3: Coupled frameworks with heterogeneity-awareness and knowledge/physics:** we will leverage the accumulated physical knowledge to enhance the heterogeneity-aware learning framework so as to better discover the data heterogeneity amongst physical systems and improve the performance with limited data.

- **Benchmark data creation and evaluation/testing:** We will evaluate the performance of the new frameworks on important ES problems selected by ES experts to facilitate infusion and adoption of the methods.
  - Crop mapping and land cover monitoring
  - Cloud masking
  - Hydrology: water temperature and streamflow monitoring

Heterogeneity-Aware Learning: Example

- One size AI does not fit all

Which is snow?

- Runn of Kutch, Gujarat, India: 255, 255, 255
- White Sands, NM, USA: 255, 255, 255
- Snow: 255, 255, 255

Confusion:

- Salt
- Sand
- Snow
Heterogeneity-Aware Learning: Example

- One size AI does not fit all

\[ Y = f(X, U) \]

\[ f(X) = f(X) \]

Al output

Reality

\[ Y_1 \neq Y_2 \]

\[ f(X, U_1) \neq f(X, U_2) \]

Presentation Contents

• Background and Objectives
• Technical and Science Advancements
• Summary of Accomplishments and Future Plans
• Actual or Potential Infusions and Collaborations
• Publications - List of Acronyms

Technical and Science Advancements

1. Statistically guided heterogeneity-aware learning
   • Matured the implementation and refactored the code to a well-defined collection of functions that can be easily extended to different input formats and problems
   • Designed and implemented the heterogeneity-aware random forest.
   • Completed both evaluations on synthetic and real data.

2. Knowledge/physics guided learning
   • Implemented deep learning models (fully-connected, recurrent and graph networks) and different training strategies using physical simulations (pre-training, augmentation) for different physical processes
   • Completed evaluations on small-scale watersheds

3. Coupled heterogeneity-aware and knowledge guided learning
   • Implemented the coupled model design for satellite-agnostic cloud masking.
     • Completed preliminary training & evaluation on Landsat-8, Sentinel-2 and PlanetScope.
   • Implemented the coupled model design for baseflow prediction across heterogeneous river basins.
     • Completed evaluations on 60 river basins in Pennsylvania.
Theme 1: Heterogeneity-Aware Learning

- **Technology gaps**
  - A single model cannot approximate heterogeneous distributions
  - Spatial footprints are unknown

- **General design**

![Diagram of Heterogeneity-Aware Learning Framework](image)

Heterogeneity-Aware Learning Framework

![Diagram of the framework showing spatial partitioning, network transformation, and training progress](image)

Legend:
- Overall Loss
- Task-Specific Loss

Training Loss

- After Level 2 Partitioning
- After Level 3 Partitioning
Heterogeneity-Aware Learning Framework

Current Data for Testing

- **Setup**
  - Samples from the CONUS that are equally separated by 1km: ~7.8M in total
  - 10 spectral band features x 33 timestamps + 3 topographical features
  - Scenarios for four major crops
    - Soybean, corn, wheat, rice
Example Results: Synthetic Data

- Setup
  - Manually create regions with a different relationship between X & Y

Artificial region: Swapped binary labels for soybean

Learned partitioning:

F1-score improvements:
Differences between heterogeneity-aware & single model
Example Results: Real Data

• Results for soybean

![Map showing distribution and learned partitions for soybean samples.](image)

F1-score improvements:
Differences between heterogeneity-aware & single model

Example Results: Real Data

• Improvements for soybean, corn, wheat and rice

![Maps showing F1 improvements for soybean, corn, wheat, and rice.](images)
Implementation & Code Repositories

- Matured the implementation and refactored the code to a well-defined collection of functions that can be easily extended to different input formats and problems
  - Per-pixel NN, CNN (UNet as an example) and RNN (LSTM)
  - Traditional ensemble models (not posted yet)

Example classes and functions:

```python
# Example implementation for a dense net.
# Can be used as a template.

class DenseNet(nn.Module):  # Need to add modules and functions
  def __init__(self, num_classes):
      super(DenseNet, self).__init__()
      self.layer_size = (DENSE_SIZE, num_layers - DENSE_SIZE, num_classes)
      self.model = nn.Sequential()
      self.x = DenseNet(input_size, output_size)  # DENSE_SIZE, output_size
```
Theme 2: Knowledge/Physics-Guided Learning

Technology gaps:
- Physics: Siloed physical models for different physical processes: rivers, reservoirs, …
- Machine learning: Limited observations to learn complex graph neural networks (GNNs) for accurate predictions

Approach:
- Composite simulation from siloed models & Pre-training with composite simulations
- GNN is finetuned with limited observations

Model training procedure:
- Phase 1: The model first creates composite simulations by optimizing the way to combine simulated data or different entities in the graph.
- Phase 2: The model gets pre-trained using the obtained composite simulations.
- Phase 3: The models gets fine-tuned using available observations.
Current Data for Evaluation

- The water temperature of 56 river segments around the Cannonsville and Pepacton reservoirs in Lordville, NY.
- The river segments were defined by the National Hydrologic Model (NHM).
- Simulated water temperature for streams and reservoirs:
  - Stream temperature model - PRMS-SNTemp
  - Reservoir temperature model - General Lake Model
- Meta features
  - Stream characteristics: soil property, land cover, canopy, geographic location, etc.
  - Reservoir characteristics: dam height, dam length, depth, evaluation, the area of catchment.

Example Results: Water Temperature Prediction

Accuracy of the composite physical simulations

- The graph model combines the biased simulations of water temperature for both streams and reservoirs to create a new composite simulated dataset.
- The obtained composite simulations are more accurate than the simulations produced by the process-based SNTemp model and the latest USGS data release of composite simulations (Exp-Decay).
Example Results: Water Temperature Prediction

Accuracy of the composite physical simulations

- The graph model combines the biased simulations of water temperature for both streams and reservoirs to create a new composite simulated dataset.
- The obtained composite simulations are more accurate than the simulations produced by the process-based SNTemp model and the latest USGS data release of composite simulations (Exp-Decay).

Predictive performance of the fine-tuned model

- Use the composite simulated data to pre-train the recurrent graph network model.
- Fine-tune the model with observations

![Graph showing water temperature predictions](image-url)
Example Results: Water Temperature Prediction

Predictive performance of the fine-tuned model
- Use the composite simulated data to pre-train the recurrent graph network model.
- Fine-tune the model with observations

Theme 3: Coupled Models

- Knowledge-Guided Self-Supervised Learning
- Physics-guided Meta-Learning
Background in the context of ChatGPT

- Self-supervised learning is a critical component behind large language models (LLMs), including ChatGPT

https://amitness.com/2020/05/self-supervised-learning-nlp/

Document

Previous sentence

Center Sentence Captain America tries lifting Thor's hammer

Next Sentence

https://amitness.com/2020/05/self-supervised-learning-nlp/
Coupled Models 1: Self-Supervised Cloud Masking

Background in the context of ChatGPT

- Self-supervised learning is a critical component behind large language models (LLMs), including ChatGPT
- Labeled data are:
  - Limited
  - Highly-localized

- ChatGPT ≠ Remote Sensing
  - Success is for very specific tasks
  - Hard to replicate

Data distribution in space

TB/PBs of:

Raw data already contain human labels

Often do not contain human labels
Background in the context of ChatGPT

- Self-supervised learning is a critical component behind large language models (LLMs), including ChatGPT

- Labeled data are:
  - Limited
  - Highly-localized

- ChatGPT ≠ Remote Sensing
  - Success is for very specific tasks
  - Hard to replicate

But what if we can reproduce it for remote sensing?
(for some tasks)

---

**Coupled Models 1: Self-Supervised Cloud Masking**

- Use knowledge to realize self-supervision (label-free)

![Diagram]

If Supervised: Simple
But no label
No "supervision"
Coupled Models 1: Self-Supervised Cloud Masking

- Use knowledge to realize self-supervision (label-free)

\[ F \]

Supervise with spatio-temporal dynamics

Spatio-temporal packs  Cloud masks

Freq. of change
Clouds

(All images are for the same location)

Cloud pattern:
- Temporal
  - Random & Fast
- Spatial: random

Earth surface

Less vegetation
More vegetation
Less vegetation (periodicity)

Surface pattern:
- Temporal
  - Periodic & Slow
- Spatial: gradual

Time

Feb., Year 1  Aug., Year 1  Feb., Year 2

Dependency: Only observable once clouds are removed
Coupled Models 1: Self-Supervised Cloud Masking

- Use knowledge to realize self-supervision (label-free)

\[ F \]

\[ ST \text{-similarity} \]

Cloud-free composite generation:

\[ I_{com}^{I} = \left( \sum_{i \epsilon T_1} I_i \otimes (1 - M_i) \right) \odot \left( \sum_{i \epsilon T_2} (1 - M_i) \right) \]

- Cloud-free composite generation:

Spatial similarity: increases as the spatial overlap between two windows increases
Preliminary Evaluation

- 6 locations in 5 continents

<table>
<thead>
<tr>
<th>Test area</th>
<th>PML#</th>
<th>UNet</th>
<th>UNet-DA</th>
<th>D3</th>
<th>D3-DA</th>
<th>Kmeans</th>
<th>HD</th>
<th>DEC</th>
<th>Auto-CM</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1</td>
<td>0.959*</td>
<td>0.964 (0.985)</td>
<td>0.961</td>
<td>0.775 (0.904)</td>
<td>0.920</td>
<td>0.829</td>
<td>0.723</td>
<td>0.668</td>
<td>0.966</td>
</tr>
<tr>
<td>L2</td>
<td>0.891</td>
<td>0.908 (0.961)</td>
<td>0.919</td>
<td>0.857 (0.902)</td>
<td>0.810</td>
<td>0.849</td>
<td>0.694</td>
<td>0.923</td>
<td>0.940</td>
</tr>
<tr>
<td>Mean</td>
<td>0.925</td>
<td>0.936 (0.973)</td>
<td>0.940</td>
<td>0.816 (0.905)</td>
<td>0.865</td>
<td>0.839</td>
<td>0.709</td>
<td>0.795</td>
<td>0.953</td>
</tr>
</tbody>
</table>

Landsat

<table>
<thead>
<tr>
<th>Test area</th>
<th>FMS2</th>
<th>UNet</th>
<th>UNet-DA</th>
<th>D3</th>
<th>D3-DA</th>
<th>Kmeans</th>
<th>HD</th>
<th>DEC</th>
<th>Auto-CM</th>
</tr>
</thead>
<tbody>
<tr>
<td>L3</td>
<td>0.741</td>
<td>0.552 (0.686)</td>
<td>0.655</td>
<td>0.081 (0.457)</td>
<td>0.460</td>
<td>0.643</td>
<td>0.482</td>
<td>0.678</td>
<td>0.737*</td>
</tr>
<tr>
<td>L4</td>
<td>0.675</td>
<td>0.754 (0.92)</td>
<td>0.883</td>
<td>0.669 (0.794)</td>
<td>0.565</td>
<td>0.808</td>
<td>0.814</td>
<td>0.885</td>
<td>0.913</td>
</tr>
<tr>
<td>Mean</td>
<td>0.708</td>
<td>0.653 (0.803)</td>
<td>0.769</td>
<td>0.375 (0.626)</td>
<td>0.512</td>
<td>0.726</td>
<td>0.648</td>
<td>0.782</td>
<td>0.825</td>
</tr>
</tbody>
</table>

Sentinel-2

<table>
<thead>
<tr>
<th>Test area</th>
<th>Fmask</th>
<th>UNet</th>
<th>UNet-DA</th>
<th>D3</th>
<th>D3-DA</th>
<th>Kmeans</th>
<th>HD</th>
<th>DEC</th>
<th>Auto-CM</th>
</tr>
</thead>
<tbody>
<tr>
<td>L5</td>
<td>0.818 (0.923)</td>
<td>0.899</td>
<td>0.55 (0.772)</td>
<td>0.855</td>
<td>0.823</td>
<td>0.297</td>
<td>0.893</td>
<td>0.930</td>
<td></td>
</tr>
<tr>
<td>L6</td>
<td>0.563 (0.907)</td>
<td>0.904</td>
<td>0.81 (0.811)</td>
<td>0.800</td>
<td>0.753</td>
<td>0.772</td>
<td>0.825</td>
<td>0.898*</td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>0.69 (0.915)</td>
<td>0.902</td>
<td>0.675 (0.791)</td>
<td>0.827</td>
<td>0.788</td>
<td>0.534</td>
<td>0.859</td>
<td>0.914</td>
<td></td>
</tr>
</tbody>
</table>

PlanetScope

Additional Evaluation: Sentinel-2 (Expanding)

- Dataset: currently 20 locations (from Sentinel-2 Cloud Mask Catalogue)
  - Image size: 1022 x1022
  - Performed additional quality check & filtering
  - Self-adaptive thresholding (showing the best for machine learning methods)

Examples of removed data
(questionable "truth")
Additional Evaluation: Sentinel-2 (Expanding)

- Dataset: currently 20 locations (from Sentinel-2 Cloud Mask Catalogue\(^1\))
  - Image size: 1022 x 1022
- Performed additional quality check & filtering
- Self-adaptive thresholding (showing the best for machine learning methods)

Examples of Results

<table>
<thead>
<tr>
<th></th>
<th>Auto-CM</th>
<th>UNet-5</th>
<th>UNet-10</th>
<th>Fmask</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Accuracy</strong></td>
<td>0.95</td>
<td>0.63</td>
<td>0.9</td>
<td>0.92</td>
</tr>
<tr>
<td><strong>Average F1</strong></td>
<td>0.9</td>
<td>0.68</td>
<td>0.83</td>
<td>0.84</td>
</tr>
<tr>
<td><strong>Overall F1</strong></td>
<td>0.96</td>
<td>0.76</td>
<td>0.92</td>
<td>0.93</td>
</tr>
</tbody>
</table>

Additional Evaluation: Landsat-8 (Expanding)

- More evaluation (expanding):
  - Dataset: currently 20 locations (from Landsat 8 Level-1 Collection 1&2 \(^1,2\))
  - Image size: ~7000x7000 to 8000x8000
- Self-adaptive thresholding (showing the best for machine learning methods)

Examples of Results

<table>
<thead>
<tr>
<th></th>
<th>Auto-CM</th>
<th>UNet-5</th>
<th>UNet-10</th>
<th>Fmask</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Accuracy</strong></td>
<td>0.98</td>
<td>0.91</td>
<td>0.97</td>
<td>0.95</td>
</tr>
<tr>
<td><strong>Average F1</strong></td>
<td>0.92</td>
<td>0.87</td>
<td>0.91</td>
<td>0.89</td>
</tr>
<tr>
<td><strong>Total F1</strong></td>
<td>0.97</td>
<td>0.88</td>
<td>0.96</td>
<td>0.93</td>
</tr>
</tbody>
</table>

1. https://www.sciencebase.gov/catalog/item/60536863d34e7eb1cb3ebfb1
2. https://www.sciencebase.gov/catalog/item/61015b2fd34ef8d7055d6395
Additional Evaluation: PlanetScope (Expanding)

- Dataset expansion for PlanetScope

  ![Image] ![Cloud mask] ![Cloud-free composite]

- Example results

Coupled Models 2: Physics-guided Flow Prediction

Baseflow Prediction
- **Physics-based:** Various models have been developed under different assumptions for different scenarios & unclear which one to use
- **Machine learning:** Limited samples & hard to generalize to heterogeneous scenarios

Approach
- Leverage physical/mathematical equations to guide the ML model.
- Automatically select which physical model to use with meta-learning.

Physics Models/Knowledge
- Various equations for different scenarios:
  - $Y_1 = b(x)$
  - $Y_2 = c \exp(c \cdot x)$
  - $Y_3 = d \cdot x + e$ (linear models)
  - $Y_4 = f \cdot x + g$ (polynomial models)

Meta-Learning
- Which physical models to use to guide the training?

Physics-guided baseflow predictions
Evaluation: Baseflow Prediction

- **Dataset:**
  - **Baseflow labels:** obtained through the hydrograph separation from streamflow.
  - **Input features:** daily weather data from NCA-LDAS, and hydrological conditions from the GAGES-II dataset.
  - **Time period:** 01/01/1987-09/18/2016 (training), 09/18/2006-07/27/2016 (testing).
  - **Study region:** 60 river basins scattered in Pennsylvania. Each basin has only one station.

**Example results**

- **Sparse training data**
  - Test the model performance using only 0.5% data from the training period.
  - Comparison with the standard LSTM: better NSE in 58 out of 60 basins.
Localized example results

- Data annotations are often collected from a small number of locations while most locations are poorly observed or completely unobserved.
- Test the model performance on a half of basins with less training data
  - 50%, 10%, 5%, 0.5%

### NSE Comparison

<table>
<thead>
<tr>
<th>100%</th>
<th>50%</th>
<th>10%</th>
<th>5.00%</th>
<th>0.50%</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM</td>
<td>PGMTL-Contrastive</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Presentation Contents

- Background and Objectives
- Technical and Science Advancements
- Summary of Accomplishments and Future Plans
- Actual or Potential Infusions and Collaborations
- Publications - List of Acronyms
Summary

1. Statistically guided heterogeneity-aware learning
   • Matured the implementation and refactored the code to a well-defined collection of functions that can be easily extended to different input formats and problems
   • Designed and implemented the heterogeneity-aware random forest.
   • Completed both evaluations on synthetic and real data.

2. Knowledge/physics guided learning
   • Implemented deep learning models (fully-connected, recurrent and graph networks) and different training strategies using physical simulations (pre-training, augmentation) for different physical processes
   • Completed evaluations on small-scale watersheds

3. Coupled heterogeneity-aware and knowledge guided learning
   • Implemented the coupled model design for satellite-agnostic cloud masking.
     • Completed preliminary training & evaluation on Landsat-8, Sentinel-2 and PlanetScope.
   • Implemented the coupled model design for baseflow prediction across heterogeneous river basins.
     • Completed evaluations on 60 river basins in Pennsylvania.

Plans Forward

Final phase goal: Increase TRL to level 5

1. Heterogeneity-aware learning
   • Mature documentation for heterogeneity-aware learning GitHub code.
   • Continue to improve extensibility and flexibility of the code.
   • Finetune design decisions & more evaluation scenarios

2. Knowledge/physics guided learning
   • Extend the evaluation scale to the entire Delaware River Basin and another large-scale river basin in Houston.
     • Mature code & documentation on Github.

3. Coupled models (heterogeneity-awareness + physics)
   • Extend the evaluation scale: Data validation & comparison
   • Extend the evaluation diversity (cloud masking): Small satellite data
   • Finetune designs and perform larger-scale training
   • Mature code & documentation, and post on Github
Presentation Contents

- Background and Objectives
- Technical and Science Advancements
- Summary of Accomplishments and Plans Forward
- Actual or Potential Infusions and Collaborations
- Publications - List of Acronyms

Actual or Potential Infusions and Collaborations

- Summary of actual or potential infusions
  - Technology transfer: transferring technology via open-sourcing, or public domain release
  - Transition: transitioning into applications in polar research, carbon monitoring and sustainability (within Earth Science), as well as transportation research (outside Earth science)

- Summary of actual or potential collaborations
  - Dr. Xiaopeng Song: Agriculture monitoring
  - Dr. Chris Justice: Agriculture monitoring
  - Dr. George Hurtt: Ecology and carbon monitoring
  - Dr. Sinead Farrell: Sea ice monitoring
  - Dr. Xinyue Ye (TAMU): Urban flood risk estimation
  - Dr. Xianfeng Yang (UMD): Civil engineering and transportation
  - Kara Stuart (Aquaya): Poverty mapping; sustainability
  - Caroline Delaire (Aquaya): Poverty mapping; sustainability
  - Jacob Zwart (USGS): Stream modeling
Presentation Contents

- Background and Objectives
- Technical and Science Advancements
- Summary of Accomplishments and Plans Forward
- Actual or Potential Infusions and Collaborations
- Publications - List of Acronyms

Publications

- **Journal / Conference Papers**
  - [SDM'23] Shengyu Chen, Yiqun Xie, Xiang Li, Xu Liang and Xiaowei Jia. Physics-Guided Meta-Learning Method in Baseflow Prediction over Large Regions. SIAM International Conference on Data Mining (SDM'23), 2023. (acceptance rate: 27.4%) Best Application Paper Award.
  - [SDM'23] Xiaowei Jia, Shengyu Chen, Can Zheng, Yiqun Xie, Zhe Jiang, Nasrin Kalantar. Physics-guided Graph Diffusion Network for Combining Heterogeneous Simulated Data: An Application in Predicting Stream Water Temperature. SIAM International Conference on Data Mining (SDM'23), 2023. (acceptance rate: 27.4%)
- **Other (Talks/Posters)**
Acronyms
List of Acronyms

- ES: Earth science
- ML: Machine learning
- DL: Deep learning
- GNN: Graph neural network
- GDN: Graph diffusion network
- SSL: Self-supervised learning
Detection of Artifacts & Transients in Earth Science Observing Systems with Machine Learning & Visualization

Yehuda Bock¹
Angelyn Moore² (Co-I, Institutional PI)
Umaa D. Rebbapragada² & Joe T. Roberts² (Co-I, Is)
¹Scripps Institution of Oceanography
²Jet Propulsion Laboratory/California Institute of Technology

AIST-21-0093 Annual Technical Review
July 21, 2023

Team: Zhen Liu, Steven Lu, Fred Calef (JPL); Roland Hohensinn (ETH Zurich/SIO); Katherine Guns, Anne Sullivan, Rohith Rachala (SIO); Jayme Laber, Ivory Small (NOAA/NWS); Jonathan Weiss (NOAA/PTWC)

Objectives
- Discriminate transients from steady, state, tidal, and atmospheric motions using machine learning (ML) models, identify their sources, and any significant discrepancies from models of the underlying physical processes.
- Significantly reduce the time and effort spent in detecting and repairing artifacts in geodetic data and derived products using ML models. Achieving the most robust, precise and accurate science requires rigorous CoI/Co-I. ML models can improve this process.
- Building on a mature open-source, interactive, public data portal developed by us to view displacement time series and metadata, extend functionality to visualize higher-order spatio-temporal Earth Science Data Records (ESDs), interface with ML methods to iteratively improve the knowledge gap between raw observations, transients, and models.

Approach
- Create Prototype Transient & Artifact Continuum Learning System (TACLS) ML software suite. Precisely define the kinds of transients and artifacts that are sought.
- Leverage our archive of thousands of labeled artifacts and physical transients, and acquired expertise in creating calibrated and validated ESDs from thousands of GNSS stations and 30 years of data to train ML models.
- Leverage and expand the open-source visualization software, MAV2, released and maintained on the GoHub platform to serve as an interactive environment for exploring and mining ESDs, aided by ML methods.

Co-Investigators:
- Angelyn Moore, Zhen Liu, Joe Roberts, Umaa Rebbapragada (JPL); Jayme Laber, Ivory Small (NWS); Jonathan Weiss (PTWC)

Key Milestones
- Consolidate labeling of transients and artifacts in geodetic data and derived products (3-4 months)
- Prototype TACLS (1 year) TRL=5
- Prototype MAV2 interactive environment (1 year) TRL=5
- Enhance MAV2/TACLS to include continuous learning, demonstrate science test cases with NWS/WFOs and PTWC (2 years): TRL=6
- Infusion of software to NWS/WFOs and PTWC (3 years): TRL=7

TRLnew=4; TRLold = 7
Background and Objectives

- Discern transients from steady state crustal and atmospheric motions using machine learning (ML) methods and identify their sources and any significant discrepancies from models of the underlying physical processes.
- Significantly reduce the time and effort spent in Cal/Val by using ML models to detect and repair artifacts in geodetic data and derived products. ML models can improve a process that is mostly performed independently and inconsistently by groups of students and researchers.
- Build on a mature open-source, interactive, public data portal developed by our group to view displacement time series and metadata, extend functionality to visualize higher-order spatiotemporal Earth Science Data Records (ESDRs); interface with ML methods to iteratively improve the knowledge gap between raw observations, transients and models.

Synergisms:
- AIST thrust area: Analytic Collaborative Frameworks (ACF)
- NASA Focus areas: Earth Surface and Interior; Applied Sciences; NISAR
- AIST goals: Earth System Digital Twins; New Observing Strategies

Approach

- Leverage our archive of thousands of labeled artifacts and physical transients, and acquired expertise in creating calibrated and validated Earth Science Data Records from thousands of GNSS stations and up to 30 years of data, & InSAR/GNSS imagery to train ML models.
- Create Transient & Artifact Continuous Learning System (TACLS) ML software suite. Precisely define the kinds of transients and artifacts that are sought.
- Expand the open-source visualization software, MGViz, released and maintained on the GitHub platform to serve as an interactive environment for exploring and mining ESDRs, aided by ML methods.
Detecting Slow Slip Events in Cascadia
24 Episodic Tremor and Slip (ETS) Events since 1994

ETS events occur about every 14 months with a duration of several weeks

ETS affects hundreds of GNSS stations allowing us to model the events in space & time
ML Approaches

<table>
<thead>
<tr>
<th>ML Approach</th>
<th>Transients/Artifacts</th>
<th>Data Sets</th>
<th>Implementation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unsupervised</td>
<td>Coseismic offsets</td>
<td>Displacement time series -</td>
<td>Isolation Forest, Clustering</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cleaned</td>
<td></td>
</tr>
<tr>
<td>Supervised</td>
<td>Coseismic offsets</td>
<td>Historical database of</td>
<td>Random Forest, Graph Neural Networks⁰</td>
</tr>
<tr>
<td></td>
<td></td>
<td>coseismic offsets</td>
<td></td>
</tr>
<tr>
<td>Unsupervised</td>
<td>Artifacts (outliers,</td>
<td>Displacement time series -</td>
<td>Factor Graphs with NUV², Priors -</td>
</tr>
<tr>
<td></td>
<td>offsets)</td>
<td>raw</td>
<td>linear state models</td>
</tr>
<tr>
<td>Supervised</td>
<td>Artifacts (outliers,</td>
<td>Historical database of</td>
<td>Gradient Boosting</td>
</tr>
<tr>
<td></td>
<td>offsets)</td>
<td>artifacts</td>
<td></td>
</tr>
<tr>
<td>Unsupervised</td>
<td>Slow slip events</td>
<td>Displacement Residuals</td>
<td>Isolation Forest</td>
</tr>
<tr>
<td>Supervised</td>
<td>Slow slip events</td>
<td>Displacement Residuals, Slip</td>
<td>Random Forest, Convolutional neural</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Models</td>
<td>network¹</td>
</tr>
<tr>
<td>Unsupervised</td>
<td>Atmospheric rivers</td>
<td>Troposphere delay time</td>
<td>Isolation Forest</td>
</tr>
<tr>
<td></td>
<td></td>
<td>series</td>
<td></td>
</tr>
<tr>
<td>Supervised</td>
<td>Atmospheric rivers</td>
<td>Historical database of</td>
<td>Random Forest, Convolutional neural</td>
</tr>
<tr>
<td></td>
<td></td>
<td>troposphere delay,</td>
<td>network¹</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GNSS/InSAR imagery</td>
<td></td>
</tr>
</tbody>
</table>

¹Additional models under investigation ²Normal priors with unknown variance

Detecting Movement of Land-Falling Atmospheric River (AR)
High-Rate Troposphere Delays from GNSS Observations

Extreme weather observed by a network of GNSS stations cause significant delays in GNSS travel time allowing us to measure total water vapor content in the troposphere. NOAA/NWS uses these data to issue flash flood warnings.

Following the manual analysis in Wang et al. (2019), we automate detection of peak ZTD values as the AR moves across the terrain. The important parameters for meteorology are the velocity of travel and spatial extent.

Katherine Guns
TACLS Architecture Progress

- Significant progress on the supervised learning pipeline
- Internal dev MGViz can now display TACLS detections
- Detection of tropospheric events has started

1. Supervised learning pipeline is now green.
2. TACLS GNSS classifier is now green.
3. TACLS anomaly detector on tropospheric data (PWV) is in progress (yellow)
4. MGViz software now displays TACLS detections. User annotations pending (yellow)

Umaa Rebbapragada

Supervised Learning Experimental Methodology: Validation using Mw7.1 2019 Ridgecrest Earthquake

- East and North daily displacement time series cleaned of artifacts ("Raw_M") and outliers from MEaSUREs ESESES Western North America project archive
- Focused on 385 stations in SoCal that recorded the 2 coseismic events in July 2019 (inclusive of some non-Ridgecrest events)
- Data are divided into 21-day windows; slid 1 day at a time
- Windows containing a coseismic offset has a positive label; otherwise, it is a negative label
- Training Set: 1999 through 2010; Test: 2019
- Extracted time series features, selected most informative ones using statistical tests of relevance to the class labels
- Classifier: 200-tree Random Forest
- % of trees that voted positive classification is reported as probability of positive classification
Supervised Learning Ridgecrest (385 stations) – East Component

ROC Curve

- True Positive Rate (TPR) vs. False Positive Rate (FPR)
  - Curve that hugs top left corner performs best.
  - We are able to discover \( \geq 60\% \) detections with close to 0 false positives.

Precision-Recall Curve

- Recall (TPR) vs. Precision (% of correct positive predictions)
  - Curve that hug top right corner performs best.
  - At 60% recall (TPR), our precision is \(~40\%\).

Distribution of probabilities

- Our probabilities appear to be well calibrated.

Map of the 385 Southern California stations.
The probability for each day is the average over 21 windows containing that day, thus you will see probabilities increase before the event day.

April 4, 2019

Steven Lu
Supervised Learning Ridgecrest (385 stations) – East Component

Map of the 385 Southern California stations. The probability for each day is the average over 21 windows containing that day, thus you will see probabilities increase before the event day.

July 6, 2019 – M7.1

Probabilities from station p595 (adjacent to Ridgecrest, CA) for all time windows covering the event date.
Supervised Learning Ridgecrest (385 stations) – North Component

ROCurve

Our ROC and P-R curves look similar to the results for the East component.

North appears to have a higher number of high probabilities detections.

Supervised Learning Ridgecrest (385 stations) – North Component

Map of the 385 Southern California stations. The probability for each day is the average over 21 windows containing that day, thus you will see probabilities increase before the event day.

April 4, 2019

LA basin stations were affected by the Ridgecrest event primarily in the North component.

Steven Lu
Supervised Learning Ridgecrest (385 stations) – North Component

Map of the 385 Southern California stations. The probability for each day is the average over 21 windows containing that day, thus you will see probabilities increase before the event day.

July 6, 2019 – M7.1

LA basin stations were affected by the Ridgecrest event primarily in the North component.

Steven Lu

Supervised Learning Ridgecrest (385 stations) – East Component

We created a movie over year 2019 of our detections over the 385 Southern California stations. Dots disappear if a station did not have data on that day. The probability for each day is the average over 21 windows containing that day, thus you will see probabilities increase before the event day.

2019.5055 = July 4th 2019

Steven Lu
Supervised Learning Ridgecrest (385 stations) – North Component

We created a movie over year 2019 of our detections over the 385 Southern California stations. Dots disappear if a station did not have data on that day. The probability for each day is the average over 21 windows containing that day, thus you will see probabilities increase before the event day.

2019.5055 = July 4th 2019

The stations in the LA basin experienced more activity in the North direction.

Steven Lu

Offset Detection using Linear State Space Models

We define the state, \( X \), of our line of best fit by its level (l) and slope (s)

\[
X = \begin{bmatrix}
l_N \\
l_E \\
s_E
\end{bmatrix}
\]

An offset corresponds to a sharp jump in one or more level components of the state

Roland Hohensinn & Hugo Umbers
Offset detection using linear state space models

The state update at each timestep

\[ X_{k+1} = AX_k + BU_k + \Delta_k \]

With our state vector, \( X \), defining the level (l) and slope (s) of the line of best fit

The observation at each time step

\[ Y_k = CX_k + Z_k \]

We allow input jumps to instantaneously change the level and slope of the line

\[ U = \begin{bmatrix} u_l \\ u_s \end{bmatrix} \]

We incorporate some state noise, allowing the slope of the line to change slightly over time.

\[ \Delta_k = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \]

We also incorporate some observation noise to allow the model to ignore noisy fluctuations in the data.

\[ Z_k = \begin{bmatrix} Z_l \\ Z_s \end{bmatrix} \]

Roland Hohensinn & Hugo Umbers

---

Representation as Factor Graphs with NUV Priors

- We can represent this model as a factor graph to help compute the necessary quantities
- Estimates for state and observation noise are made based on empirical data
- Jump inputs, \( U(k) \), have a NUV prior (Normal with an Unknown Variance) which is learned over time
  - This promotes sparsity
  - The resulting sparse jump input vector becomes our estimate of when offsets occurred.

Roland Hohensinn & Hugo Umbers
Sample Results – Coseismic Offset at Station *lnmt in Mojave desert*

The model determines that the best explanation for the data in July 2019 is a level (jump) input. Hence an offset is detected here.

--

Roland Hohensinn & Hugo Umbers

Sample Results – Coseismic Offset & Slope Changes at Station *wrna in LA Basin*

Here the model detects features around a coseismic event in July 2019:
1. Level jump (offset) clearly detected
2. Slope jumps/changes detected in the months preceding and following the offset

Roland Hohensinn & Hugo Umbers
Summary of Accomplishments and Future Plans

- Current state:
  - Supervised and unsupervised pipeline code is complete. Can focus on running ML experiments now
  - Coseismic offset results are reasonable given complexity of data. Can turn focus to other types of transients
  - Cluster analysis shows promising results for unsupervised transient detection
  - Refined training displacement time series data set for artifact detection models

- Work planned - ML:
  - TACLS will spend Y2Q1 focused on leveraging data from neighboring stations to improve ML performance
  - Experiment with other state-of-the-art deep learning models in Y2Q2 to improve supervised learning performance
  - Focus on slow slip events and extreme weather experiments with current code base.

- Work Plan - MGViz Next Quarter:
  - Multi-site box selection tool for map interface
  - Annotate events on MGViz
  - Usability improvements on MGViz for visualizing TACLS results
Actual or Potential Infusions and Collaborations

- **Proposal infusion partners:** NOAA NWS Forecast Offices and NOAA NWS Pacific Tsunami Warning Center are both challenged by insufficient situational awareness that inadequately supports watch and warning decisions. The MGViz/TACLS platform will address both by providing:
  - ML-backed displays to improve discovery and tracking of severe weather transients
  - ML-based detection of slow slip transients across the Pacific basin
  - Our proposed infusion partners have been engaged with the project throughout Y1

- **Summary of actual and potential collaborations**
  - NOAA NWS Co-I Jayme Laber (Los Angeles/Oxnard Weather Forecast Office) and Ivory Small, San Diego Weather Forecast Office
  - NOAA NWS Pacific Tsunami Warning Center Co-I Dr. Jonathan Weiss
  - Collaborator Dr. David Sandwell, director of SIO InSAR group
  - Co-I’s engaged with NASA Earth Science R&A focus areas (ESI, including Space Geodesy Task, & Weather) and NASA Applied Science Program
  - Co-I’s engaged with EarthScope Consortium (previously UNAVCO)

---

Thank you!
Questions?
Publications


Hohensinn R. & Y. Bock (2022), Sensitivity Analysis of Western U.S. Velocities from Daily GNSS Displacements, Abstract G35B-0336, American Geophysical Union (AGU) Fall Meeting, Chicago, 14 December.


Acronyms

- **API**: Application Programming Interface
- **AR**: Atmospheric River
- **TACLS**: Transient and Artifact Continuous Learning System (TACLS)
- **ESDR**: Earth Science Data Records
- **ETS**: Episodic Tremor and Slip
- **GIS**: Geographic Information System
- **GNSS**: Global Navigation Satellite System
- **InSAR**: Interferometric Synthetic Aperture Radar
- **ML**: Machine Learning
- **MGViz**: MMGIS GNSS Visualizer
- **NUV**: Normal priors with unknown variance
- **PTWC**: Pacific Tsunami Warning Center
- **NOAA**: National Oceanic and Atmospheric Administration
- **NOS**: National Ocean Service
- **NWS**: National Weather Service
- **WFO**: Weather Forecast Office
- **LOS**: Line of Sight (InSAR)
- **MMGIS**: Multi-Mission Geographic Information System
- **ESESES**: Extended Solid Earth Science ESDR System
- **SSE**: Slow Slip Event
- **ZTD**: Zenith Troposphere Delay
- **IWV**: Integrated Water Vapor
- **TPR**: True Positive Rate
- **FPR**: False Positive Rate
- **TRL**: Technology Readiness Level
- **JPL**: Jet Propulsion Laboratory
- **SiO**: Scripps Institution of Oceanography
Knowledge Transfer for Robust GeoAI Across Space, Sensors and Time via Active Deep Learning

Saurabh Prasad (PI, University of Houston)
Melba Crawford (Co-I, Purdue University)

AIST-21-0051 Review Meeting
July 23, 2023

Objective
Develop a robust and effective strategy for Generative Adversarial Learning (GAN) based knowledge transfer across a geospatial sensor web.
Develop an active learning approach to systematically leverage additional ground-reference data for improved knowledge transfer.
Prototype and host the GAN framework on a commercial cloud for greater adoption among the geospatial community and stakeholders.
Validate performance with benchmark geospatial datasets and the relevancy scenario data.

Key Milestones
• Deep Knowledge Transfer via GANs 06/23
• Active Deep Learning 06/23
• Prototype on Commercial Cloud 12/24
• Validate with Benchmark & Relevancy Scenario 12/24

Approach
• Develop optics inspired and Bayesian inference approaches within our GAN framework to impart invariances to confounding sources of disparity between heterogenous sensors and to facilitate effective model/knowledge transfer across sensors, space and time and model uncertainty.
• Develop and demonstrate knowledge transfer for multi-branch source/target networks that can work with multiple source/target domain sensor nodes (data fusion).
• Develop semi-supervised and active learning capability within our GAN framework.

Co-I: Melba Crawford, Purdue University
Collaborator: Xiong Zhou, Amazon AWS AI
Collaborator: Johny Park, Wabash Heartland Innovation Network
Presentation Contents

- Background and Objectives
- Technical and Science Advancements
- Summary of Accomplishments and Future Plans
- Actual or Potential Infusions and Collaborations
- Publications - List of Acronyms

Background / Objectives

Objectives:
- Knowledge Transfer across sensors, space and time
- Knowledge can be deep features from sensor-specific deep networks (e.g. transfer learning), or raw imagery data (e.g. super-resolution)
- This can also address “nuisance factors” in GeoAI tasks.
- A successful implementation will facilitate seamless learning across sensors, space and time.
Background / Objectives

Sources of Variability:

- Sensors/Viewpoints/Imaging Geometry
- Geographical
- Illumination

One could use classical approaches, such as...

Application use-cases that will be studied in this project:

- Benchmark data for urban land-cover mapping and related applications (e.g., building footprints, disaster mapping etc.)
- Precision-agriculture data for high-throughput phenotyping over agricultural farms at and near Purdue University.

Background/Objectives (2)

Alternatively, we could leverage generative adversarial networks...

Technology: Built on a framework of generative adversarial learning
Presentation Contents

- Background and Objectives
- Technical and Science Advancements
- Summary of Accomplishments and Future Plans
- Actual or Potential Infusions and Collaborations
- Publications - List of Acronyms

Multi-Scale Super-Resolution using Generative Adversarial Learning

Related Work

- GAN-based Network:
  - GAN-based super-resolution (SRGAN): originally proposed for single RGB image super-resolution

  ![GAN-based super-resolution](image)

  - Advantage: learning-based supervised model, can reconstruct realistic texture details with a single low-resolution RGB image.
  - Disadvantage: requires hundreds of thousands of training RGB image pairs (low- and high-resolution)* cover a wide range of objects/scenarios to learn how to reconstruct detailed texture; can only handle 3 band RGB images.

Multi-Scale Super-Resolution using Generative Adversarial Learning

Methodology

- We extended and evaluated **GAN-based RGB-guided multispectral super-resolution**: 
  - Adopt the advantage of SRGAN: use the **GAN-based network** learn how to sharpen the edges/detailed textures 
  - Avoid the disadvantage of SRGAN: instead of training with hundreds of thousands of training images, we use a **high-resolution grayscale** image as one of the input to provide the spatial information 
  - Instead of 3 band RGB images, it can accommodate **multispectral** images and minimize the spectral distortion at the same time (modified GAN loss function)

**Methodology**

GAN-based RGB-guided multispectral image super-resolution network

- Input of the generator: low-resolution multispectral (**LRMS**) + high-resolution grayscale (**HRGS**) 
- Output of the generator: reconstructed high-resolution multispectral (**GenMS**) 

- Discriminator: differentiate the real high-resolution multispectral (**HRMS**) and the reconstructed high-resolution multispectral (**GenMS**)
Multi-Scale Super-Resolution using Generative Adversarial Learning

Methodology

Two networks developed/investigated: **MS SRGAN** and **MS ESRGAN**:
The MS ESRGAN adopts residual in residual dense block (RRDB) in the generator, while MS SRGAN only uses residual block (RB).

**Advantage:** with multi-level residual networks, RRDB has a deeper structure which helps learning detailed textures.

The MS ESRGAN uses the Relativistic average Discriminator (RaD), while MS SRGAN uses the standard discriminator (StD).

**Advantage:** the RaD helps the network reconstruct more realistic images than StD.

\[
D_{\text{rad}}(x, y) = \sigma(C(\text{real}) - \exp(C(\text{fake})) \to 1 \\
D_{\text{rad}}(x, y) = \sigma(C(\text{fake}) - \exp(C(\text{real})) \to 0
\]

Experiments and Results

The input data were generated using images from **one sensor on manned aircraft cover an urban area**

Model: **MS ESRGAN**

**HRMS:** high-resolution multispectral orthomosaic (60 cm GSD)

**LRMS:** down sampled (2.4 m GSD)

**HRGS:** formed by converting RGB bands of the HRMS to grayscale (60 cm GSD)
Experiments and Results

Experiments and results:
- MS ESRGAN trained and tested with subarea of the multispectral orthomosaic
- Using 50% as training, 50% as testing: both were randomly cropped from the corresponding area
  - 1000 training patches were cropped from the training part
  - 200 testing patches were cropped from the testing part

Experiments and Results

Experiments and results:
Model performance over the testing area

LRMS  Reconstructed MS (MS ESRGAN)  Ground reference
Experiments and Results

Experiments and results:
Model performance over the testing area

LRMS  
Reconstructed MS (MS ESRGAN)  
Ground reference

Error maps (x5)

- Red: positive error value
- Blue: negative error value
Multi-Scale Super-Resolution using Generative Adversarial Learning

Experiments and Results

Experiments and results:
Error maps (x5)

- Red: positive error value
- Blue: negative error value

Ground reference

Absolute difference per pixel over 200 testing images

- The first 8 bands have relatively smaller error than the last 8 bands
- There is an outlier image shows bigger error than the others
Transfer Learning of Models across Sensors/Space/Time

Approach

**Notation**

Let $S = \{(x^1_s, y^1_s) \mid x^1_s \in \mathbb{R}^{m_x \times n_x}, y^1_s \in \mathbb{R}^{m_y \times n_y} \}$ be the source domain. Also let $T = \{ (T_1, T_2) \}$ be the target domain, where the labeled target domain $T_1 = \{(x^2_t, y^2_t) \mid x^2_t \in \mathbb{R}^{m_x \times n_x}, y^2_t \in \mathbb{R}^{m_y \times n_y} \}$, and $T_2 = \{(x^2_t, y^2_t) \mid x^2_t \in \mathbb{R}^{m_x \times n_x}, y^2_t \in \mathbb{R}^{m_y \times n_y} \}$ is the unlabeled target domain.

- Use a few labeled samples from target domain, $T_1$, in addition to $S$ to solve the optimization problem:
  
  $$ \arg\min_{E, D} \mathbb{E}_{(x, y) \sim \mathcal{S}} \left[ \ell(E(x), y) \right] $$

- Incorporating $T_1$ alleviates negative transfer with a reasonable assumption of availability of a few labeled samples from the target domain for many applications.

- The discriminator model, $E_D$, $D$, can be learned by optimizing:
  
  $$ \arg\min_{E_D, D} \mathbb{E}_{(x, y) \sim \mathcal{S}} \left[ \ell(E_D(x), y) \right]$$

  where $y^2_t = \text{Total num of classes} + 1$

- The feature extractor of the target model, $E_T$, can be learned by optimizing:
  
  $$ \arg\min_{E_T} \mathbb{E}_{(x, y) \sim \mathcal{S}} \left[ \ell(E_T(x), y) \right]$$

  where $y^2_t = \text{#Total num of classes}$


Transfer Learning of Models across Sensors/Space/Time

Experimental Results

<table>
<thead>
<tr>
<th>May (Source)</th>
<th>July (Target)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Overall Acc (std)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Adaptation</td>
<td>100% (26%)</td>
</tr>
<tr>
<td>ADDA</td>
<td>100% (23%)</td>
</tr>
<tr>
<td>Ours (Unsupervised)</td>
<td>99% (62%)</td>
</tr>
<tr>
<td>Ours (Semi-supervised)</td>
<td>99% (74%)</td>
</tr>
</tbody>
</table>

Symbols definition:

- Feature extractor
- Discriminator
- Labels
- Unlabeled examples
Transfer Learning of Models across Sensors/Space/Time

Experimental Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Sensing/Imagery</th>
<th>Healthy grass</th>
<th>Stressed grass</th>
<th>Commercial</th>
<th>Highway</th>
<th>Railway</th>
<th>Parking lot 2</th>
<th>Mean Overall Acc (std)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No adaptation</td>
<td>36%</td>
<td>100%</td>
<td>51%</td>
<td>0%</td>
<td>95%</td>
<td>0%</td>
<td>47% (2)</td>
<td></td>
</tr>
<tr>
<td>Unsupervised ADDA</td>
<td>47%</td>
<td>97%</td>
<td>63%</td>
<td>49%</td>
<td>81%</td>
<td>0%</td>
<td>56% (7)</td>
<td></td>
</tr>
<tr>
<td>Ours (Unsupervised)</td>
<td>39%</td>
<td>100%</td>
<td>57%</td>
<td>0%</td>
<td>93%</td>
<td>0%</td>
<td>48% (6)</td>
<td></td>
</tr>
<tr>
<td>Semi-supervised ADDA</td>
<td>39%</td>
<td>97%</td>
<td>83%</td>
<td>77%</td>
<td>53%</td>
<td>37%</td>
<td>64% (9)</td>
<td></td>
</tr>
<tr>
<td>Ours (Semi-supervised)</td>
<td>61%</td>
<td>95%</td>
<td>80%</td>
<td>72%</td>
<td>69%</td>
<td>35%</td>
<td>68% (8)</td>
<td></td>
</tr>
</tbody>
</table>

Can GANs be made “attentive”?

- Preliminary work on an attention module within GANs

Transformers and Multi-Channel Image Analysis

Nuances relative to multi-channel image analysis

- CNNs are not effective at capturing long-range dependencies
- Frames/patch-sizes for typical CNNs used with moderate resolution multispectral/hyperspectral imagery are small (e.g. 5x5, 7x7 etc.)
- This limits the CNN design (e.g. depth/width of the network)
- Transformers on the other hand do not exhibit the inductive bias property naturally provided by CNNs

To address this, we are developing a hybrid CNN/Transformer based Spectral Transformer for multispectral/hyperspectral classification.
• Inspired by the Swin Transformer architecture, a hierarchical spectral Transformer is designed to assemble discriminative features from the spectral dimension.

• A CNN-mixer module is introduced to remedy the lack of inductive bias inherent in vision Transformers to capture local contextual features of HSI.
Going beyond Convolutional Neural Networks: Transformers and GeoAI

Interpretability (Preliminary results)

Transfer Learning of Models across Sensors/Space/Time

Prompting (Preliminary work)

Idea: Incremental model updates to address errors (Results below are with a Swin UNet and continual updates to the ground truth).

Implemented as a GUI to allow user interactions
Presentation Contents

• Background and Objectives
• Technical and Science Advancements
• Summary of Accomplishments and Future Plans
• Actual or Potential Infusions and Collaborations
• Publications - List of Acronyms

Summary of Accomplishments and Future Plans

• Current Accomplishments
  • Backbone Networks:
    • Incorporating Vision Transformers for robust GeoAI (Ongoing)
  • GAN based Transfer
    • Super-Resolution (Completed)
    • Model Transfer (Completed with basic networks, will be implemented with improved backbone networks)
• Future Plans
  • Enhancing Vision Transformers to incorporate spatial-spectral information and multi-modal information
  • Understanding and Addressing GAN Stability
  • Improving GAN based Model Transfer with Advanced Backbones
  • Semi-Supervised Knowledge Transfer and Prompting
  • Validating with Relevancy Scenarios
Presentation Contents

- Background and Objectives
- Technical and Science Advancements
- Summary of Accomplishments and Plans Forward
- Actual or Potential Infusions and Collaborations
- Publications - List of Acronyms

Actual or Potential Infusions and Collaborations

- **Technology Infusions**
  - Multiscale Sensing and Domain Adaptation
    - Agriculture Tech companies: Operate large scale plant breeding experiments over extended environments
    - U.S. Forest Service, USDA
    - Cloud Computing

- **Current Collaborations**
  - Wabash Heartland Innovation Network (WHIN) (https://whin.org/)
    - Fuse UAV airborne data with multi-county airborne data acquired by IntelinAir (https://www.intelinair.com/)
  - Bayer Crop Science and Corteva Agriscience
    - Domain adaption over extended regions for classification and prediction models

- **Potential Collaborations**
  - Amazon AWS / Microsoft / Planet
  - DoD
Presentation Contents

• Background and Objectives
• Technical and Science Advancements
• Summary of Accomplishments and Plans Forward
• Actual or Potential Infusions and Collaborations
• Publications

Publications

• Journal / Conference Papers

• Dissertations
  • Taojun Wang, Deep Neural Networks and Transfer Learning for Crop Phenotyping using Multi-modality Remote Sensing and Environmental Data, Purdue University, 2023
A hosted analytic collaborative framework for global river water quantity and quality from SWOT, Landsat, and Sentinel-2

Colin Gleason (PI, UMass)
Subhransu Maji (Co-I, UMass)
Suresh Vannan (Co-I, JPL)
Nikki Tebaldi (Co-I, JPL)
John Gardner (Co-I, Pitt)
Tamlin Pavelsky (Co-I, UNC)

AIST-21-0037 Annual Technical Review
July 21 2023

Objective

- Create a novel Analytic Collaborative Framework (ACF) for global rivers, integrating SWOT (RADAR), Landsat, and Sentinel 2 (Optical) remote sensing datasets
- Improve the accuracy of water quality datasets
- Create an entirely cloud-based workflow to generate water quality datasets for rivers produced from harmonized (RADAR + Optical) imagery.
- Transition the ACF to JPL/PO.DAAC for sustained operation with engineering support
- Open source all software

Approach

- Build on existing ACF hosted in AWS
- Dockerize all components for cross platform compatibility
- Pair subject matter experts with software engineers
- Include transition partner PO.DAAC from day 1
- "Hub and spokes" task management

Key Milestones

1. SWOT Launch 11/22
2. Setup AWS Cloud environment organized around global rivers 1/23
3. Prototype data production workflow 08/23
4. Hydrologic sediment algorithm embedded and tested 03/24
5. Global computer vision embedded and tested 08/24
6. Full transition of ACF to PO.DAAC 01/25
7. Software open sourced and released on public github 07/25

TRLIni = 3 TRLCurrent = 5

Co-Is/Partners: Tamlin Pavelsky, UNC, John Gardner, Pitt, Suresh Vannan JPL, Subhransu Maji, Nikki Tebaldi, UMass
Presentation Contents

- Background and Objectives
- Technical and Science Advancements
- Summary of Accomplishments and Future Plans
- Actual or Potential Infusions and Collaborations
- Publications - List of Acronyms

River discharge: one of SWOT’s revolutions

Per 21 days

Altenau et al., WRR, 2021
At each river, we calculate discharge when SWOT was overhead

SWOT does not measure discharge; we must calculate it at this global scale to meet mission requirements

Total Reaches: 213,485

Discharge: the basic idea

Severely underconstrained inverse problem
Rivers are more than just water

Sediment:
- Erosion, deposition, delta formation, scour, siltation, nutrient transport, hydropower efficiency...

Co-I Gardner et al., 2023

Optical satellites can retrieve SSC:

Sediment suspended in the water column

Co-I Gardner et al., 2023 ERL
Rivers are more than just water

Limitations of existing hydrology approaches:
Preselected features, radiometric calibration, inability to detect sources or sinks: point based estimates from space

Background / Objectives

- Combining SWOT’s water quantity with optical water quality will lead to unprecedented monitoring and science
- Computer Vision to improve hydrologic water quality algorithms
- AWS based computational engine - massive scale
- PODAAC/LPDAAC data backends - SWOT and optical data
- PODAAC hosting data outputs

‘Confluence’
Presentation Contents

- Background and Objectives
- Technical and Science Advancements
- Summary of Accomplishments and Future Plans
- Actual or Potential Infusions and Collaborations
- Publications - List of Acronyms

What is Confluence?

**SWOT**

- The Public
- Confluence: AIST software
- SwoRD at UNC
- SDS at JPL
- PO.DAAC
- Discharge Data product
- The SwoRD of Science
- Hydrologic data products not included in the PDD

*SWOT: SWOT a priori River Database
SDS: Science Data System
ST: Science team
E.g., Manning’s n, Chezy friction factor*
Confluence

SWOT lives!

All modules at right have been tested on calval orbit

All run in AWS-completely cloud based operation

Confluence

- Water quantity high initial TRL- 10+ years of legacy and SWOT funding

- Water quality (SSC) low initial TRL-proven in theory but not ready for the cloud

- CV is mature, but hasn’t been applied to this problem
From shallow to deep learning: CV

Hydro SSC: Works!
- Preselects features
- Point-based
- ‘Shallow’ ML

CV SSC: Should work
- Feature discovery
- Context-based
- ‘Deep’ ML

Confluence will deploy both: community value

First: benchmark Hydro SSC at scale in cloud

Software target

Replacing with MobileNetv3 improves speed 30% with no loss of accuracy

From shallow to deep learning: CV
### From shallow to deep learning: CV

Our water masking model predicts water pixels more accurately with 30% speedup in prediction time

<table>
<thead>
<tr>
<th></th>
<th>Baseline (DeepWaterMap)</th>
<th>Ours (MobileNet)</th>
</tr>
</thead>
<tbody>
<tr>
<td># Trainable Parameters (↓)</td>
<td>37,200,150</td>
<td>1,138,486</td>
</tr>
<tr>
<td>Inference Time with CPU (↓)</td>
<td>82.69 s</td>
<td>63.80 s</td>
</tr>
<tr>
<td>F1 Score (↑)</td>
<td>91.42%</td>
<td>91.98%</td>
</tr>
<tr>
<td>Precision (↑)</td>
<td>95.50%</td>
<td>94.19%</td>
</tr>
<tr>
<td>Recall (↑)</td>
<td>88.16%</td>
<td>89.86%</td>
</tr>
</tbody>
</table>

↓ means lower is better
↑ means higher is better

---


**From shallow to deep learning: CV**

Our water masking model predicts water pixels more accurately with 30% speedup in prediction time

![Image of water masking model predictions]

---

**Backend advances**

- Integration with Fargate SPOT for substantially reduced processing costs
- Allows for mid module checkpointing making the pipeline robust to data irregularities and unforeseen edge cases
- Straightforward deployment using a GUI or IoC methods
- Both CV and Hydro SSC models will be parallelized at the tile level
Presentation Contents

• Background and Objectives
• Technical and Science Advancements
• Summary of Accomplishments and Future Plans
• Actual or Potential Infusions and Collaborations
• Publications - List of Acronyms

Summary of Accomplishments and Future Plans

• We're slightly ahead of schedule and have resources and personnel in place

• SWOT has launched- we’re awaiting ‘reprocessed’ data to see if our water quantity engine behaves as expected
  • Currently operational with SWOT data, but not meaningful

• We’ll produce global SSC from both hydrology and CV approaches in the next year
Summary of Accomplishments and Future Plans

• Future computer vision problems will additionally aim to do the following:
  • Identify sources and sinks of sediments using the TSS predictions as basis
  • Interpolate TSS predictions in space and time for missing data points (possibly due to cloud cover or data collection limitations)

• Further improving efficiency and runtime in full sediment prediction pipeline for deployment

Technical and Science Advancements in Year 1

OPEN SOURCE PIPELINE BENCHMARK VALIDATE DEPLOY

SWOT Q

Hydro SSC

CV SSC

PODAAC ARCHIVAL

ROBUST CLOUD SOFTWARE
Actual or Potential Infusions and Collaborations

• PO.DAAC integration underway - we’re pulling data from PO.DAAC and they’re preparing to ingest the SoS

• Will present Confluence at the SWOT Science Team meeting Sep. 2023 in Toulouse to the SWOT mission team

• We’re supporting SWOT calval by testing SWOT data through Confluence

• Approached about including dissolved CO₂, Nitrogen, and Phosphorous
  • Would make Confluence a ‘river information system’
List of Acronyms

- SWOT- Surface Water and Ocean Topography mission
- AWS – Amazon web services
- SSC- Suspended Sediment Concentration
- CV- computer vision
- PODAAC/LPDAAC – NASA distributed data access and archive centers
- SoS- SWORD of Science
- SWORD- SWOt River Database
- Calval- Calibration and validation
- IoC- Infrastructure as Code
Deep learning for Environmental and Ecological Prediction, eValuation and Insight with Ensembles of Water quality

(DEEP-VIEW)

PI: Stephanie Schollaert Uz\textsuperscript{1},
Troy Ames\textsuperscript{1}, J. Blake Clark\textsuperscript{1,2}, Dirk Aurin\textsuperscript{1,3}

\textsuperscript{1}NASA Goddard Space Flight Center
\textsuperscript{2}University of Maryland Baltimore County, \textsuperscript{3}Morgan State University

How can we improve coastal water quality data products to address societal challenges locally?

- Challenge – large estuary, increased land use and runoff, increased aquaculture and recreation while water quality management resources shrinking, maximize potential of upcoming NASA hyperspectral missions
- Interagency Chesapeake Bay working group activities increase coordination, collaboration on data collection, calibration activities, new research
- DEEP-VIEW machine learning development to fuse features detected through multiple sources, including hyperspectral satellites, to capture greater variability
Motivation: resource manager need

<table>
<thead>
<tr>
<th>Parameter name</th>
<th>Water Quality Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fecal coliform</td>
<td>&lt;14 MPN median per 100ml</td>
</tr>
<tr>
<td>Bacteriological Escherichia coli</td>
<td>&lt; 410 count per 100ml</td>
</tr>
<tr>
<td>Dissolved oxygen</td>
<td>&gt; 5 mg/l</td>
</tr>
<tr>
<td>Temperature</td>
<td>&lt; 90°F/32°C</td>
</tr>
<tr>
<td>pH</td>
<td>6.5 - 8.5</td>
</tr>
<tr>
<td>Turbidity</td>
<td>&lt;150 nephelometer turbidity units</td>
</tr>
<tr>
<td>Color</td>
<td>&lt; 75 platinum cobalt units</td>
</tr>
<tr>
<td>Water clarity</td>
<td>&gt; 13% (tidal fresh)</td>
</tr>
</tbody>
</table>

Maryland shellfish harvesting water quality criteria

- Can we improve coastal remote sensing to assist resource managers?
- Exploring relationship between satellite data and classification labels: temperature, turbidity, phytoplankton pigments, pollutants

Integration of Observations and Models into Machine Learning for Coastal Water Quality

PI: Stephanie Schollaert Uz, NASA GSFC

Objective
- Continue development and validation of modular framework to integrate data from multiple satellites and models to identify water quality problem areas in the Chesapeake Bay.
- Initial performance goals are >80% accuracy for detection of poor water quality (exceeding thresholds for indicators, e.g. turbidity, harmful algal blooms, pollutants).
- Technology includes feature extraction by machine learning with multiple satellite sensors, physical models, and in situ sampling.
- Improved capability prepares to exploit hyperspectral sensing by future NASA missions, i.e. PACE GLIMR, SBG

Approach
- Apply NASA data, science, and technology to support interagency partners (e.g. state agencies, NOAA) in their operations toward the development of a decision support tool for shellfish aquaculture:
  1. Collect and analyze all available in situ and remotely sensed data relevant to Chesapeake Bay water quality.
  2. Collect and analyze absorption and fluorescence properties of water constituents at hyperspectral resolution for select sites.
  3. Train an ACF ML to identify features that resulted in shellfish bed closures.
  4. Refine and validate the ML against current conditions.

Co-Is/Partners: Troy Ames, GSFC; Blake Clark, UMBC/GSFC; Marjorie Friedrichs, VIMS; Chris Brown, NOAA; John McKay, MDE

Key Milestones
- Create interface to data modules (12/22)
- Optimize feature encoders (01/23)
- Develop feature fusion module (03/23)
- Optimize temporal fusion (06/23)
- Machine Learning validation (TRL 5) (06/24)
- Transition ML for Bias Correction (06/25)

Modular framework for detecting water quality features from multi-sensor segmentation using remote sensing and in situ data

TRL_{in} = 3  TRL_{current} = 4
Satellite Sensors
- Landsat 8 OLI
- Sentinel 2 MSI
- Aqua-MODIS
- Sentinel 3 OLCI
- DESIS
- PRISMA

In-Situ Data
- Dissolved Oxygen
- Organics
- Phytoplankton
- Harmful Algae
- High Bacteria/fisheries impacts
- Physics

Multi-sensor and spectral agnostic ML Architecture trained on in-situ and satellite data

Data QA and QC and historical analysis

*Uncertainty quantification in in-situ and satellite data ongoing

Feed into ML Model

Identification of features that relate to poor water quality
1. Single threshold variables
2. Multi-variable threshold
3. Presence of human health related microbes

Integration with additional causal earth system variables from many observations

Quantify Uncertainty for Error Accounting and Propagation

Set up assimilation capability with VIMS CBEFS Model for initial testing. Develop methodology for using combined ML and biogeochemical model for ecological forecast purposes.

*Completed
*Actively in development

AIST18: initially developed ML using satellite data and process model

Using 3-D Virginia Institute of Marine Science (VIMS) model as label data, initial architecture trained on optical satellite data input.

a) Target vs. predicted for every image bin
b) Surface predicted dissolved oxygen
c) Surface target dissolved oxygen
d) Predicted vs target vertical contour from the center of map
e) Predicted cross-section values at depth
f) Target cross-section at same location

Shellfish harvesting threshold: DO < 5mg/L
Deep learning for Environmental and Ecological Prediction, Evaluation and Insight with Ensembles of Water quality (DEEP-VIEW)

Validating Hyperspectral DESIS Satellite Data

Atmospherically correcting DESIS hyperspectral surface reflectances in Acolite using https://github.com/acolite/acolite

Once these are validated with in situ data (from AERONET-QC and boat), integrate into new machine learning architecture

- 235 bands from 400-1000 nm
- 30 m resolution
- 2020-2022 (156 images)
Unsupervised Feature Training

- Unsupervised learning using Autoencoder architecture on spectral data (2003-2021)
- Training sets consisting of 50K, 100K, 250K water pixels
- Trained encoder can then be used for training additional decoder(s) to predict water quality.

Original spectrum compared to predicted spectrum from 8-feature vector (inset)

Supervised Feature Training using In situ data

- Supervised learning using decoder module on trained spectral features
  - Correlate learned features with in situ data within 2-6 hours (weighted)
  - Thousands of in-situ matchups
- Utilize longer history of MODIS data and in-situ matchups for transfer learning to other satellite sources
MODIS/In situ Trained ML Output vs In situ Data

- **KD**
  - $R^2 = 0.598$
  - Samples: 339

- **CHLA**
  - $R^2 = 0.486$
  - Samples: 216

- **SECCHI**
  - $R^2 = 0.689$
  - Samples: 432

- **TSS**
  - $R^2 = 0.604$
  - Samples: 219

- **CDOM_440**
  - $R^2 = 0.671$
  - Samples: 19

- **DO**
  - $R^2 = 0.269$
  - Samples: 202

- **SALINITY**
  - $R^2 = 0.517$
  - Samples: 204

- **TP**
  - $R^2 = 0.562$
  - Samples: 223

**DESIS Satellite Data Unsupervised/Supervised Feature Training**

- Currently only 8 scenes have in situ matchups
- Unsupervised learning using Autoencoder architecture on hyperspectral pixel data
- Supervised learning using MODIS trained model
- Exploit longer history of MODIS data and in-situ matchups for transfer learning to DESIS model

**Original spectrum (177 channels) vs reconstructed spectrum from an 8-feature vector (inset)**
Kriging In Situ Fields to Increase Labeled Data for ML

Kriging method uses observations weighted by distance in monthly gridded climatology

- Download CBP Data, sort by location, date, and extract surface values
- Fit three bounded theoretical (model) variogram functions and find best model using $R^2$
- Calculate semi-variogram using water-distance and CBP data grouped by month
- Calculate spatially distributed daily values using CBP observations, theoretical variogram, and kriging. Hold out 5% for test
- Assign each CBP station to a triangular mesh point for water-distance based distance matrix
- Assess model skill with 5% hold out test data. Mask points outside of auto-correlation range beyond spatial limits of model utility
- Combine with satellite remote sensing reflectance

MODIS In-situ Training + Kriging Interpolated Values

Experimental (semi) variograms and theoretical variogram models
MODIS In-situ Training + Kriging Interpolated Values

**Kd 2002-2022**
- 547 Individual Days
- Average 15 samples per day
- 8211 input observations
- 554 test observations

**Secchi Depth**
March 14, 2019
- Kriging product masked by ½ of Sill variance (strict)
- ML prediction using individual image from the same day pixels matched within 50 m
MODIS In-situ Training + Kriging Interpolated Values

**Secchi depth and Kd for four initial images**

- Upper plots are ML predictions vs. observations each day
- Lower plots are ML predictions vs. Kriging predictions aggregated over the four images

**Summary:** many preparatory activities are reducing data gaps, quantifying uncertainties, developing DEEP-VIEW to serve resource managers as well as providing a new method for exploiting upcoming hyperspectral satellite data more broadly, e.g. process model assimilation

- Challenges remain – land adjacency for area of greatest interest, clouds and atmospheric correction of satellite data, sparse matchups within 2 hours
- Methodology developed here will be transitioned to open science cloud for interdisciplinary, e.g. land-water research, and scaling to other locations
- Transition from ADAPT to SMCE cloud services
Back up slides

## Project Schedule

<table>
<thead>
<tr>
<th>Task</th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Create Interface to data modules</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Satellite data processing</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Optimize Feature Encoders</td>
<td>v1</td>
<td>v2</td>
<td>v3</td>
<td>v4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prioritize features with MDE,VDH</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Continue sampling and analysis</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Develop Feature Fusion Module</td>
<td>v1</td>
<td>v2</td>
<td>v3</td>
<td>v4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Optimize Temporal Fusion</td>
<td>v1</td>
<td>v2</td>
<td>v3</td>
<td>v4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ML validation</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assess TRL, review, report</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discussions with VIMS team</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Transition DEEP-VIEW to VIMS</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Analyze and optimize transition</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DEEP-VIEW into MARACODS CBEFS for Bias Correction</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>
**Additional Team Members: Students and Postdocs**

**Shannon McDonnell**, defended Ph.D. spring 2023, University of Maryland; chemistry sampling and optical characterization (field and lab) – AIST18

**Samantha Smith**, Student intern, summer 2021, spring 2022, BS, UMD, started on SSAI contract fall 2023. Chemistry and data analysis of in situ and satellite scenes for optical indicators of run-off *will present at Interagency Chesapeake Bay workshop at GSFC (Jul, 2023), IOCS conference in St. Petersburg, FL (Nov, 2023)

**William Daniels**, Student intern, summer 2022, fall 2022, Masters student, Northwestern University; Computer Science – Machine learning development. Paper accepted to IGARSS 2023, delivering in person oral presentation in Pasadena, CA.

**Nathan Duchez**, Student intern, summer 2023, Undergraduate, USNA, Electrical Engineering – automating in situ data graphics and data transmission to the cloud for near-real-time access by ML, presenting at Interagency Chesapeake Bay workshop at GSFC (Jul, 2023)

**Morgaine McKibben**, NPP, Complementary project, ‘Optical discrimination of phytoplankton taxa in estuarine waters: application of a hyperspectral, bio-optical algorithm to aid Chesapeake Bay resource managers’ presented at EGU 2023 in Vienna, Austria (Apr, 2023), Interagency Chesapeake Bay workshop at GSFC (Jul, 2023), IOCS conference in St. Peters berg, FL (Nov, 2023)

---

**Journal / Conference Papers**


**Dissertations**


**Other**
### Acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADAPT</td>
<td>Advanced Data Analytics PlaTform</td>
</tr>
<tr>
<td>CBEFS</td>
<td>Chesapeake Bay Environmental Forecast System</td>
</tr>
<tr>
<td>CBP</td>
<td>Chesapeake Bay Program</td>
</tr>
<tr>
<td>CDOM</td>
<td>Colored Dissolved Organic Matter</td>
</tr>
<tr>
<td>CNN</td>
<td>Convolutional Neural Network</td>
</tr>
<tr>
<td>CSOSIP</td>
<td>Commercial Smallsat Data Acquisition Program</td>
</tr>
<tr>
<td>DEEP-VIEW</td>
<td>Deep learning for Environmental and Ecological Prediction, Evaluation and Insight with Ensembles of Water quality</td>
</tr>
<tr>
<td>DESIS</td>
<td>DLR (German Space Agency) Earth Sensing Imaging Spectrometer</td>
</tr>
<tr>
<td>EIS</td>
<td>Earth Information System</td>
</tr>
<tr>
<td>HAB</td>
<td>Harmful Algal Bloom</td>
</tr>
<tr>
<td>HICO</td>
<td>Hyperspectral Imager for the Coastal Ocean</td>
</tr>
<tr>
<td>LSTM</td>
<td>Long Short Term Memory</td>
</tr>
<tr>
<td>MODIS</td>
<td>Moderate-resolution Imaging Spectrometer</td>
</tr>
<tr>
<td>MSI</td>
<td>Multispectral Imager</td>
</tr>
<tr>
<td>NCCS</td>
<td>NASA Center for Climate Simulations</td>
</tr>
<tr>
<td>NWQMC</td>
<td>National Water Quality Monitoring Council</td>
</tr>
<tr>
<td>OLCI</td>
<td>Ocean and Land Color Instrument</td>
</tr>
<tr>
<td>OLI</td>
<td>Operational Land Imager</td>
</tr>
<tr>
<td>PRISMA</td>
<td>(Italian) Hyperspectral Precursor of the Application Mission</td>
</tr>
<tr>
<td>Rhos</td>
<td>Top-of-atmosphere reflectance minus Rayleigh</td>
</tr>
<tr>
<td>Rs</td>
<td>Remote sensing reflectance</td>
</tr>
<tr>
<td>SAA</td>
<td>Space Act Agreement</td>
</tr>
<tr>
<td>SMCE</td>
<td>Science Managed Cloud Environment</td>
</tr>
<tr>
<td>SST</td>
<td>Sea-Surface Temperature</td>
</tr>
<tr>
<td>S2</td>
<td>Sentinel-2 A&amp;B</td>
</tr>
<tr>
<td>S3</td>
<td>Sentinel-3 A&amp;B</td>
</tr>
</tbody>
</table>
A Formal Study on Machine Learning (ML) and Geospatial Regression Methods for Processing Green House Gases (GHG) and Air Pollutants

Dr. Chaowei Yang (PI, George Mason University)
Dr. Thomas Huang (Co-I, NASA JPL)
Dr. Mohammad Pourhomayoun (Co-I, California State University)

AIST-QRS/23-0002 Annual Technical Review
07-21-2023

Team: Anusha Srirenganathan/GMU, Seren Smith/GMU, Dr. Qian Liu/GMU, Dr. Sina Hasheminassab/NASA JPL, Joe Roberts/NASA JPL, Kevin Marlis/NASA JPL
**Objective**

- Conduct a formal study on ML and geospatial methods for air pollutants simulation, retrieval or prediction using relevant training datasets
- Develop and configure a ML open-source package including recommended optimum model parameters and configurations
- Integrate this package in NASA AIST’s Air Quality Analytic Collaborative Framework (AQACF) and FireAlarm Digital Twin
- Prepare pollutant dataset including well-quantified uncertainties
- Engage high-school and community college students, esp. from underserved communities.

**Approach**

- Formalize a list of parameters for tuning models
- Prepare a suite of training datasets for training ML algorithms for Methane (CH₄), Ozone (O₃), PM₂.₅, SO₂, and NO₂ analysis
- Formalize a list of methods for uncertainty quantification and accuracy assessments
- Benchmark the various methods on relevant computing environments for simulation, retrieval and prediction
- Analyze the results to identify the optimal model parameters for models
- Integrate the package with the AQACF and FireAlarm systems
- Release the package as open source for open science initiative

**Key Milestones**

- Formalize models, parameters and uncertainty 07/23
- Assemble training datasets including quality control and calibration matrix 08/23
- Complete the models/pollutants study 12/23
- Optimize model parameters and configurations 02/24
- Integrate the package with AQACF and with the FireAlarm Digital Twin 04/24
- Release the package Open Source 05/24

**Architecture and Workflow**

**Co-Is/Partners:** M. Pourhomayoun, CSU-LA; T. Huang, JPL; A.S. Malarvizhi, and S. Smith, J. Liu, Q.Liu, GMU
Presentation Contents

• Background and Objectives

• Technical and Science Advancements

• Summary of Accomplishments and Future Plans

• Actual or Potential Infusions and Collaborations

• Publications - List of Acronyms
Background and Objectives

Background:
• Air pollution is a global environmental issue with detrimental effects on human health and the environment, such as killing 7M human lives each year.
• **Accurate and reliable assessment** of air pollutant concentrations is essential for effective air quality management and policy-making.
• **High spatiotemporal resolution air quality retrieval and analysis data** are essential for investigating air quality patterns and better mitigation strategies.
• AI/ML models are increasing used in air pollutants and greenhouse gases analytics
• There is **no systematic study on how to best use AI/ML models in the analytics**
• A lack of professionals in this specific field of earth science information technology

Objectives:
• Conduct **a formal study** on ML and geospatial methods for air pollutants simulation, retrieval or prediction using relevant training datasets
• Develop and configure **a ML open-source package** including recommended optimum model parameters and configurations
• **Integrate this package** in NASA AIST’s Air Quality Analytic Collaborative Framework (AQACF) and FireAlarm Digital Twin
• Prepare **pollutant/ GHG (training) dataset** including well-quantified uncertainties
• Engage **high-school and community college students**, esp. from underserved communities.
Presentation Contents

• Background and Objectives
• Technical and Science Advancements
• Summary of Accomplishments and Future Plans
• Actual or Potential Infusions and Collaborations
• Publications - List of Acronyms
Technical and Science Advancements

- Uncertainty Quantification and Error Representation
- Dataset Outlier Removal
- Sensor Data Calibration
- Multi-source pollutants, GHG, and AOD Fusion and Downscaling/Upscaling
- Prepare high quality training datasets
- PM$_{2.5}$ Retrieval
Uncertainty Quantification

- AI/ML models are subject to noise and model inference errors, it is crucial to evaluate their reliability. Representing **uncertainty** in a trustworthy manner is crucial for any AI/ML models.
- Sources of uncertainty
  - Data: Noisy and incomplete data.
  - Physical models: Misspecification and stochasticity
  - Neural networks: Architecture, hyperparameters, and overparametrization
  - Posterior inference: Uncertainty arises from the process of posterior inference.
- Two types of uncertainty
  - **Aleatoric Uncertainty (AU)** - Arises from noisy data and cannot be reduced.
  - **Epistemic Uncertainty (EU)** - Arises from noisy and limited data, as well as neural network overparametrization.

Predictive Uncertainty (PU) can be given as the sum of Aleatoric Uncertainty (AU) and epistemic Uncertainty (EU)

$$PU = AU + EU$$
Uncertainty Modeling

• Construct a distribution \( p(u|x, D, H) \) to predict the value of \( u \) at any new location \( x \), where
  – \( D = \{(x_1, u_1), (x_2, u_2), \ldots, (x_N, u_N)\} \) is paired noisy observations
  – \( H \) is model

• Aleatoric uncertainty
  – \( u(x) \) produced by the data-generating process contains both a deterministic part \( u_c(x) \) as well as some additive noise \( \varepsilon_u \)
    \[
    u(x) = u_c(x) + \varepsilon_u
    \]
  – The factorized Gaussian likelihood function to construct a model \( u_\theta(x) \) that captures the deterministic part as well as noise
    \[
    p(u|x, \theta) = \mathcal{N}(u|u_\theta(x), diag(\Sigma_u^2)) = \prod_{d=1}^{D_u} \frac{1}{\sqrt{2\pi\sigma_u^2}} \exp\left(-\frac{(u_d - u_\theta(x)_d)^2}{2\sigma_u^2}\right)
    \]

• Epistemic uncertainty
  – Uncertainty regarding the NN parameters \( \theta \)
  – Using Bayes’ rule, the posterior \( p(\theta|D) \) is given as
    \[
    p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}.
    \]

Source: Abdar, M et al. (2021), Psaros, A. F. et al. (2023)
# Methods for Uncertainty Quantification

- **Uncertainty quantification using Bayesian techniques**

<table>
<thead>
<tr>
<th>Method name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monte Carlo (MC) dropout</td>
<td>Standard training with NN parameters set to zero in each step</td>
</tr>
<tr>
<td>Markov chain Monte Carlo (MCMC)</td>
<td>Iterative sampling for approximate inference</td>
</tr>
<tr>
<td>Variational Inference (VI)</td>
<td>Optimization-based approximation of the posterior distribution</td>
</tr>
<tr>
<td>Laplace Approximation (LA)</td>
<td>Approximating a distribution with a Gaussian near its mode</td>
</tr>
</tbody>
</table>

- **Uncertainty quantification using Ensemble techniques**

<table>
<thead>
<tr>
<th>Method name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deep ensembles (DEns)</td>
<td>Training multiple networks for improved predictions and uncertainty estimation.</td>
</tr>
<tr>
<td>Snapshot ensembles (SEns)</td>
<td>Training and ensembling models at different epochs for improved predictions and uncertainty estimation.</td>
</tr>
<tr>
<td>Stochastic weight averaging (SWA) and SWA-Gaussian (SWAG)</td>
<td>leveraging stochastic weight averaging and Gaussian approximation</td>
</tr>
</tbody>
</table>

*Source: Abdar, M et al. (2021), Psaros, A. F. et al. (2023)*
## Metrics for Uncertainty Quantification

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Quality Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Root Mean Squared Error (RMSE)</td>
<td>Measures prediction accuracy, and the value approximates the average deviation of the predicted values from the measured values</td>
</tr>
<tr>
<td>R-squared ($R^2$)</td>
<td>Indicates how well the model captures the variability in the data</td>
</tr>
<tr>
<td>Average Standard Error (ASE)</td>
<td>Measures model precision, a tendency to produce narrow predictive distributions closely centered around the predicted value</td>
</tr>
<tr>
<td>Mean Standardized Error (MSE)</td>
<td>Measures model bias on a standardized scale so that it is comparable across datasets with different values and units</td>
</tr>
<tr>
<td>Mean Predictive Likelihood (MPL)</td>
<td>Measures predictive capability</td>
</tr>
<tr>
<td>Relative l2 Error (RL2E)</td>
<td>Measures accuracy</td>
</tr>
<tr>
<td>Root Mean Squared Calibration Error (RMSCE)</td>
<td>Measures statistical consistency between predictions and data</td>
</tr>
<tr>
<td>Prediction Interval Width (PIW)</td>
<td>Measure prediction sharpness and dispersion</td>
</tr>
<tr>
<td>SD Coefficient of Variation (SDCV)</td>
<td></td>
</tr>
<tr>
<td>Normalized Inner Product between Variances (NIP-G); KL divergence between distributions (KL-G)</td>
<td>Measure the inner product between the predicted variance and the variance of gold standard</td>
</tr>
</tbody>
</table>

**Source:** Abdar, M et al. (2021), Psaros, A. F. et al. (2023)
PurpleAir Sensor Data Outlier Removal

Steps to remove outliers

1. **Identification of Malfunctioning Sensors:** Use 5-hour moving standard deviation (SD) to assess sensor performance. Discarded records with 5-hour SD of zero.

   \[
   SD(t-n, t+n) = \frac{\sqrt{T}}{2n+1} \sum_{i=1}^{2n+1} (x_i - \mu)^2
   \]

2. **Discard apparent outliers:** Remove minute-level values exceeding, e.g., 500 \(\mu g/m^3\).

3. **Filtering using Mean Absolute Deviation (MAD):** Calculate MAD by channel within a calendar month and define threshold to discard outliers.

   \[
   \text{MAD} = b \times \text{median} \left( |X_i - \bar{X}| \right)
   \]
   \[
   X_i < \bar{X} - 3 \times \text{MAD} \text{ or } X_i > \bar{X} + 3 \times \text{MAD}
   \]

4. **Evaluate dual-channel agreement within a given month:** Use coefficient of determination \((R^2)\), mean absolute error (MAE), and mean absolute percentage error (MAPE) as statistical indicator and set thresholds, respectively.

   \[
   \text{MAE} = \frac{\sum_{i=1}^{n} |PM2.5_A - PM2.5_B|}{n} ; \text{MAPE} = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{PM2.5_A - PM2.5_B}{PM2.5_A} \right|
   \]

**Source:** Lu, Y et al. 2021
PurpleAir Outlier Removal

Location: CA, Rancho Dominguez
Time: 2017-11-28 ~ 2022-03-15

Raw data:
- PearsonR: 0.18
- R square: 0.03
- data points: 27266

1. Maximum value control:
- PearsonR: 0.73
- R square: 0.54
- data points: 27241

2. Remove single channel:
- PearsonR: 0.73
- R square: 0.54
- data points: 27240

3. Remove malfunction:
- PearsonR: 0.76
- R square: 0.57
- data points: 27174

4. Dual channel agreement:
- PearsonR: 0.77
- R square: 0.59
- points: 26463
Sensor Data Calibration

\[ PM_{2.5,\text{calibrated}} = a_0 + a_1 PM_{2.5,\text{raw}} + a_2 PM^2_{2.5,\text{raw}} + a_3 T + a_4 RH \]

1. \( PM_{2.5,\text{raw}} = PM_{\text{CF}_1} \), particulate matter in cubic feet
2. \( T \) = temperature, in Fahrenheit
3. \( RH \) = relative humidity
4. \( a_0, a_1, a_2, a_3, a_4 \) = scaling factors, vary on location in the United States

*This calibration algorithm is typically applied via a linear regression, however, using ML we will explore further possibilities.*
Accomplishments: Purple Air Calibration

Results are based on the outlier removed PurpleAir PM$_{2.5}$

<table>
<thead>
<tr>
<th>Model type</th>
<th>Train-test split</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>80%-20%</td>
</tr>
<tr>
<td>Linear Regression</td>
<td>0.79</td>
</tr>
<tr>
<td>Decision Tree Regressor</td>
<td>0.66</td>
</tr>
<tr>
<td>Random Forests</td>
<td>0.81</td>
</tr>
<tr>
<td>k-Nearest Neighbors</td>
<td>0.81</td>
</tr>
<tr>
<td>XGBoost</td>
<td>0.82</td>
</tr>
<tr>
<td>Ordinary Least Square Regression (OLS)</td>
<td>0.71</td>
</tr>
<tr>
<td>Lasso Regression</td>
<td>0.73</td>
</tr>
</tbody>
</table>

Why did XGBoost perform the best?:
- Gradient boosting on decision trees
- Usage of lasso and ridge regularizations
- Minimizes loss function
- Effective at parallelization

\[
obj(\theta) = \sum_{i}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)
\]
Accomplishments: ML models and XGBoost Systematic Study (Seren)

- Ran a 2939-run sweep, calculating RMSE, $R^2$, and other parameters
- Best score with raw data: $R^2$ of 0.8211, RMSE of 3.538
- Found most impactful factors aside from PM_CF_1: Humidity, then Temperature

Name: worldly-sweep-137
eta: 0.1
max_depth: 8
n_estimators: 90
rmse: 3.538
r2_score: 0.8211
Accomplishments: Hyperparameter Tuning (Seren)

- **XGBoost**
  - *eta*: the learning rate/shrinkage parameter, controls step size of trees
  - *max_depth*: the maximum depth for any individual decision tree while boosting
  - *n_estimators*: the number of boosting rounds performed during training

- **PyTorch**:
  - *epochs*: the number of times the model will iterate over the training dataset
  - *batch_size*: number of data samples processed in a single pass
  - *learning_rate*: the step size

- **Tensorflow**:
  - Like above, utilizes *learning rate, batch size, epochs*

Name: worldly-sweep-137
t
eta: 0.1
max_depth: 8
n_estimators: 90
rmse: 3.538
r2_score: 0.8211

Parameter importance with respect to r2_score

<table>
<thead>
<tr>
<th>Config parameter</th>
<th>Importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>eta</td>
<td></td>
</tr>
<tr>
<td>max_depth</td>
<td></td>
</tr>
<tr>
<td>n_estimators</td>
<td></td>
</tr>
</tbody>
</table>

Top: displays XGBoost run details with highest \( R^2 \)
Bottom: shows importance of each config parameter
Multisource AOD data fusion and downscaling

<table>
<thead>
<tr>
<th>Data</th>
<th>Satellite</th>
<th>Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Satellite derived AOD</td>
<td>Terra and Aqua MODIS MAIAC, GOES-16</td>
<td>AOD at 550 nm</td>
</tr>
<tr>
<td>Assimilated AOD</td>
<td>MERRA-2</td>
<td>total AOD at 550 nm and column mass density of black carbon, dust, organic carbon, sea salt, sulfate</td>
</tr>
<tr>
<td>Meteorological variables</td>
<td>ERA-5 global reanalysis</td>
<td>Planetary Boundary layer Height (BLH), surface-level temperature, total precipitation, surface air pressure, and near-surface wind speeds at eastward/northward components</td>
</tr>
<tr>
<td>Daily Cloud Fraction</td>
<td>Terra and Aqua MODIS</td>
<td>Cloud Fraction</td>
</tr>
</tbody>
</table>

1. Imputation model to predict missing values of daily Aerosol Optical Depth (AOD)

\[
A_m(s_p, d_i) = f(Z(s, d_i), X(s_i), Y(s_i), XY(s_i), DOY(d_i), DOW(d_i))
\]

where \(Z(s_i, d_i)\) is a set of spatiotemporal covariates

2. Using multi-sourced AOD, spatiotemporally downscale the fused AOD

3. Calibrate and validate the fused AOD to reduce bias using AERONET
### NO$_2$, SO$_2$, and O$_3$ Data Processing

<table>
<thead>
<tr>
<th>Data</th>
<th>Satellite</th>
<th>Input Resolution</th>
<th>Processes</th>
<th>Result Resolution</th>
</tr>
</thead>
<tbody>
<tr>
<td>NO$_2$ Tropospheric Column Density (S5P_L2__NO2___HiR)</td>
<td>Sentinel-5 TROPOMI</td>
<td>5.5 km x 3.5 km x 101.5 minutes (Level 2 Data)</td>
<td>1. Retrieve using GES DISC Subsetter</td>
<td>0.5 km x 0.5 km (1-3 observations per day for study area)</td>
</tr>
<tr>
<td>SO$_2$ Tropospheric Column Density (S5P_L2__SO2___HiR)</td>
<td></td>
<td></td>
<td>2. Resample to uniform grid</td>
<td></td>
</tr>
<tr>
<td>Ozone Total Column (S5_L2__O3_TOT_HiR)</td>
<td></td>
<td></td>
<td>3. Interpolate to increase spatial resolution</td>
<td></td>
</tr>
</tbody>
</table>
Presentation Contents

• Background and Objectives

• Technical and Science Advancements

• Summary of Accomplishments and Future Plans
  • GHG list and analytical progress
  • Uncertainty quantification
  • In-situ data collection and service
  • PM$_{2.5}$ Data outlier removal and calibration
  • ...

• Actual or Potential Infusions and Collaborations

• Publications - List of Acronyms
## GHG and Air Pollutant Progress

<table>
<thead>
<tr>
<th>Pollutants</th>
<th>Source Identification</th>
<th>Obtain &amp; Preprocess</th>
<th>Training Dataset</th>
<th>Analytics &amp; Model Training</th>
<th>Optimum parameter identification</th>
</tr>
</thead>
<tbody>
<tr>
<td>PM$_{2.5}$</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>NO$_2$</td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>SO$_2$</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ozone (O$_3$)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Methane (CH$_4$)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Model and Platform Preparation

<table>
<thead>
<tr>
<th>Model Type</th>
<th>RStudio is compatible?</th>
<th>PyTorch is compatible?</th>
<th>TensorFlow is compatible?</th>
<th>Sci-Kit is compatible?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear Regression</td>
<td>y</td>
<td>y</td>
<td>y (deprecated)</td>
<td>y (needs Euclidean coordinates)</td>
</tr>
<tr>
<td>Decision Tree Classifier</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>y</td>
</tr>
<tr>
<td>Decision Tree Regressor</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>y</td>
</tr>
<tr>
<td>Support Vector Machines</td>
<td>y</td>
<td>y</td>
<td>y (deprecated)</td>
<td>y</td>
</tr>
<tr>
<td>Random Forests</td>
<td>y</td>
<td>n</td>
<td>y</td>
<td>y</td>
</tr>
<tr>
<td>K-Nearest Neighbors</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>y</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>y</td>
<td>N</td>
<td>y</td>
<td>y</td>
</tr>
<tr>
<td>Convolutional Neural Network</td>
<td>y</td>
<td>Y</td>
<td>y</td>
<td>n</td>
</tr>
<tr>
<td>K-Means Clustering</td>
<td>y</td>
<td>y</td>
<td>y (deprecated)</td>
<td>y</td>
</tr>
<tr>
<td>XGBoost</td>
<td>y</td>
<td>n</td>
<td>y</td>
<td>y</td>
</tr>
<tr>
<td>Support Vector Regression</td>
<td>y</td>
<td>y</td>
<td>y (deprecated)</td>
<td>y</td>
</tr>
<tr>
<td>Deep Neural Networks</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>n</td>
</tr>
<tr>
<td>LSTM/Deep Convolutional LSTM</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>n</td>
</tr>
<tr>
<td>Spatiotemporal Regression</td>
<td>y</td>
<td>n</td>
<td>y</td>
<td>y (needs Euclidean coordinates)</td>
</tr>
<tr>
<td>Recurrent Neural Networks</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>n</td>
</tr>
<tr>
<td>Weighted long short-term memory neural network extended model (WLSTME)</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>n</td>
</tr>
<tr>
<td>Ordinary Least Square Regression (OLS)</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>y</td>
</tr>
<tr>
<td>Lasso Regression</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>y</td>
</tr>
</tbody>
</table>
## Model and Platform Preparation

<table>
<thead>
<tr>
<th>Model Type</th>
<th>Calibration</th>
<th>Downscaling</th>
<th>Upscaling</th>
<th>Fusion</th>
<th>Forecasting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear Regression</td>
<td>is compatible?</td>
<td>SC, PT</td>
<td>SC, PT</td>
<td>SC, PT</td>
<td>SC, PT</td>
</tr>
<tr>
<td>Decision Tree Classifier</td>
<td>SC, PT, TF</td>
<td>SC, PT, TF</td>
<td>SC, PT, TF</td>
<td>SC, PT, TF</td>
<td>SC, PT, TF</td>
</tr>
<tr>
<td>Decision Tree Regressor</td>
<td>SC, PT, TF</td>
<td>SC, PT, TF</td>
<td>SC, PT, TF</td>
<td>SC, PT, TF</td>
<td>SC, PT, TF</td>
</tr>
<tr>
<td>Support Vector Machines</td>
<td>SC, PT</td>
<td>SC, PT</td>
<td>SC, PT</td>
<td>SC, PT</td>
<td></td>
</tr>
<tr>
<td>Random Forests</td>
<td>SC, TF</td>
<td>SC, TF</td>
<td>SC, TF</td>
<td>SC, TF</td>
<td></td>
</tr>
<tr>
<td>K-Nearest Neighbors</td>
<td>SC, PT, TF</td>
<td>SC, PT, TF</td>
<td>SC, PT, TF</td>
<td>SC, PT, TF</td>
<td>SC, PT, TF</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>SC, TF</td>
<td>SC, TF</td>
<td>SC, TF</td>
<td>SC, TF</td>
<td></td>
</tr>
<tr>
<td>Convolutional Neural Network</td>
<td>PT, TF</td>
<td>PT, TF</td>
<td>SC, TF</td>
<td>SC, TF</td>
<td></td>
</tr>
<tr>
<td>K-Means Clustering</td>
<td>SC, PT, TF</td>
<td>SC, PT, TF</td>
<td>SC, PT, TF</td>
<td>SC, PT, TF</td>
<td>SC, PT, TF</td>
</tr>
<tr>
<td>XGBoost</td>
<td>SC, PT, TF</td>
<td>SC, PT, TF</td>
<td>SC, PT, TF</td>
<td>SC, PT, TF</td>
<td>SC, PT, TF</td>
</tr>
<tr>
<td>Support Vector Regression</td>
<td>SC, PT</td>
<td>SC, PT</td>
<td>SC, PT</td>
<td>SC, PT</td>
<td></td>
</tr>
<tr>
<td>Deep Neural Networks (DNN)</td>
<td>PT</td>
<td>PT, TF</td>
<td>SC, PT</td>
<td>SC, PT</td>
<td></td>
</tr>
<tr>
<td>LSTM/Deep Convolutional LSTM</td>
<td>PT</td>
<td>PT, TF</td>
<td>SC, PT</td>
<td>SC, PT</td>
<td></td>
</tr>
<tr>
<td>Spatiotemporal Regression</td>
<td>PT, TF</td>
<td>SC, PT</td>
<td>SC, PT</td>
<td>SC, PT</td>
<td></td>
</tr>
<tr>
<td>Recurrent Neural Networks</td>
<td>PT</td>
<td>PT, TF</td>
<td>SC, PT</td>
<td>SC, PT</td>
<td></td>
</tr>
<tr>
<td>Weighted long short-term memory neural network extended model (WLSTME)</td>
<td>PT</td>
<td>PT, TF</td>
<td>SC, PT</td>
<td>SC, PT</td>
<td></td>
</tr>
<tr>
<td>Ordinary Least Square Regression (OLS)</td>
<td>SC, PT</td>
<td>SC, PT, TF</td>
<td>SC, PT, TF</td>
<td>SC, PT, TF</td>
<td>SC, PT, TF</td>
</tr>
<tr>
<td>Lasso Regression</td>
<td>SC, PT</td>
<td>SC, PT, TF</td>
<td>SC, PT, TF</td>
<td>SC, PT, TF</td>
<td>SC, PT, TF</td>
</tr>
</tbody>
</table>

Python Packages: SC – Scikit-learn; PT – PyTorch; TF - TensorFlow
Data Collection:
- Using PurpleAir API to collect data from their website.
- Using multi-thread and some other strategy to speedup data download performance.
- After downloading data to local memory, script combines data as a table and store the entire table into database.

<table>
<thead>
<tr>
<th>Sensors</th>
<th>pollutants</th>
<th>Time start</th>
<th>Time end</th>
<th>Data volume</th>
<th>Quality controlled</th>
<th>Geographic coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>PurpleAir</td>
<td>PM$<em>{1.0}$, PM$</em>{2.5}$, PM$_{10.0}$</td>
<td>Depends on sensor setup time</td>
<td>Present</td>
<td>2 TB</td>
<td>1. Remove above 500</td>
<td>Worldwide</td>
</tr>
<tr>
<td>EPA</td>
<td>NO$<em>{2}$, SO$</em>{2}$, O$<em>{3}$, PM$</em>{2.5}$, PM$_{10}$, CO</td>
<td>1980-01-01</td>
<td>Present</td>
<td>185 GB</td>
<td>None</td>
<td>Nationwide</td>
</tr>
<tr>
<td>AirNow</td>
<td>PM$_{2.5}$</td>
<td>2018-01-01</td>
<td>2023-06-12</td>
<td>2GB</td>
<td>None</td>
<td>Nationwide</td>
</tr>
<tr>
<td>Aeoronet</td>
<td>AOD</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Accomplishments: Students Engagements

Jared Trigg, NVCC
Theodore Trefonides, Dartmouth
Amiyah Stukes, William & Mary
Hannah Zook, NVCC
Pratyush Muthukumar, CSU-LA
Janmesh Kalra, CSU-LA

Not Pictured...
Phoebe Pan, ASSIP
Climate Change and Drought

Climate Moisture Index (CMI)

Under different climate scenarios (RCP 2.6 and 8.5) and timeframes

REFERENCE PERIOD
1981–2010

CMI (cm)

Very Wet
Wet
Moist
Dry
Very Dry
Extremely Dry

Provincial Boundaries

RCP 2.6
RAPID EMISSIONS REDUCTIONS

2071–2100

CONTINUED EMISSIONS INCREASES

2011–2040

2041–2070

2071–2100

RCP 8.5

Head Fire Intensity
Intensité du front
2023-06-05

0 - 10 kW/m
10 - 500
500 - 2000
2000 - 4000
4000 - 10000
10000 - 30000
> 30000
Nil / s.o.

Map created at 19:34 (UTC) on 2023-06-12
Carte créée le 2023-06-12 (UTC) à 19:34

climate_moisture_index_update RCP85_1140.gif (1140 × 881) (canada.ca)
Canada wildfire and air quality?

Photos: Canadian wildfires bring smoke, low air quality to eastern US
Smoke from the fires continues to blanket the East Coast.

Photos: Canadian wildfires bring smoke to eastern US (bostonherald.com)
GOES-16 ABI AOD for June 7, 2023
Accomplishments: Sensor Data Exploration
1-2 slides

Tracking Canadian wildfire smoke with PurpleAir Sensors

- Efficient retrieval and database storage of PurpleAir data for analysis
- **Sensor table** holds basic information about the sensor
- **Reading table** holds minute-level record for every sensor
A Spatiotemporal Statistical Analyses

Aggregation

Dimension Reduction
1. **Formalized the list of models** that capture the complex relationships between pollutant concentrations and various factors like meteorological conditions, emissions, and geographical features.

2. Formalized a list of methods for **uncertainty quantification** and accuracy assessments in air pollutant retrieval, and prediction. This involved identifying and implementing techniques such as sensitivity analysis to effectively quantify and manage uncertainties.

3. Identified **key model parameters** and their ranges to optimize model performance and improve accuracy.

4. **AQ pattern/event analysis**: Analyze the changes of spatiotemporal patterns of air pollution due to major AQ event over study regions such as the **Canadian wildfire to put the research in context.**
Future Plans

- Continue gather data source, access and preprocess for training data
- Continue model training on the 4 platforms and for various data analytics
- Develop data-fusion techniques to integrate data from multiple sources into a unified dataset. This may involve spatial and temporal alignment, harmonization of variable units, and handling missing or inconsistent data.
- Fine-tune the models by adjusting hyperparameters, such as learning rates, regularization techniques, or network architectures, to improve their performance and generalization capabilities.
- Utilize cross-validation techniques to assess model stability and avoid overfitting.
- Model evaluation - Evaluate the trained models using separate validation datasets to assess their predictive accuracy and generalizability.
- Prepare open-source software and release
Presentation Contents

- Background and Objectives
- Technical and Science Advancements
- Summary of Accomplishments and Plans Forward
- Actual or Potential Infusions and Collaborations
- Publications - List of Acronyms
Actual or Potential Infusions and Collaborations

• **Adopted big data technologies and a cloud computing platform** to process and handle the vast amounts of air quality data efficiently. The ML models are trained on large datasets with ease using **distributed computing frameworks**.

• Provided an opportunity to **transfer knowledge and expertise** from researchers to the students, enabling them to gain practical experience, learn new skills, and contribute to the project.

• Ensuring **replication and reproducibility**, the formal study on AQ analysis will be integrated with the Harvard KNIME spatiotemporal data analytics workflow system to be used by 13 spatial data labs and

• The system developed will be made accessible as **open-source software**, facilitating the widespread adoption and utilization of the research findings.
FireAlarm System Integration

Mission Statement

AQACF aims to provide the scientific foundation for the development of air quality data applications and decision support tools for scientists, policymakers, resource managers, and the public. The program aims to improve the understanding of atmospheric composition, air quality, and their interactions with climate and ecosystems, and to develop innovative solutions to air quality challenges. AQACF is part of NASA's Earth Science Division, which aims to better understand Earth's systems and changes to inform decisions that affect the future of the planet.

- [https://ideas-digitaltwin.jpl.nasa.gov/aqacf/](https://ideas-digitaltwin.jpl.nasa.gov/aqacf/)
- Adding API to serve the data and analytics of in-situ data
The City of L.A. is Applying Data and Machine Learning to Understand Urban Air Quality

The City of Los Angeles is in a unique situation to be an urban proving ground to look at how to better understand, predict, and mitigate the issues of air pollution for 4 million citizens. In partnership with NASA, California State University Los Angeles (CSULA) and OpenAQ, the City has embarked on an Air Quality Project with a goal of helping to mitigate the effects of air pollution through interventions that have measured results.

The Predicting What We Breathe project looks at the time-series measurements of satellite and ground data and applies machine learning to uncover patterns that may not be discernible to human analysts. Enhancing human understanding and prediction of air quality will help inform local governments and others on appropriate measurements, analytics, predictive algorithms and mitigation strategies that are useful for dealing with air quality variability.

The work done here is funded by a NASA grant for Advanced Information Systems and Technology.
New NASA Instrument to Study Air Pollution

Provide Calibrated In-Situ Datasets (with Thomas and Sina)

MAIA | Home Page (nasa.gov)
Harvard Spatial Data Lab

Provide the air quality data analytics workflow for 13 affiliated international labs of spatiotemporal innovation with KNIME integration and fusion.
Presentation Contents

• Background and Objectives

• Technical and Science Advancements

• Summary of Accomplishments and Plans Forward

• Actual or Potential Infusions and Collaborations

• Publications - List of Acronyms


List of Acronyms

- ML - Machine Learning
- AQACF - Air Quality Analytic Collaborative Framework
- GHG - Green House Gases
- AIST - Advanced Information Systems Technology
- AQ - Air Quality
- EPA - Environmental Protection Agency
- OLS - Ordinary Least Square Regression
- SVR - Support Vector Regression
References


Acknowledgements

• Dr. Jacqueline Le Moigne/AIST provided funding support, Ben Smith helped with connections.
• Dr. Daniel Q. Duffy and NSF Spatiotemporal I/UCRC for the computing infrastructure.
• ESIP Air Quality Cluster.
• Jeanne Holm, LA, collaborated on use cases.
• Karen Moe collaborated on Chevy Township use case.
• NSF START intern program provided funding for community college students.
• Many Spatiotemporal I/UCRC members commented on the study.
Stochastic Parameterization Of An Atmospheric Model Assisted By Quantum Annealing

Alexandre Guillaume (PI)
Jet Propulsion Laboratory, California Institute of Technology

AIST-18/21-005 Annual Technical Review
07/21/2023, 1:30pm PDT

Team listing: Youngmin Seo (co-I, JPL), Marcin Kurowski (co-I, JPL), Boualem Khouider (collaborator, UV)

Objective
Our goal is to create a quantum computing framework to enable a new stochastic parameterization of the boundary layer clouds constrained by remote sensing data. This development could improve the parameterization of shallow clouds in climate models.

Open Earth Science to Ising spin systems advances in theory, algorithms and hardware (other than quantum annealing) for future work. The work of Giorgio Parisi, laureate of the Nobel prize in Physics 2021, exemplifies the benefits of exploiting spin systems theories for climate studies.

Key Milestones
- Quantum BM retrieval 07/23
- Conventional (MCMC) forward model 07/23
- Reverse QA forward model 01/24
- Model validation 01/24

Approach
We will implement a quantum and conventional algorithms:
- AG: Implement a Boltzmann Machine (BM) algorithm to retrieve stochastic parameters from remote sensing observations, using a quantum annealer.
- YS: Implement a conventional Markov Chain Monte Carlo (MCMC) algorithm to simulate the stochastic model forward, i.e. generate clouds coverage.
- AG: Implement a quantum annealing (QA) optimization with memory, using reverse quantum annealing, to simulate the stochastic model forward, i.e. generate clouds coverage.

Co-I/collaborator: Youngmin Seo (JPL), Marcin Kurowski (JPL), Boualem Khouider (UVIC)
Background / Objectives

- General circulation models, GCM, are the most advanced tools to simulate the response of the global climate to increasing greenhouse concentration. There are many processes in the climate systems that occur at scales smaller than GCM (grid) resolved scales. Those unresolved processes are represented in numerical models with semi-empirical formulae known as parameterization. Improving climate model predictions will thus require new improved parameterizations.

- A quantum annealer (QA) is a physical realization of the Ising model, a regular lattice of interacting magnetic moments pointing up or down.

- Khouider and collaborators have developed different stochastic models through the years relating the Ising model to atmospheric models [KB2019]. Simulating a mesoscale domain of a 100 km at a 1 km resolution for a two-dimensional lattice requires the exploration of a trade space of dimension $2^{10^4} \approx 10^{3010}$. Typically, Markov chain Monte Carlo (MCMC) methods are used to simulate such large spaces.

- We propose to use a quantum annealer to simulate stochastic models of the atmosphere based on the Ising model [KB2019] to take advantage of the speedup enabled by quantum mechanics and algorithmic advances in the field of Ising model theory.
Background/Objectives (2)

- Our goal G1 is to create a quantum computing framework to characterize a stochastic parameterization of the boundary layer clouds constrained by remote sensing data. Using the formal similarity of both, the stochastic parameterization and the quantum hardware, to a regular lattice known as the Ising model, we can take advantage of the quantum computing efficiency.

- Our second goal G2 is to further the use of the Ising model in climate science and promote the use of theories, algorithms and new hardware designed to solve this model when it can benefit climate science.

- The proposed work will demonstrate that QC can be used to improve observation-based parameterizations used in climate models.

The computing framework we are proposing to implement couples large scale dynamics with sub-grids parameterization while using remote sensing observations to do so. These elements are similar to what the decadal survey [2018] lists as essential ingredients for progress: the representation of physical processes in parameterizations, coupling of Earth-system components and the use of observations with advanced data assimilation algorithms. Our proposed work follows a similar methodology as listed in the decadal survey.

Presentation Contents

- Background and Objectives
- Technical and Science Advancements
  - Summary of Accomplishments and Future Plans
  - Actual or Potential Infusions and Collaborations
  - Publications - List of Acronyms
1Technical and Science Advancements

**Ising model**

\[ H_P = \sum_i h_i \sigma_i^z + \sum_{i,j} J_{i,j} \sigma_i^z \sigma_j^z \]

\[ E(\mathbf{x}) = -\sum_i h_i x_i - \sum_{i,j} J_{i,j} x_i x_j \]

\[ p(\mathbf{x}) = \frac{1}{Z} e^{-\frac{E(\mathbf{x})}{T}} \]

\[ Z = \sum_x e^{-\frac{E(x)}{T}} \]

D-Wave quantum annealer:

\[ QA(h, J) \rightarrow \{\mathbf{x}\} \]

---

2Technical and Science Advancements

**Restricted Boltzmann Machine (RBM)**

\[ E(\mathbf{v}, \mathbf{h}) = -\sum_{i=1}^m \sum_{j=1}^n w_{ij} v_i h_j - \sum_{j=1}^n b_j h_j - \sum_{i=1}^m c_i v_i \]

\[ (\mathbf{v}, \mathbf{h}) \in \{0, 1\}^{m+n} \]

\[ p(\mathbf{v}, \mathbf{h}) = \frac{1}{Z} e^{-E(\mathbf{v}, \mathbf{h})} \]

\[ \ln L(\theta|\mathbf{v}) = \ln p(\mathbf{v}|\theta) = \ln \frac{1}{Z} \sum_{h} e^{-E(\mathbf{v}, \mathbf{h})} = \ln \sum_{h} e^{-E(\mathbf{v}, \mathbf{h})} - \ln \sum_{h} e^{-E(\mathbf{v}, \mathbf{h})} \]

\[ \sum_{v \in S} \frac{\partial \ln L(\theta|\mathbf{v})}{\partial w_{ij}} = \alpha \langle (v_i h_j)_{\text{data}} - (v_i h_j)_{\text{model}} \rangle \]

Sample with D-Wave QA
Technical and Science Advancements

MODIS cloud classified and water

JPL restricted Boltzmann machine (RBM)

30 x 30 pixels
Technical and Science Advancements

JPL Quantum Restricted Boltzmann Machine (QRBM)

12 x 13 = 156 pixels

This Boltzmann machine has:
• 156 visible units
• 156 hidden units

Advantage_system6.2:
• 5614 qubits
• 40106 couplers

156x156 = 24336 connections,
(24336 couplers)

4627 used ➔ 4627/312 ≈ 14.8 physical qubits/logical qubit ~ 15

Our QRBM is the largest (to our knowledge)
Technical and Science Advancements

**Data pipeline**

Goal: collect remote sensing data \( \sigma_{\text{obs}} \) to form a training to learn the stochastic model parameters.

We have implemented a pipeline that:

- Download MODIS data and generate cloud masks and classes.
- Download water products from ERA5 and GOES.
- Delete original data (for size reduction)

This pipeline is coded in Matlab and Python.

Technical and Science Advancements

**MODIS cloud classed and water**

Black: no data
Blue: partial data
Green: no low clouds
Orange: spatial restriction to avoid lands
Red: selected for learning set
Technical and Science Advancements

It takes approximately 12 days to download 1 year of data ➔ automation necessary

Presentation Contents

- Background and Objectives
- Technical and Science Advancements
- Summary of Accomplishments and Future Plans
- Actual or Potential Infusions and Collaborations
- Publications - List of Acronyms
Summary of Accomplishments and Future Plans

• In the last quarter of year 1 (10 months in reality):
  1. We have implemented a quantum restricted Boltzmann machine, QRBM, using remote sensing data to invert the stochastic model. This finished the first milestone called Boltzmann Machine, BM, which in effect contained two subtasks, a classical and quantum RBM.
  2. We have implemented a software “pipeline” to collect remote sensing data from MODIS, ERA5 and GOES. This data is used to form training data to invert the stochastic model and get its parameters \( \{ h_i(q), J_{ij}(q) \} \). This has never been done before.

• In the last six month of year 2 (8 months in reality):
  • We will implement the quantum version of the forward stochastic model using quantum annealing and reverse quantum annealing. The forward model generates a cloud cover from a set of parameters \( \{ h_i, J_{ij} \} \). This will be the first time, that this model will take parameters extracted from observations i.e., \( \{ h_i(q_{obs}), J_{ij}(q_{obs}) \} \).
  • We will integrate all components and validate the model framework:
    • Fine-tune both the pipeline: MODIS cloud classes versus GOES cloud top height to discriminate low clouds. Fine-tune the data: filter images by similarity to form training set.
    • Validate model on a separate portion (in space and/or time; the validation region or regions) of the observational data.
  • We will write an article.

Presentation Contents

• Background and Objectives
• Technical and Science Advancements
• Summary of Accomplishments and Plans Forward
• Actual or Potential Infusions and Collaborations
• Publications - List of Acronyms
Actual or Potential Infusions and Collaborations

- This project is an Early Stage Technology (EST), with a TRL_{in}=2 and TRL_{out}=3. The TRL at the end of the project will be too low for an infusion. We have however a plan for a follow on that would bring the concept and computational framework to a more mature TRL. The goal is to infuse the stochastic parameterization framework into a global circulation model (GCM) and demonstrate its efficiency. We laid out our vision in a 2-pages white paper.

Presentation Contents

- Background and Objectives
- Technical and Science Advancements
- Summary of Accomplishments and Plans Forward
- Actual or Potential Infusions and Collaborations
- Publications - List of Acronyms
### Acronyms

#### List of Acronyms

- **QA**: Quantum Annealing (or Quantum Annealer)
- **QC**: Quantum Computing
- **GCM**: General circulation model
- **BM**: Boltzmann Machine
- **RBM**: Restricted Boltzmann Machine
- **QRBM**: Quantum Restricted Boltzmann Machine
- **MODIS**: Moderate Resolution Imaging Spectroradiometer
- **GOES**: Geostationary Operational Environmental Satellites
- **MCMC**: Markov Chain Monte Carlo
SLICE: Semi-supervised Learning from Images of a Changing Earth

Brian Wilson (PI)
Edwin Goh, Alice Yepremyan, Kai Pak, Diego Martinez, Brian Bue
Jinbo Wang, Ben Holt
Grace Vincent, Isaac Ward (interns)
Andy Thompson, Ed Armstrong (collaborators)

AIST-21-0025 Annual Technical Review
July 21, 2023

Objective

- Develop a scalable self- and semi-supervised learning (SSL) framework for Earth Science datasets with limited labels
- Perform self-supervised pretraining on unlabeled global data products to generate reusable, task agnostic learned representations
- Finetune pretrained weights on classification and segmentation tasks from high-resolution simulation (e.g., pre-SWOT), SAR and oceanography data
- Demonstrate ability of SSL pre-trained models to perform few-shot classification, segmentation and phenomena detection for ocean surface phenomena using SST, SSH, SSC, and SAR datasets.

Approach

- Identify benchmark tasks/datasets for classification, segmentation and phenomena detection
  - Global ECCO simulation LLC4320 at 1/48-degree resolution and hourly output (SRS) on AMES Pleiades.
  - MODIS L2 SST/SSC images
  - Synthetic SWOT SSH L2 based on LLC4320
  - Sentinel-1A/B SAR
- Adapt SLICE framework to interface with AIST-funded ACF’s
- Develop efficient data preprocessing/augmentation pipelines for global, multi-domain data products
- Leverage multi-cloud and NASA HPC for large-scale SSL pretraining and finetuning on benchmark tasks

Key Milestones

- Compare semi vs. plain supervised learning on segmentation & detection problems.
  (04/23)
- Release v1 of SLICE platform
  (06/24)
- Submit science papers on eddy properties from SSH, and eddy detection from SAR
  (08/24)
- Implement data augmentation & training pipeline in ACF environment at PO.DAAC
  (11/24)
- Begin beta testing of SLICE for SWOT and other analysis at PO.DAAC
  (01/25)
- Submit paper on classification & segmentation performance analysis from SSL models
  (04/25)
- Small-scale eddy database delivered to SWOT
  (06/25)

TRL$_{init}$ = 3  TRL$_{current}$ = 4  TRL$_{final}$ = 6 (after 3 years)
The Challenge of Large-Scale Scientific Imagery

- Decade-scale remote sensing datasets for Earth, ocean, atmospheric, & planetary sciences
  - Primarily unlabeled
  - Impossible to properly label at scale
  - Crowd-sourcing is difficult, and still falls short
  - Varying data platforms and acquisition rates

Top-Level ML Technology Goals

- **Investigate** cutting-edge Self-Supervised Learning (SSL) algorithms and compare their performance with supervised learning baselines for Earth imagery problems under limited labels.
- **Develop** a scalable, flexible cloud-based framework for performing computer vision tasks on decadal scales, with preconfigured workflows and pluggable algorithms.
- **Train and publish** optimized SSL and Foundation models for Earth observations — inspired by GPT, CLIP and SAM — for tasks in Earth image classification, segmentation, and phenomena detection.
- Apply SLICE to study ocean surface phenomena - detecting ocean eddies at meso and sub-meso scales & deriving eddy properties and associated heat fluxes.
- Generalize and apply the system to other decade-scale imagery datasets — land, ice, atmosphere, Mars, other planets, etc.
SLICE Objectives:
“Semi-supervised Learning from Images of a Changing Earth”

• Establish the SLICE framework and platform for applying scalable semi-supervised computer vision models to Earth and Planetary imagery
  • Running in an AWS Cloud and Supercomputing environment
  • Easily adopted as a reusable platform by NASA data centers, mission science teams and NASA PIs.

• Investigate and characterize the accuracy of multiple SSL Foundation models (SimCLRv2, MoCo, MAE, ViT) on a variety of relevant remote sensing tasks with minimal labels.

• Build and publish self- and semi-supervised learning models with a focus on the upper ocean small-scale processes in anticipation of several on-going and potentially future NASA missions (i.e. SWOT, ODYSEA, and PACE).
  • Build a database of small-scale eddies for SWOT measurement validation
  • Fine-tune pretrained SSL models to derive heat flux from sea surface observations
    • SSH, SST, SSC (sea surface current), wind
  • Fine-tune pretrained SSL semantic segmentation models to identify eddy morphologies for potential deployment at scale on full Sentinel-1 scenes

Detecting Sub-Mesoscale Eddies with Computer Vision

• SAR detects submesoscale eddies primarily via alignment of marine slicks by circulation flow field /fine structure, SST via surface temperature gradients
• Differences in detected SAR eddies (6) and SST (2) likely related to time of generation and energetics
• Little known about properties at depth, generation, lifespan, vertical heat exchange
Fine-Tuning Computer Vision Foundation Models

- Detecting submesoscale ocean eddies from Sentinel-1 SAR data (Grace)
  - Pre-trained & Tuned two contrastive models:
    - Simple framework for contrastive learning of visual representations (SimCLR)
    - Momentum Contrast (MoCo)
  - Evaluated precision & recall

- Reconstructing Sea surface temperature (SST) to fill data gaps (Edwin)
  - Trained a Masked Auto Encoder (MAE) model
  - Masking out a percentage of image patches
  - Evaluated accuracy and coherence

Summary of Year 1 Accomplishments

Submesoscale SAR eddy identification from SAR images
- Assembled:
  - unlabeled dataset with 100k SAR images
  - labeled dataset from Mediterranean and Pacific Ocean (~30k images)
- Conducted contrastive pretraining using Momentum Contrast (MoCo) and SimCLR
- Finetuned and evaluated pretrained models to classify SAR images into eddies vs. not eddies

Sea surface temperature (SST) reconstruction under clouds
- Assembled dataset of 1 million SST tiles from ECCO LLC4320 simulation
- Trained a masked autoencoder (MAE) on SST tiles to reconstruct masked regions
- Submitted Ocean Science journal manuscript (https://doi.org/10.5194/egusphere-2023-1385) with the following contributions:
  - Mixed masking ratio technique
  - Additional evaluation metrics specific to SST (spatial correlation, spectral coherence)
- Trained MAE on SSH tiles from SWOT simulation to reconstruct SSH under SWOT nadir mask

Ocean vertical heat flux retrieval with deep learning
- Assembled dataset of 3.4 million SSH tiles from ECCO LLC4320 simulation to mimic SWOT SSH retrieval
- Developed pipeline to train a model to predict average ocean vertical heat flux for each tile at 40m
Contrastive Pretraining for SAR Ocean Eddy Detection

Unsupervised pretraining on unlabeled SAR tiles
- Maximize agreement of learned representations for the same image under different data augmentations
- Use contrastive loss
  - MoCo and SimCLR
- Goal: learn task-agnostic features/representations

Sentinel-1 SAR Data Pipeline

Contrastive learning outperforms supervised learning

- The contrastive learning techniques resulted in superior performance for both test sets
  - SAR-based contrastive pretraining techniques are comparable to and in some instances outperform their ImageNet counterparts
- Contrastive learning techniques minimize the performance degradation across domain-shifts
  - Location, resolution, and spatial coverage of tile

Contrastive learning improves sample efficiency

- SAR based contrastive pretraining improves Mediterranean and California eddy classification performance in a limited-label system
- With 20% of available finetuning labels, both SAR based contrastive pretraining retain greater than 91% of peak Mediterranean performance
Apply trained model on unseen images to generate submesoscale eddy database

- Ran inference on 5,000 randomly-sampled images from Pacific Ocean and sent predictions to expert (Holt) for validation
- Current model captures **55% of all eddies** in the sample dataset
  - 97% of all non eddy tiles
- Can use this information to improve finetuning results
  - Increase number of positive samples
  - Identify difficult instances

---

**Eddy definitions are subjective**

False Negative Example
SAR for Eddy Detection Next Steps

- Quality check the current finetuning dataset
- Experiment with different SSL pretraining methods
  - Analyze the impact of different augmentations (e.g., leverage SAR speckle as a pretraining task)
- Investigate the dataset variations
  - Change in spatial domains and resolutions
- Evaluate current datasets for distribution, seasonality, and longevity

Size distribution of eddies detected in the Southern California Bight from 1992 to 1998

Density of submesoscale eddies during (a) winter (b) spring (c) summer (d) autumn.

Total number of eddies observed: 107

MAESSTRO: Masked Autoencoders for Sea Surface Temperature Reconstruction under Occlusion
SST Reconstruction - Motivation

- Monitor and forecast climatic variations, e.g., El Niño and La Niña\(^1\)
- Used for atmospheric and numerical weather prediction models (e.g., validation, as boundary conditions, etc.)\(^1\)
- Enable tracking of marine flora/fauna:
  - coral bleaching
  - sea turtles (migration and nesting)

Multi-Scale Ultra High-Resolution Sea Surface Temperature (MUR SST)

---

SST Reconstruction - Motivation

- MUR SST: global, 1-km, daily gridded SST product (L4)
- Attended the Grid SST hackathon (NASA PO) in Nov 2022 to kickstart a successor to MUR SST
- Identified masked autoencoders (MAE) as a viable self-supervised learning approach to reconstruction

Masked Autoencoders (MAE) for SST Reconstruction

- Concept: train a model to reconstruct missing pixels
  - Self-supervised because the image itself is the ground truth “label”
- Original developed by Facebook Research for self-supervised pretraining
  - i.e., for use with supervised fine-tuning on downstream tasks (e.g., classification, segmentation)
- Uses vision transformer (ViT) backbones
Training MAE on Simulated Sea Surface Temperature Data

- Start by training on clean, cloud-free, high-resolution SST
  - Use LLC4320 high-res ocean simulation (1/48° grid spacing)
- Split ocean into ~500km x 500km tiles (256x256 pixels)
  - Remove land and under-sea ice
- Result: 1,000,809 tiles
  - Training set (2012): 750,362
  - Validation set (2011): 250,447
- Train for 300 iterations/epochs on this dataset
- Loss function: mean squared error across all masked pixels

Evaluation Metrics

- Root mean squared error (RMSE)
  - Express error in terms of °C
- Spatial correlation - Pearson's correlation coefficient
- Spectral coherence
  - Fourier transform of convolution between ground truth and predicted SST
  - \( C(k) = \frac{|\text{CSD}(\hat{T}, \hat{T})|^2}{\sigma^2(\hat{T})\sigma^2(\hat{T})} \)
  - \( k \): wavenumber
  - \( \hat{T} \): temperature (predicted)
MAESSTRO achieves ≤ 0.2°C error (globally) with up to 80% of original SST removed

- Test MAESSTRO on example tile from a different numerical simulation (LLC2160)
  - 80% of pixels randomly masked
- MAESSTRO reconstruction captures SST front and two small filaments near the bottom
- Reconstructed gradient is also sharper than cubic interpolation
  - Highlights small-scale frontal structures
MAESSTRO outperforms cubic interpolation and Kriging across all LLC2160 tiles.

Lowest mean RMSE

Highest correlation
MAESSTRO on example VIIRS SST tile

- MAESSTRO demonstrates a global consistent performance on simulated data at two resolutions (≈2 km and ≈4 km grids) and on the satellite SST data at a pixel resolution of 750 m

Summary of Year 1 Accomplishments

Submesoscale SAR eddy identification from SAR images
- Assembled:
  - unlabeled dataset with 100k SAR images
  - labeled dataset from Mediterranean and Pacific Ocean (~30k images)
- Conducted contrastive pretraining using Momentum Contrast (MoCo) and SimCLR
- Finetuned and evaluated pretrained models to classify SAR images into eddies vs. not eddies

Sea surface temperature (SST) reconstruction under clouds
- Assembled dataset of 1 million SST tiles from ECCO LLC4320 simulation
- Trained a masked autoencoder (MAE) on SST tiles to reconstruct masked regions
- Submitted Ocean Science journal manuscript (https://doi.org/10.5194/egusphere-2023-1385) with the following contributions:
  - Mixed masking ratio technique
  - Additional evaluation metrics specific to SST (spatial correlation, spectral coherence)
- Trained MAE on SSH tiles from SWOT simulation to reconstruct SSH under SWOT nadir mask

Ocean vertical heat flux retrieval with deep learning
- Assembled dataset of 3.4 million SSH tiles from ECCO LLC4320 simulation to mimic SWOT SSH retrieval
- Developed pipeline to train a model to predict average ocean vertical heat flux for each tile at 40m
Summary of Accomplishments and Future Plans

Submesoscale SAR eddy identification from SAR images
- Pretrained models using Momentum Contrast (MoCo) and SimCLR on an unlabeled dataset of 100k SAR images, and then finetuned and evaluated these models for the classification of SAR images into eddies and non-eddies
- Future plans: Data product which consists of classification of the SAR eddies.

Sea surface temperature (SST)/ Sea surface Height (SSH) reconstruction under clouds
- Assembled dataset of 1 million SST tiles from ECCO LLC4320 simulation, trained a masked autoencoder (MAE) on SST tiles to reconstruct masked regions and trained the MAE on SSH tiles from SWOT simulation to reconstruct SSH under SWOT nadir mask.
- Future plans: we plan to extend our methodology through further training and evaluation using high-resolution SST datasets from VIIRS and MODIS sensors as well as other supporting satellite measurements including lower-resolution microwave SST and sea surface height from conventional altimeters and the just-launched Surface Water and Ocean Topography (SWOT) mission

Ocean vertical heat flux retrieval with deep learning
- Assembled a dataset of 3.4 million SSH tiles from ECCO LLC4320 simulation, designed a pipeline to train a model for predicting average ocean vertical heat flux at 40m for each tile.
- Future plans: continue testing heat flux prediction pipeline in preparation for SWOT data availability in August.

Actual or Potential Infusions and Collaborations

- SWOT (Transition)
  - SWOT swaths contain 20km gap nadir and the the long temporal gaps which can be filled using MAESTRO for reconstruction. Filling this gap could potentially reveal ocean eddies that would have otherwise been missing.
  - Use SWOT data for heat flux and vertical velocity retrieval
- PODAAC (Infusion)
  - MAESTRO can be used to address the challenge of filling gaps in high-resolution (1km) sea surface temperature (SST) fields caused by cloud cover, which often results in gaps in the SST data and/or blurry imagery in blended SST products.
  - Engage PO.DAAC to apply MAESTRO for level-4 high-resolution SST product as a pilot experiment toward the next generation high-res SST gridded product in the cloud.
  - Submesoscale eddy database based on SAR images.
- Open-source (Technology transfer)
  - Publish datasets, models, and weights for use by SWOT team and research communities at large
Publications

- GHRSSST abstract submitted
- Two IGARSS papers / presentations
  - Reconstruction of Sea Surface Temperature Under Clouds Using Masked Autoencoders
  - Unsupervised SAR Images for Submesoscale Oceanic Eddy Detection
- Manuscript submitted to European Geosciences Union (EGU) Ocean Science Journal
  - MAESTRO: Masked Autoencoders for Sea Surface Temperature Reconstruction under Occlusion
- Other papers in contrastive learning:

List of Acronyms

- CV  Computer Vision
- DL  Deep Learning
- SSL  Self-Supervised Learning
- SimCLR  Simple framework for Contrastive Learning of visual Representations
- DINO  self-distillation with no labels
- EsViT  Efficient self-supervised Vision Transformer
- MAE  Masked Auto Encoder
- MVIT  Multi-scale Vision Transformer
- ViT  Vision Transformer
- SAM  Segment Anything Model
- SAR  Synthetic Aperture Radar
- SST  Sea Surface Temperature
- SSH  Sea Surface Height
- SSV  Sea Surface Velocity
Thematic Observation Search, Segmentation, Collation and Analysis (TOS²CA) system

Ziad S. Haddad (PI, Jet Propulsion Laboratory, California Institute of Technology)
Brian Knosp (Co-I, JPL/CalTech)
Svetla Hristova-Veleva (Co-I, JPL/CalTech)
Quoc Vu (Co-I, JPL/CalTech)
Karen Yuen (Co-I, JPL/CalTech)
George Chang (Co-I, JPL/CalTech)
Charles Thompson (Co-I, JPL/CalTech)

AIST-18/21-0015 Annual Technical Review
2023-07-21

Objective
Develop a user-driven data-centric system that can identify, collate, serve and analyze diverse Earth system data relevant to a given phenomenon of interest to the Earth System Observatory.

TOS²CA will not only facilitate the curation and analysis of data from disparate sources simultaneously, it will also enable scientists to:
- Establish science-traceability requirements
- Quantify detection thresholds
- Define uncertainty requirements
- Discover conditional relations between process variables

This prototype is intended to become autodidactic, and expandable to assist decision-makers in applications.

Approach
The components of TOS²CA include:
1) a user-driven phenomenon-definition function ("TOS²CA-1");
2) an automated curation function ("TOS²CA-2"); and
3) a user-friendly visualization and data exploration toolkit.

TOS²CA-1: The mapping (detection & tracking) of the occurrences of a phenomenon using relevant Program-of-Record datasets, is done by adapting ForTracCC (program to find connected components of an inequality and track their overlap over time).

TOS²CA-1: The data curation is developing APIs and adapting software already developed for the NASA-ESA "Multi-mission Algorithm and Analysis Platform" for the eco-systems community.

Co-Is/Partners: Brian Knosp, Svetla Hristova-Veleva, Quoc Vu, George Chang, Karen Yuen, Charles Thompson, JPL; Randy Sawaya, Primer.ai

Student Interns: Jason Eriksen (graduated 6/23); Phillip Le

Chronology of Occurrences • of user-specified phenomenon

Key Milestones
- Design of Chronology of Occurrences generator TOS²CA-1 12/22
- Design of data curator TOS²CA-2 04/23
- Design of analysis and visualization modules TOS²CA-a+v 08/23
- Implementation and testing of TOS²CA-1 06/23
- Implementation and testing of TOS²CA-2 10/23
- Implementation and testing of TOS²CA-a+v 01/24
- Integration and testing 06/24

TRL₀ = 3  TRL_current = 4
Background and Objectives

The main goal is to create a system that enables a scientist to improve our understanding of a phenomenon (hurricanes, earthquakes, volcano eruptions, heat waves, droughts, landslides, etc.) by identifying, empirically, relations between geophysical variables (obtained from different observation systems) that the data can reveal by joint analysis of the observations for many/all occurrences of the phenomenon.

Examples:
- What atmospheric conditions lead to droughts? Specifically in the US West? What, in these conditions, is most related to the duration of the drought, or to its spatial extent?
- What gas and aerosol emissions precede volcanic eruptions? For a subclass of volcanoes?
- What in the atmospheric state is most correlated with the longevity of convective storms? With their size? With their productivity (surface precipitation)? In Africa versus South America?
- (How) are sea surface temperatures (anomalously high or low) related to the strength of the monsoon? In SouthEast Asia vs India vs West Africa vs South America vs North America?
- Over what time scales and/or over what range of intensities does surface precipitation affect landslides?

All these can be answered empirically with a system that can curate data on user-specified variables, organized according to the event occurrences of a phenomenon defined by the user, then enable the user to visualize and analyze the data.
Technical and Science Advancements

- We have designed, implemented and tested the path-1 phenomenon definition tool
  - “path-1” is the approach where the user chooses
    - an observable (such as surface temperature, relative humidity, IR brightness temperature, vertically-integrated ice-water, near-surface wind speed, land-surface displacement – these are examples that we will use in the illustrations to come), and
    - a method to constrain it (such as “less than a user-specified threshold”, or “greater than 2 standard deviations from the monthly mean”), and
    - a regional domain and a time interval,
    and TOS²CA finds the connected components of the solution set at each time step, then tracks these components in time (by detecting intersections at consecutive times) to produce a chronology of occurrences of the phenomenon

The following slides illustrate path-1 in a very interesting case relevant to the Atmosphere Observing System (AOS) and the Investigation of Convective Updrafts (INCUS), where, in order to study convective storms, the latter need to be defined, and detected, and their chronologies need to be listed. However, to date, there is no agreed method to do this:
  - One way relies on the IR brightness temperature (values < 235K indicate a high cloud that is probably either deep or the anvil produced by a deep storm) – but this would not detect storms that do not reach 10,000m AMSL or that have not yet reached that height...
  - Another, more physical way is to rely on the vertically-integrated ice water being greater than a threshold (e.g. 0.1 kg/m² or 0.2 kg/m² – need to test) – but that may depend on the source of the ice water data …
Before illustrating path-1, it is important to highlight the development of three previously unplanned tools to help users “visualize” the outcome of path-1:

1. A listing of the table of contents of the outcome of the path-1 process (i.e. a listing of the storms or droughts or heatwaves, or more generally the occurrences of whatever phenomenon the user specified within the user region and time interval)
2. A display of all the occurrences in the region at a single user-chosen time
3. A display of the chronology of a single occurrence throughout the time interval

These three capabilities allow the user to “see” the outcome of the phenomenon-definition portion of their effort much as we have illustrated in our schematic representation:

This is to allow the user to gauge how constrictive (or generous) their specified threshold was, and to decide if the user might be more interested in a smaller or larger time interval or spatial domain

These 3 tools are not “visualization” capabilities per se, since they only apply to the phenomenon itself and not to any of the variables that the user will want to have curated into the phenomenon occurrences.

The actual visualization capabilities will be described in more detail in slide 26 …

Major development of the “anomaly-detection” Python software library to support path-1 has completed.

This software was designed to modularly:
1) Read data sets with an appropriate reader
2) Submit data to ForTraCC
3) Output a netCDF-4 file of geospatially and temporally located mask indices
4) Plot the originally requested data variable with mask overlays

This library includes readers for multiple data sets, such as GPM and MERRA-2. We currently have operators for:
• Equal To
• Less Than
• Less Than or Equal To
• Greater Than
• Greater Than or Equal To
• Standard Deviation

Though this work has completed, we will be adding additional data readers as requested, and resolving some minor open issues. We will also be updating as needed to support any emerging data curation and visualization requirements.
The TOS²CA Phenomenon Definition page collects all relevant information needed by the anomaly-detection library to create the mask indices.

This includes the temporal and geospatial bounds, as well as the dataset and variable of interest.

Data is read out of NASA’s Earthdata Cloud (i.e., AWS S3), or through other DAAC services, and passed to ForTraCC.

Once ForTraCC generates the masks, the user receives an email with an AWS S3 location where they can pick up their netCDF-4 mask file and any plots.

Additionally, they can visit the Phenomenon Definition Viewer page which contains:
1. Job parameters
2. Table of Contents of identified anomalies
3. Map showing anomaly footprints
When an anomaly is selected from the Table of Contents, users can step through footprints of the anomaly to see its lifecycle, using the controls below the map.

In this example, we move between two timesteps.

Clicking a footprint on the map will bring up a data plot. This plot shows the anomaly masks at that timestep, overlaid on the data that was used to generate the mask.
Data plots can also be accessed from a TOS²CA AWS S3 bucket.

In this example, we see MERRA-2 vertically-integrated ice water anomalies, using a (user-selected) threshold of 0.1 g/m² (which is probably too sensitive and ends up connecting a large swath of cloud).

Anomalies 5 and 7 are properly tracked over 3 hours.

Another example of a number of the IR brightness temperature anomalies identified over an hour and a half time period (half an hour before the illustration on the previous slide).

The outlined detections are for a (user-selected) threshold of 235K.
Technical and Science Advancements  TOS²CA-1

- We have also designed and almost completed the implementation of the path-2 phenomenon definition tool
  - “path-2” is the approach where the user chooses a pre-defined list of pre-detected occurrences of an already-tracked phenomenon such as “Topical Cyclone”, “Earthquakes”, “Volcanoes”, “Landslides”, listed by location (e.g. the location of the eye of the hurricane), then allows the user to
    - define the domain of interest (the equivalent of the path-1 connected components, by choosing the threshold value for a single constraining variable Vc (more on this in the following slides)
    - choose a subset in space and time (e.g. only Philippines volcanoes, or only Atlantic hurricanes from 2015 to the present), and TOS²CA will then find the relevant domains and chronology of occurrences as in path-1

Allowing the user to produce a domain of interest from the pre-defined listing is crucial because that is the criterion for the existence or non-existence of relations between different process variables, for example:
  - Anomalous CO₂ and/or SO₂ and/or aerosol emissions have been observed before volcanic eruptions, but one would not look for these correlations in general, instead one would look in regions and time where the surface displacement (Vc in this case) is anomalous
  - For hurricanes, one wants to understand processes “in the vortex” as opposed to “in the environment of the storm”, as well as relations between the two, so … one needs to identify where the vortex ends and where “the environment” begins. This requires thresholding the right constraining variable Vc (wind speed w? rate of decrease of w with radial distance?)

Technical and Science Advancements  TOS²CA-1

- path-2 in the case of hurricanes:
  - From a pre-defined list of pre-detected occurrences of Topical Cyclones, listed by location (e.g. the location of the eye of the hurricane), then system needs to allow the user to
    1. define the domain of interest (the equivalent of the path-1 connected components, by choosing the threshold value for a single constraining variable
    2. choose a subset in space and time

For each hurricane, use its lifecycle depiction as provided by IBTrACS

This includes estimates of the:
- Storm center
- Maximum wind speed
- The radius of the 34kt, 50kt, 64kt winds
- Minimum sea level pressure
  *but no information on the edge*
Technical and Science Advancements  

- e.g. Topical Cyclone Laura of 2020: Path-2 needs to define the domain of interest (the equivalent of the path-1 connected components), by choosing the threshold value for a single constraining variable – which can be estimated from IBTrACS information.
Technical and Science Advancements  TOS²CA-1

- e.g. Topical Cyclone Laura of 2020: Path-2 needs to define the domain of interest (the equivalent of the path-1 connected components), by choosing the threshold value for a single constraining variable – which can be estimated from IBTrACS information

Moral of the story:
- IBTrACS provide the location and intensity of the hurricanes at 6h intervals. Included is also the information on the radius of the 34kt, 50kt, 64kt winds.

- Different geophysical variables show
  - Different sizes for the same hurricane
  - Different asymmetries!!!!

So – where is the edge of the hurricane??

⇒ TOS²CA-1 path-2 allows the user to compare the effects of different definitions

The next four slides illustrate three options …
Radius of the 34 kt winds

- 200 km distance from storm center
- 400 km distance from storm center
- 500 km distance from storm center

Radius of the 34 kt winds
For volcanoes, landslides and earthquakes, we decided to generate our own reference dataset of the constraining variable, namely the "surface displacement".

Currently, "surface displacement" can be generated using an existing NASA-funded system called ARIA (Advanced Rapid Imaging and Analysis project). ARIA generates radar interferograms between repeat passes of the ALOS, Sentinel-1A/B, and starting in 2024 the NISAR radars. ARIA tools can then be used to extract, from the interferograms, the time series of surface displacement (sampled in the user-specified region every 30m, and every 6 days over the user-specified time interval). ARIA archives its calculated time series so that future requests can directly access data for which the time series have already been processed.

We are developing the pipeline to convert a TOS2CA user request into an ARIA request, and retrieve the data (whether it is archived, or needs to be generated to fulfill the request).

We were hoping to have tested the case of the La Palma volcano in time for this review – it is the analysis of this case that allowed us to develop the requirements, scripts and pipelines for this new "surface displacement" variable … this is currently in progress.
Technical and Science Advancements

- All the previous discussion concerned the Phenomenon Definition portion of the project.
- In addition, we are implementing the Data Curation portion, and we have defined the requirements for the Visualization and Data Analysis portion.
- The Data Curation is developing the scripts required to locate, extract and collate the user-requested data at two times surrounding the instant \( t_0 \) of each spatial mask produced by the phenomenon definition: the closest time to the mask previous to \( t_0 \), and the closest time following \( t_0 \). This is in recognition of the fact that these requested data will almost never be available on the same regular sampling as the phenomenon-defining variable, and of the fact that if the user requests more than one data variable (such as \( \text{CO}_2 \) and aerosol loading) these variables will also almost certainly not be available on the same spatial or temporal grid.
- The Visualization and Data Analysis will need to start with a spatial re-sampling of the curated data to a common grid, along with a time interpolation of the re-sampled data at the “before” and “after” times so that the final results are data at the same reference times and on the same reference grid.
- In addition to displaying the resulting interpolated re-sampled data, the processing will allow the formation of scatter plots of any pair of variables, along with the derivation of regression curves (with conditional variances) of one variable on another. These are the capabilities that we plan to include in our analysis toolkit to allow the user to quantify (to first order) any relations between the variables of interest.

Typical goal of the visualization / data analysis: Are \( \text{SO}_2 \) and aerosols related in the yellow areas?

Outcome of the phenomenon definition: chronologies of occurrences – masks (yellow areas):

<table>
<thead>
<tr>
<th>Year/Start Date</th>
<th>Year/End Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>2017-01-28</td>
<td>2017-06-15</td>
</tr>
<tr>
<td>2018-01-31</td>
<td>2018-03-17</td>
</tr>
<tr>
<td>2018-08-18</td>
<td>2018-11-21</td>
</tr>
<tr>
<td>2019-01-13</td>
<td>2019-03-24</td>
</tr>
</tbody>
</table>

Outcome of the data curation: \( \text{SO}_2 \) and aerosol data within each mask:

- \( \text{SO}_2 \) (nearest neighbors in time)
- Aerosol loading

VISUALIZATION and DATA ANALYSIS toolkit to be implemented:
1) Re-sample and interpolate \( \text{SO}_2 \) and aerosols (with option to choose "nearest-neighbor instead of interpolation), in order to enable
2) Plotting one vs the other
3) Plotting the conditional histogram of one (conditioned on the other being in a user-specified range)
4) Finding regressions (linear and others) of one on the other
Presentation Contents

• Background and Objectives
• Technical and Science Advancements
• Summary of Accomplishments and Future Plans
• Actual or Potential Infusions and Collaborations
• Publications - List of Acronyms

Summary of Accomplishments and Future Plans

• Phenomenon definition path-1 implemented
• Phenomenon definition path-2 requirements being defined with single constraining variable,
  will implement in year 2
• Data curation started, should be completed for current data types by the end of calendar 2023
• Visualization and data analysis requirements defined, will implement in year 2
• Start defining requirements for phenomena requiring more than a single variable to define
  (e.g. “atmospheric river of aerosols”)

• Define community education effort, geared towards gathering feedback on
  user-specified options and data analysis toolkit
• Define community outreach effort, to identify launch customers or early adopters
• Prepare presentations for AGU, ESTO forum, NASA Hyperwall

• Start infusion into INCUS project, and knowledge sharing with AOS mission (see slide 30)

• Start architecting natural-language front-end for both Phenomenon Definition and Data
  Curation so that neither requires specific (and restrictive) options menus (which we have
developed from interviews off-line with subject-matter experts)
Actual or Potential Infusions and Collaborations

• Infusion into INvestigation of Convective UpdraftS (INCUS) project, currently in phase B, to
  1) help INCUS evaluate different definitions of convective storms, and
  2) Help INCUS determine effect of environment on storms in the past 2 decades
  We have already given a top-level marketing description of the TOS2CA system to the INCUS PI (Prof. Sue van den Heever, Colorado State U) along with first results of testing with IR and Ice-Water-Path from reanalysis, and INCUS will definitely be the earliest adopter

• Knowledge sharing with the AOS project, to explore the possibilities of conducting empirical studies of correlations between cloud, storm and precipitation processes using the “program of record”

• Potential collaborations
  - ARIA and OPERA projects
  - Future OPERA element producing Vertical Land Motion estimates
  - Space Test Program Houston 8 (STP H-8)
  - Surface Deformation and Change mission (currently in phase A)
  - Surface Topography and Vegetation (currently an incubator)
  - Surface Biology and Geology
  (the last 4 are intrinsically about joint analysis of diverse data)
Presentation Contents

- Background and Objectives
- Technical and Science Advancements
- Summary of Accomplishments and Plans Forward
- Actual or Potential Infusions and Collaborations
  - Publications - List of Acronyms

Publications

- None so far
List of Acronyms

- AGU  American Geophysical Union
- ALOS  Advanced Land Observing Satellite (carrying L-band synthetic aperture radars)
- AMSL  Above Mean Sea Level
- AOS  Atmosphere Observing System mission
- ARIA  Advanced Rapid Imaging and Analysis project
- NISAR  NASA ISRO Synthetic Aperture Radar mission
- OPERA  Operational Products for End-Users (White House mandated, NASA funded)
- SBG  Surface Biology and Geology mission
- SDC  Surface Deformation and Change mission
- STV  Surface Topography and Vegetation mission incubator
Open Climate Workbench (OCW) to support efficient and innovative analysis of NASA's high-resolution observations and modeling data

PI: Huikyo Lee, Jet Propulsion Laboratory

Objective

- We propose to develop OCW as an Analytic Collaborative Framework (ACF) that can power the processing flow of large and complex Earth science datasets and advance the scientific analysis of those datasets;
- Migrate the RCMED database (RCMED) to Amazon Web Service (AWS) and provide observational datasets for the upcoming fifth National Climate Assessment;
- Optimize the scientific workflows for common operations, applying data compression techniques and autonomic runtime system;
- Integrate cross-disciplinary topological data analysis (TDA) for statistically robust analyses of high-resolution datasets;
- Provide a comprehensive web service and supporting documentation for end users.

Key Milestones

- Release of OCW v1.6 with a TDA module
- Release of OCW v1.8 and the Cloud RCMED
- Release of OCW v2.0 with optional data compression, distributed data processing, advanced documentation, and examples
- Transfer of OCW v2.0 to ARC, SMHI, DOE LBNL, and Utah State U.

Approach

As an open-source ACF for climate scientists, OCW v2.0 will run on AWS Cloud with special emphasis on developing two use cases: air quality impacts due to wildfires and elevation-dependent warming.

- Scalable cloud database
- Data reduction framework and workflow optimization
- Topological data analysis
- Interactive web service and advanced documentation

Co-Is/Partners: Alexander Goodman, JPL; Michael Garay, JPL; Colin Raymond, UCLA; Valerio Pascucci, U. of Utah; Manish Parashar, U. of Utah; Yulia Gel, UT Dallas

The overall architecture of OCW v2.0. Blue text represents technology components that will be developed in the proposed effort. Red boxes highlight key infrastructure to provide useful resources to OCW users.
Presentation Contents

• Background and Objectives
• Technical and Science Advancements
• Summary of Accomplishments and Future Plans
• Actual or Potential Infusions and Collaborations
• Publications - List of Acronyms

Open Climate Workbench (OCW) v2.0

Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement by the United States Government or the Jet Propulsion Laboratory, California Institute of Technology.
Objectives

- We propose to develop OCW as an Analytic Collaborative Framework (ACF) that can power the processing flow of large and complex Earth science datasets and advance the scientific analysis of those datasets.
  - Migrate the Regional Climate Model Evaluation System database (RCMED) to Amazon Web Service (AWS) and provide observational datasets for the upcoming fifth National Climate Assessment;
  - Optimize the scientific workflows for common operations, applying data compression techniques and autonomic runtime system;
  - Integrate cross-disciplinary topological data analysis (TDA) for statistically robust analyses of high-resolution datasets;
  - Provide a comprehensive web service and supporting documentation for end users

Two use cases

- **Relevancy Scenario 1: elevation-dependent warming**
  - NASA’s enabling tool to support the US National Climate Assessment
  - Assessment of changes in extreme heat stress for complex topography

- **Relevancy Scenario 2: air quality impacts due to wildfires**
  - Comprehensive analyses of satellite observations and simulations with and without wildfire emissions
  - Impact of climate change on fire danger over the contiguous US
  - Weather conditions influencing wildfire conditions in the Western US
  - "Fire Alarm Digital Twins" led by Thomas Huang
  - "Innovative geometric deep learning models for onboard detection of anomalous events" led by Yulia Gel

- In both use cases, we work with the NASA Earth eXchange (NEX) team from the Ames Research Center.
Presentation Contents

• Background and Objectives

• Technical and Science Advancements

• Summary of Accomplishments and Future Plans

• Actual or Potential Infusions and Collaborations

• Publications - List of Acronyms

AWS Resources

We have deployed and made use of the following AWS cloud computing resources.

• Compute:
  ▪ AWS EC2 instance (m5.2xlarge, us-west-2)
    ▪ Primarily used for prototyping data analyses
  ▪ JupyterHub Server (https://ocw-jpl.org/) on AWS Elastic Kubernetes Service (EKS): similar to the Science Managed Cloud Environment (SMCE) for the Earth Information System (EIS)
    ▪ Running on custom docker image stack with all necessary packages for the team (including OpenVISUS and OCW)
    ▪ Will be the main platform for team members to perform interactive analyses and visualizations
    ▪ K8s Austoscaler is enabled so workflows can be scaled up for much larger datasets

• Storage
  ▪ AWS S3 bucket (mounted as a filesystem to instance via S3-FUSE)
  ▪ EFS shared network storage disks
Scenario 1: Elevation-dependent warming

- Large increases in the environmental stress index (ESI; combination of daily maximum temperature, relative humidity, and shortwave radiance) across the US from the late 20th to the late 21st centuries

- Ongoing research: factors influencing elevation profiles of heat stress for multiple regions and models

- Does a regional change represent trends at individual grid points in each of the seven National Climate Assessment (NCA) regions?

The value added by high spatial resolution

- Climate simulations and satellite observations at finer spatial scales are important for studying climate change at local/regional scale, especially in regions of complex topography.

- It is not well-known whether the benefits of high-resolution datasets outweigh the increased costs in computational power and data storage.

- The tools for quantitatively evaluating the benefits currently don’t exist.

- Q: How high should be the spatial resolution of the new data products from future NASA missions and climate simulations to better understand elevation-dependent warming?
Optimization of workflows: Federated Data Staging

- **Goal**: Enable seamless unified access to a range of data sources by providing a single data access handle and request API.

- **Approach**: Federate instances of the DataSpaces data-staging service to create a directory service for finding data across the federation.
  - Ability to query a subset of an array rather than the whole dataset.
  - Highly scalable, supports the access to datasets on HPC and Cloud using distributed indexing.

- **Status**: Proof of concept implementation integrating external data source for the exemplar use case.
  - Earth science data for wet bulb temperature analyses.
  - Geospatial imagery for FIRE-D access (submitted to NeurIPS 2023).

Operating on Federated Data

- DataSpaces ([https://dataspaces.sci.utah.edu/](https://dataspaces.sci.utah.edu/)) is designed to support dynamic interaction and coordination patterns between scientific applications.
  - DataSpaces permits operations on staged data.
  - Operations across multiple datasets.
  - User-defined data manipulations.

- Perform data operations closer to data.

- Currently, we are extending this capability to support the federated execution of python scripts.
  - Python scripts to calculate wet bulb temperature is executed at the data source.
  - Adding features such as caching Numba compilations.

A variety of sources and consumers share a distributed data storage namespace.
OpenViSUS: Dynamic spatial resolution

Surface pressure data in IDX format

**psl.read** *(time = t, quality)*

Numpy Masked Array

---

**A Seamless Approach for Evaluating Climate Models Across Spatial Scales (Chang et al., 2023)**

- In both observed and simulated precipitation, as the spatial resolution increases, so does the spatial variance in precipitation.
Scenario 2: Air quality impacts due to wildfires

Aerosol optical depth simulated by GISS-E2-1-G

Doubled emissions from wildfires
Control

Difference

Evaluating spatial structures of simulated aerosols: Application of topological data analysis (TDA)

- In terms of the latent topology, two aerosol optical depth (AOD) maps from NASA's MISR and a climate model can be compared via Wasserstein distance (i.e., optimal transport) between their respective persistence diagrams ($D_1$ and $D_2$).

- To the best of our knowledge, there is no quantitative metric to measure similarity/difference in spatial patterns between Earth science datasets at different spatial resolutions.

- Persistent homology (PH), one of TDA tools, allows us to compress two-dimensional AOD maps from MISR and models into persistence diagrams.
Bayesian weighting of climate models for the NCA5 report

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
<th>Domain</th>
<th>Source</th>
<th>Years</th>
</tr>
</thead>
<tbody>
<tr>
<td>TS</td>
<td>Surface temperature</td>
<td>global/land</td>
<td>HadCr</td>
<td>1950–2015</td>
</tr>
<tr>
<td>TS anomaly</td>
<td></td>
<td>global/land</td>
<td>NASA/GISS</td>
<td>1950–2015</td>
</tr>
<tr>
<td>PS</td>
<td>Precipitation</td>
<td>global/land</td>
<td>REGCM with mask</td>
<td>1950–2015</td>
</tr>
<tr>
<td>RN</td>
<td>Snow precipitation</td>
<td>subpolar lat</td>
<td>GISS, GFDL, equivalent</td>
<td>1950–2015</td>
</tr>
<tr>
<td>SSTU</td>
<td>Sea surface temperature</td>
<td>global</td>
<td>USSTOPEX</td>
<td>2000–2015</td>
</tr>
<tr>
<td>SLUT</td>
<td>Sea surface temperature</td>
<td>global</td>
<td>USSTOPEX</td>
<td>2000–2015</td>
</tr>
<tr>
<td>SM</td>
<td>Global run-off</td>
<td>global</td>
<td>DA</td>
<td>1950–2015</td>
</tr>
<tr>
<td>GMT</td>
<td>Global mean temperature</td>
<td>global</td>
<td>HadCRUT5</td>
<td>1850–1900/1985–2015</td>
</tr>
<tr>
<td>Tnx</td>
<td>Warmest night</td>
<td>CONUS+CANADA</td>
<td>Livneh, Hutchinson</td>
<td>1950–2011</td>
</tr>
<tr>
<td>Txx</td>
<td>Warmest day</td>
<td>CONUS+CANADA</td>
<td>Livneh, Hutchinson</td>
<td>1950–2011</td>
</tr>
</tbody>
</table>

Bayesian weighting of climate models based on climate sensitivity

Elias C. Massoud1, Huijun Liu2, Adam Tomcak2, Michael Wolter3

1 Computational Sciences and Engineering Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA
2 Air Resources Laboratory, California Institute of Technology, Pasadena, CA, USA
3 U.S. Geological Survey, Southeast Climate Adaptation Science Center, Raleigh, NC, USA
4 Department of Applied Ecology, North Carolina State University, Raleigh, NC, USA
5 Applied Mathematics and Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

* Corresponding author: massoud@ornl.gov

Abstract

National and international climate assessments are used to synthesize information about climate change and its impacts. Recently, the incorporation of models with high climate sensitivity has resulted in climate change projections that are potentially biased. Various methods have been proposed to alleviate this "hot-model" problem, such as using model simulations or simply rejecting models from the ensemble that are deemed too sensitive. Here, we utilize Bayesian model combination (BMC) to efficiently alleviate the "hot-model" problem without the need to reject any models from the ensemble. By producing a probability distribution of the model weights, the BMC framework allows for estimation of the probability distribution, and therefore uncertainty, of the predicted quantity. We use the weights generated with BMC to project future global mean surface temperature and its uncertainty, and we achieve similar results as previous published works. When using weighted multi-model ensemble approaches, we find that future global mean surface temperatures is expected to increase by an average of 2°C for a low emission scenario (RCP2.6) and by an average of 4°C for a high emission scenario (RCP8.5). These estimates are lower than those published in the Fifth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC-AR5), regardless of the level of climate sensitivity in each model. Our results showcase BMC as a path forward to project future climate change and its impacts.

PI has participated in the Federal Adaptation and Resilience Group (FARG) meetings.

We provided the 5th National Climate Assessment with weights (ωi) to linearly combine N climate models (Mi, M2, ..., MN): ∑ωiMi

OCW will include a BMA package with an example to reproduce Massoud et al. (2023)

An interactive web interface

- https://vizus.cs.usu.edu/app/earth-loca/

Credit: Isaac Cho, an assistant professor at the Utah State University
Fire Weather Index (FWI) climatology in summer

- High to extreme fire danger (> 50) is concentrated in the Southwestern US.
- Higher FWI values in the future when compared to historical observations.
- By considering the interplay between weather and land-surface conditions, FWI becomes a more reliable predictor for fire risk assessment and wildfire management.

Credit: Klariza Madrazo, a senior undergraduate at the Cal State U. LA

Changes to Onset of Fall Wind & Rain Influencing Wildfire Conditions in the Western US

- How likely is strong wind before the onset of rainy season in the western US? The cooccurrence of strong wind and dry conditions can lead to explosive fire growth.
- What is the change in likelihood of strong fall wind before rain according to climate models?

Credit: Graham Taylor, a Ph. D. candidate at the Portland State University
Presentation Contents

- Background and Objectives
- Technical and Science Advancements
- Summary of Accomplishments and Future Plans
- Actual or Potential Infusions and Collaborations
- Publications - List of Acronyms

Summary

- Infrastructure:
  - AWS instance and storage for the project

- Technology development to facilitate studying *elevation-dependent warming*:
  - Data Staging Service to optimize workflows
  - OpenViSUS framework to compress massive datasets

- Technology development to study *air quality impacts due to wildfires*:
  - Topological data analysis toolkit

- Contribution to the Fifth National Climate Assessment:
  - Performance-based weights of climate models
  - Web-based interactive visual analytics

- Ongoing scientific research:
  - Assessment of fire danger using NASA’s statistically downscaled climate projections
  - Meteorological conditions influencing wildfires in the Western US
### Project Schedule

<table>
<thead>
<tr>
<th>Tasks</th>
<th>Year 1</th>
<th>Year 2</th>
<th>Year 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Report</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Objective 1 RCM/ED</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Migrate existing ROMED datasets to AWS</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Add new observations to ROMED and provide examples and documentation</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Milestones: Release of OCW v1.6 and the cloud ROMED (TRL 4)</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Objective 2 Data reduction</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Milestones: write up and submit a conference proceeding on ROMED</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Test the data compression with ROMED and CMIP5-5 and assess the trade-off between accuracy and resolution</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Develop the runtime system for the data reduction framework</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Milestones: write up and submit a conference proceeding on Objective 2</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Objective 3 TDA</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Develop topological descriptors with different complexities</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>TDA application to air quality impacts due to wildfires</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>TDA application to elevation-dependent warming</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Validate topological features in ORBS simulations</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Milestones: Release of OCW v1.6 with a TDA module (TRL 4)</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Objective 4 Documentation</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Create Jupiter hub notebook</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Update and manage the documents on the ROMED and OCW websites</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Manage the Jupiter hub server</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Milestones: Release of OCW v2.0 with advanced documentation and examples</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
</tbody>
</table>

- Release of OCW v1.6 with the TDA module and BMA toolkit
- Development of ACF for making observation/model comparisons more efficient and robust

### Plan Forward

- **August 16-25, 2023:** NASA Summer School on Satellite Observations and Climate Models
- **August 24, 2023:** A meeting with the NASA Earth eXchange (NEX) team at ARC and scientists from the Swedish Meteorological and Hydrological Institute
- **September 25-29:** Hands-on training on how to compare bias-corrected Regional Climate Models (RCMs) with Empirical Statistical Downscaling (ESD) for robust future climate change projections at the International Conference on Regional Climate-CORDEX 2023 (ICRC-CORDEX 2023) in Trieste, Italy
Presentation Contents

- Background and Objectives
- Technical and Science Advancements
- Summary of Accomplishments and Future Plans
- Actual or Potential Infusions and Collaborations
- Publications - List of Acronyms

Actual or Potential Infusions and Collaborations

- Continued collaboration with Lawrence Berkeley National Laboratory and Utah State University
  Application of OpenVISUS and distributed data processing
- Potential collaboration with the Swedish Meteorological and Hydrological Institute (SMHI)
  Machine learning-based statistical downscaling developed at SMHI
  Using OCW, we will evaluate the downscaled products from ARC and SMHI with a special emphasis on the value added by high spatial resolution
- Co-I Raymond will participate in collaborative evaluation and analysis of large climate-model ensemble-based heat-stress projections
- PI and several Co-Is have worked on Yulia Gel’s AIST project applying topological data analysis to evaluate climate models and study the impact of wildfires on air quality
- There are potential overlaps between our project and Seungwon Lee’s AIST project, but the science use cases & integrations are independent. The Bayesian Model Averaging (BMA) toolkit will be integrated into OCW in summer 2023 and used in Seungwon’s project.
- PI has worked on Thomas Huang’s AIST project, Fire Alarm Digital Twins, by providing the L3 TROPOMI datasets of O₃, NO₂, SO₂, and CO.
Presentation Contents

• Schedule
• Financials
• Team
• Background and Objectives
• TRL assessment
• Summary of Accomplishments and Plans Forward
• Actual or Potential Infusions and Collaborations (if any)
• Publications
• List of Acronyms

Publications

• Raymond et al. (2022), Regional and elevational patterns of extreme heat stress change in the US, Environmental Research Letters, DOI 10.1088/1748-9326/ac7343.
• Taylor et al. (2023), Projections of Large-Scale Atmospheric Circulation Patterns and Associated Temperature and Precipitation Over the Pacific Northwest Using CMIP6 Models, accepted.
• Massoud et al. (2023), Bayesian weighting of climate models based on climate sensitivity, under review
• Chen et al. (2023), FIRE-D: NASA-centric Remote Sensing of Wildfire, NeurlPS 2023 Dataset and Benchmarks, under review.
• Chang et al. (2023), A Seamless Approach for Evaluating Climate Models Across Spatial Scales, under review.
• Park et al. (2023), NEX-GDDP-FWI: Downscaled 21st century global fire weather projections, under review.
• Madrazo et al. (2023), The impact of climate change on fire danger over the contiguous United States, IGARSS 2023.
• Five presentations at the AGU fall meeting 2023
## Acronyms

**List of Acronyms**

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>OCW</td>
<td>Open Climate Workbench</td>
</tr>
<tr>
<td>ACF</td>
<td>Analytic Collaborative Framework</td>
</tr>
<tr>
<td>AeroCom</td>
<td>Aerosol Comparisons between Observations and Models</td>
</tr>
<tr>
<td>AOD</td>
<td>Aerosol Optical Depth</td>
</tr>
<tr>
<td>AWS</td>
<td>Amazon Web Service</td>
</tr>
<tr>
<td>CORDEX</td>
<td>Coordinated Regional Climate Downscaling Experiment</td>
</tr>
<tr>
<td>FWI</td>
<td>Fire Weather Index</td>
</tr>
<tr>
<td>GDDP</td>
<td>Global Daily Downscaled Projections</td>
</tr>
<tr>
<td>HEALPix</td>
<td>Hierarchical Equal-Area isoLatitude Pixelation</td>
</tr>
<tr>
<td>IGARSS</td>
<td>International Geoscience and Remote Sensing Symposium</td>
</tr>
<tr>
<td>LLNL</td>
<td>Lawrence Livermore National Laboratory</td>
</tr>
<tr>
<td>MISR</td>
<td>Multi-angle Imaging SpectroRadiometer</td>
</tr>
<tr>
<td>NCA</td>
<td>National Climate Assessment</td>
</tr>
<tr>
<td>NEX</td>
<td>NASA Earth eXchange</td>
</tr>
<tr>
<td>NOAA</td>
<td>National Oceanic and Atmospheric Administration</td>
</tr>
<tr>
<td>PH</td>
<td>Persistent Homology</td>
</tr>
<tr>
<td>RCMED</td>
<td>Regional Climate Model Evaluation System Database</td>
</tr>
<tr>
<td>SMHI</td>
<td>Swedish Meteorological and Hydrological Institute</td>
</tr>
<tr>
<td>TDA</td>
<td>Topological Data Analysis</td>
</tr>
<tr>
<td>TSU</td>
<td>Technical Support Unit</td>
</tr>
</tbody>
</table>