Thomas J. Vandal **Research Statement**

Society is already feeling the effects of climate change which are expected to worsen. Over just the past few years we have witnessed record breaking hurricanes, drastic melting of Greenland's Ice Sheet, and many other natural disasters stressing our infrastructure and ecosystems. This led the 2019 World Economic Forum to report that 7 of the 10 highest likelihood and impactful global risks are directly related to climate change, including categories such as extreme weather events and climate change mitigation and adaptation [1]. Hence, resilience to such risks through climate adaptation are of crucial importance for society to be safe and prosperous in the coming decades. General circulation models (GCMs), developed by agencies around the world, provide long-term projections of the earth's system to better understand these changes. These dynamical models encode the scientific community's best understanding of physics regulating earth's system and are executed on high performance systems. While the fundamental laws of physics are well understood, there are subprocesses and parameterizations that are less clear. Extensive computational requirements to generate simulations is a major limitation on spatial and temporal resolutions, a key aspect to modeling mesoscale processes. Recent advances in deep learning along with the massive datastores of satellite and ground based observations provides a catalyst for science advancements in the earth sciences.

Earth science data comes in many forms that include near-term weather forecasts and long-term climate projections, satellites with varying spatial, temporal, and spectral resolutions, and a sparse networks of ground based sensors. These multi-variate spatio-temporal datasets have complex structure containing physical and dynamical processes. Development of machine learning models that leverage these datasets will improve GCM modeling and satellite monitoring capabilities. **My interdisciplinary research goal is to develop machine learning tools for earth science datasets to improve our understanding of the climate system.**

Background and contributions: My research in this area during both my Ph.D. and as a research scientist has focused on *super-resolution, Bayesian deep learning, optical flow,* and *time-series data mining* with applications to *climate* and *satellite* datasets. These applications include downscaling of low resolution spatial, temporal, and spectral datasets, virtual sensing, and time-series compression and search at the petabyte scale.

The ill-posed problem of super-resolution occurs often in the earth sciences when downsampling climate models, weather forecasts, and satellite imagery. **My Ph.D. thesis developed, *DeepSD*, a scalable and uncertainty aware Bayesian super-resolution approach for statistical downscaling of climate models** [2], [3]**.** This work showed that convolutional neural networks are suitable to statistical downscaling with higher accuracy and reduced computational needs compared to long standing methods. DeepSD resulted in a **runner-up best paper award**, runner-up best student paper, and student travel award at KDD 2018, an **invited submission** to the International Joint Conference on Artificial Intelligence (IJCAI) [4], and highlighted in the journal *Nature* [5] which has led to considerable interest across machine learning and earth science communities.

At the NASA Earth Exchange (NEX) **I lead our effort to integrate artificial intelligence models into remote sensing processing pipelines developing software and pursuing research in the areas of optical flow, virtual sensing, and high-performance data mining**. In this role I supervise interns in the areas of machine learning, quantum and high-performance computing, and remote sensing. In recent work, I developed a multi-spectral optical flow approach to temporal interpolation to generate synthetic high-temporal resolution satellite imagery [6]. This technique has motivated further exploration of optical flow applied to satellite imagery for studying dense atmospheric motion. A second line of work has developed an approach to emulate atmospheric processing of satellite data using Bayesian residual convolutional neural networks [7]. Using transfer learning, we showed these models can generate synthetic data over the spectral dimension for true color imagery. I am a Co-I on a 2-Year Advanced Information Systems Technology (AIST) grant in the area of high-performance data mining where we are developing a system to compress and search over high-resolution spatial time-series. We have **applied quantum annealing to optimize discrete variational autoencoders for compression of spatial vegetation data using the D-Wave 2000Q quantum computer** housed at NASA Ames Research Center [8], [9]. For searching, we have implemented radius sketch using *Dask* with high-performance computing and visualizations on NASA's pleiades supercomputer.

The interdisciplinary nature of my work has provided opportunities to collaborate with scientists and technologists in a variety of areas including remote sensing, atmospheric science, quantum physics, and high-performance computing in government, academia, and industry settings. At NEX, I collaborate with researchers throughout NASA as well as at the National Oceanic and Atmospheric Administration (NOAA) and National Center for Atmospheric Research (NCAR). I am an elected student member of the American Meteorological Society's Artificial Intelligence committee. As a faculty member I will continue expanding my collaborations to tackle challenging earth science applications with methodological advancements in machine learning.

Super-resolution for Climate Downscaling

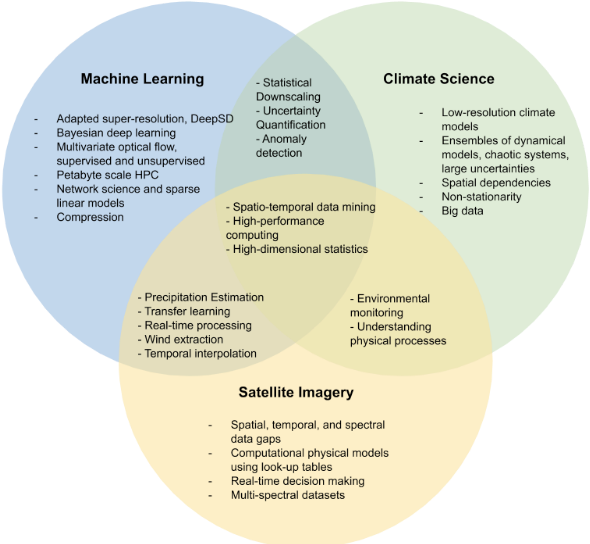
Machine Learning
- Adapted super-resolution, DeepSD
- Bayesian deep learning
- Multivariate optical flow, supervised and unsupervised
- Petabyte scale HPC
- Network science and sparse linear models
- Compression
Climate Science
- Low-resolution climate models
- Ensembles of dynamical models, chaotic systems, large uncertainties
- Spatial dependencies
- Non-stationarity
- Big data
Satellite Imagery
- Spatial, temporal, and spectral data gaps
- Computational physical models using look-up tables
- Real-time decision making
- Multi-spectral datasets
- Statistical Downscaling
- Uncertainty Quantification
- Anomaly detection
- Spatio-temporal data mining
- High-performance computing
- High-dimensional statistics
- Precipitation Estimation
- Transfer learning
- Real-time processing
- Wind extraction
- Temporal interpolation
- Environmental monitoring
- Understanding physical processes

Figure 1: Diagram of core research areas and applications. Each core category discusses data and methodological challenges where the intersections are applications of interest.

Statistical downscaling is a widely applied process to generate high resolution climate projections from coarse resolution simulations at relatively low computational cost. Many statistical and machine learning techniques have been applied to SD, from basic bias correction methods to artificial neural networks. However, these methods are trained in a pixel-wise manner, meaning that a different set of model parameters was learned for every high-resolution location and fail to leverage the spatial information available have both low- and high-resolutions. My dissertation research studied this problem in three phases. First, to understand the current state-of-the-art, **I compared seven traditional and machine learning methods for SD of precipitation under average and extreme conditions** [10]**.** In this study, I found that off-the-shelf machine learning approaches did not outperform simpler approaches. Next, **I developed a novel deep learning approach to statistical downscaling by leveraging super-resolution literature in the computer vision community** . This approach, named DeepSD, is an augmented super-resolution convolutional neural network with a topography auxiliary variable that is trained on observed and gridded datasets [2]. Experiments showed that DeepSD produced a fraction of the error when compared to traditional approaches used in our previous study. SD is an ill-posed problem and hence introduces further uncertainty into the projections. **To produce uncertainties from DeepSD, I used Bayesian Deep Learning to define a stepwise and skewed predictive prior distribution for our applications to science and engineering** [2]. Promising results found from DeepSD and uncertainty quantification continues to produce interest from the climate and data mining communities across the world.

Efficient Processing and Learning from Satellite Imagery

Data from earth science focused satellites are used by scientists around the world to study all aspects of the earth's system. The range of applications, from tracking weather systems, approximating water body size, and monitoring vegetation, are enabled through a diverse set of sensors capturing data at varying resolutions of spectral, spatial, and temporal frequencies. NASA itself has more than 18 satellites currently streaming earth science data with many others from private companies and government agencies. These datasets will be essential to understand changes in earth's system for a long time to come. Fundamental areas driving data science, including large scale data processing and statistics, have historically been applied to make better use of these datasets. However, recent advances in machine learning and high-performance computing have been slow to adapt to these remote sensing domains.

At NEX, we are developing deep learning models to improve data processing and extract new information from geostationary (GEO) satellite imagery. For example, we built an emulator for a radiative transfer model using convolutional neural networks (CNNs). Similarly, CNNs have been effective in learning "virtual sensors" that can predict unobserved spectral bands and expand the range of applications. Combinations of satellites, both GEO and low-earth orbit (LEO), are generating valuable training datasets that we can learn from to build synthetic datasets. **Recent research of ours shows that multi-spectral dense optical flow methods can be used to extract atmospheric motion vectors (winds) from sequences of satellite images.** With collaborators at JPL, NASA, and NOAA, I am exploring the applicability of optical flow for tracking water vapor and clouds along with height assignment. While results are promising, I have found that certain assumptions made for computer vision do not hold in the case of multi-variate and dependent dense flows.

Future Research Agenda

I plan to continue studying how data science can be used and further developed to solve major challenges in climate change adaptation and mitigation. Building on my doctoral studies and knowledge developed at NASA, **the overarching theme of my research will be at the intersection of spatio-temporal machine learning, remote sensing, and climate science.** I hope to continue developing solutions to problems outlined above including resolution enhancement, virtual sensors, and physical model emulation in the following ways:

Unsupervised optical flow methods for multi-variate spatio-temporal data: Earth system models encode physical processes into complex sets of partial differential equations (PDE) which are simulated in spatio-temporal variables. These PDEs are used to approximate the movement of variables such as temperature and water vapor in the atmosphere in a defined gridded and spherical data structure. In computer vision, optical flow is defined as the distribution of apparent velocities of movement of intensity patterns in an image with a strong assumption of constant pixel intensity. These ideas around optical flow can be vastly extended to handle multi-variate dynamics by weakening the constant pixel intensity assumption and incorporating well known physical processes. Furthermore, flows in multi-spectral satellite images, while correlated, vary between bands and hence also does not satisfy the intensity assumption. **My goal is to develop multi-variate and spatio-temporal unsupervised optical flow techniques for both simulated and observed datasets for problems related to extreme weather events and wildfire dynamics.**

GEO-LEO multi-satellite virtual sensors: While the earth sciences already have massive datastores, there will continue to be gaps in both historical and future datasets where synthetic data can provide relief. The GEO-LEO configuration, where high frequency geostationary data are used to generate virtual datasets trained from lower frequency low earth orbit satellites, presents an opportunity for gap filling. For example, the European Space Agency's satellite Aeolous flies a lidar sensor to measure vertical wind speeds of intense weather events. Due to its orbit, Aeolus can only capture one or two tracks through a storm per day. However, this data can be matched spatially and temporally to the closest GOES pixel to generate a training datasets. If a sufficiently accurate model can be developed, for the first time we will be able to virtually generate a dense wind dataset over large regions and high frequencies. Similar problems can be defined to generate an array of datasets using imagery from satellites such as MODIS, Landsat, and CloudSat.

Open-source software development is crucial not only for reproducing of scientific studies but also for allowing users to apply the developed approaches to their specific problems. In the case of the earth sciences, this software must run at scale to handle petabyte scale datasets. The *Pangeo* project is a group of scientists and engineers developing open-source software, such as *Xarray* and *Dask*, for these purposes. **I recently led a proposal with Pangeo which aims to develop machine learning pipelines for spatio-temporal earth science datasets on NEX into popular deep learning libraries.** Successful development of such libraries will greatly improve our ability to reproduce and share large scale machine learning experiments with a common software architecture.

Funding availability for research at the intersection of machine learning and earth sciences can come from many sources. Firstly, NASA solicits funding proposals in technology and earth sciences with an increasing focus on data science implementations and systems. I am currently developing a proposal to NASA's ROSES program with collaborators at NASA and NOAA to generate a wind dataset with optical flow. NASA has expressed interest in further funding opportunities and I look forward to extending collaborations there. I have also been actively involved in proposal writing as a Co-I and student to solicitations from Microsoft AI for Earth, Google AI, Department of Energy, and Schmidt Futures / VERSI (currently pursuing). Other agencies and organizations are also actively funding these areas such as National Science Foundation and the National Oceanic and Atmospheric Administration. As climate change increasingly affects society, I expect even more organizations to open funding in these areas.

Research plans outlined above and previous sections will be dependent on the availability of data, computing power, and funding. My experience and collaborators at NASA will continue to aid in many of these areas including the use of the NASA Earth Exchange (NEX) for computing as needed. Earth science datasets are typically provided by government agencies and are publically available which will limit data restrictions. My past experience in industry, academia, and government agencies provides a strong base for continuing long-term research in the intersection of data and earth sciences.

Citations

[1] World Economic Forum, "Global Risks Report 2019," 2019.

[2] T. Vandal, E. Kodra, S. Ganguly, A. Michaelis, R. Nemani, and A. R. Ganguly, "DeepSD: Generating high resolution climate change projections through single image super-resolution," *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2017.

[3] T. Vandal, E. Kodra, J. Dy, S. Ganguly, R. Nemani, and A. R. Ganguly, "Quantifying Uncertainty in Discrete-Continuous and Skewed Data with Bayesian Deep Learning," *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining - KDD '18*. 2018.

[4] T. Vandal, E. Kodra, S. Ganguly, A. Michaelis, R. Nemani, and A. R. Ganguly, "Generating High Resolution Climate Change Projections through Single Image Super-Resolution: An Abridged Version," *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*. 2018.

[5] L. N. Joppa, "The case for technology investments in the environment," *Nature*, vol. 552, no. 7685, pp. 325–328, Dec. 2017.

[6] T. Vandal and R. Nemani, "Optical Flow for Intermediate Frame Interpolation of Multispectral Geostationary Satellite Data," *arXiv [cs.CV]*, 28-Jul-2019.

[7] K. Duffy, T. Vandal, S. Li, S. Ganguly, R. Nemani, and A. Ganguly, "DeepEmSat: Deep Emulation of Satellite Data Mining," in *SIGKDD workshop on Fragile Earth: Theory Guided Data Science to Enhance Scientific Discovery*, 2019.

[8] A. M. Wilson, A. Michaelis, E. Rieffel, R. R. Nemani, and T. J. Vandal, "Compressing Earth science datasets with quantum-assisted machine learning algorithms," presented at the AGU Fall Meeting Abstracts, 2018, vol. 2018.

[9] M. Wilson, T. Vandal, T. Hogg, and E. Rieffel, "Quantum-assisted associative adversarial network: Applying quantum annealing in deep learning," *arXiv [cs.LG]*, 23-Apr-2019.

[10] T. Vandal, E. Kodra, and A. R. Ganguly, "Intercomparison of machine learning methods for statistical downscaling: the case of daily and extreme precipitation," *Theoretical and Applied Climatology*, vol. 137, no. 1–2. pp. 557–570, 2019.