Blue Waters Professor Allocation Annual Report: February 1, 2017 – January 31, 2018

Title: Satellite remote sensing and 3D radiative transfer modeling for improved weather and climate predictions

PI: Larry Di Girolamo, Department of Atmospheric Sciences, University of Illinois at Urbana-Champaign

Collaborator(s): Shashank Bansal, Michelle Butler, Brandon Chen, Landon Clipp, Soumi Dutta, Dongwei Fu, Yizhao Gao, Yan Liu, Yulan Hong, Jonathan Kim, David Raila, Sean Stevens, John Towns, Kandace Turner, Shaowen Wang, Yizhe Zhan, Guangyu Zhao, University of Illinois at Urbana-Champaign; Mike Garay, Jet Propulsion Laboratory; Alexandra Jones, Princeton University; Ralph Kahn and James Limbacher, NASA Goddard Space Flight Center; Lusheng Liang, SSAI and NASA Langley Research Center; Hyohyung Lee and Kent Yang, HDF Group; and Souichiro Hioki, Yi Wang, Ping Yang, Texas A&M University

Corresponding Author: Larry Di Girolamo, gdi@illinois.edu

Executive Summary:

Our multi-pronged approach for tackling key problems in weather and climate research using Blue Waters, satellite observations and 3D radiative transfer has had an incredible year of success: (1) A peer reviewed publication on a new 3D Monte Carlo Radiative transfer model; (2) a new US-Israel BSF grant to make use of this model; (3) progress on the Terra Data Fusion project with support from NASA, which aims to tackle challenges for data fusion of Terra's five instruments; (4) a new NASA grant to take advantage of the Terra fusion data for meteorological studies in SE Asia as part of the Cloud and Aerosol Monsoonal Processes Philippines Experiment (CAMP2Ex) field campaign; (5) continued enhancements to our understanding of global microphysical properties of water and ice clouds with Terra fusion data that continues to be supported by two NASA grants, with three peer-reviewed manuscripts in preparation; (6) a new initiative using the Terra fusion data to support ML (CNN) cloud detection algorithm development for NASA's MAIA mission; and (7) another peer-reviewed manuscript in preparation that describes and benchmarks the first spectrally integrating, atmospheric 3D Monte Carlo radiative transfer model. Our initial request of 180K NH was on target, but was underutilized on Blue Waters as more of our work was pushed onto ROGER due to some preferred ROGER capabilities.

Description of Research Activities and Results:

Research in weather and climate has massive societal benefits, and indeed has been one of the leading drivers for advancing supercomputing infrastructures. One of least understood and most important aspect of the weather and climate system are Earth's clouds. Clouds cover about 70% of our planet. They are one of the most interconnected components of the Earth System, playing a key role in the Earth's hydrological cycle, regulating the incident solar radiation field more than any other atmospheric variable, and acting as the most important greenhouse constituent in our atmosphere. As such, they modulate a wide range of physical, chemical, and biological processes on Earth. The Intergovernmental Panel on Climate Change (IPCC) affirms that the role of clouds remains the leading source of uncertainty in anthropogenic climate change predictions. In addition, the role of cloud microphysics and cloud-radiation interactions in the timing and intensity of weather events remains an active area of research.

To make headway in reducing uncertainty in weather and climate predictions, the World Meteorological Organization and the IPCC defined a list of Essential Climate Variables (ECVs) requiring global satellite observations (http://www.wmo.int/pages/prog/gcos). It has been established that ~2/3 of the ECVs derived from satellite do not meet accuracy requirements, therefore calling for improvements in the algorithms and technologies used by satellites. For improving algorithms, one of the key recommendations from the NRC 2007 Decadal Survey on Earth Science and Applications from Space (NRC 2007) is clear: *"… experts should… focus on providing comprehensive data sets that combine measurements from multiple sensors."* This, in part, targets NASA's flagship of the Earth Observing System called Terra. Terra was launched in 1999 and continues to collect data for Earth sciences using five instruments: the Moderate-resolution Imaging Spectroradiometer (MODIS), the Multi-angle Imaging SpectroRadiometer (MISR), the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER), the Clouds and Earth's Radiant Energy System (CERES), and the Measurements of Pollution in the Troposphere (MOPITT). Terra data is amongst the most popular NASA datasets, serving not only the scientific community, but also governmental, commercial, and educational communities.

While the need for data fusion and the ability for scientists to perform large-scale analytics with long records have never been greater, the challenge is particularly acute for Terra, given its growing data volume (>1 petabyte), the storage of different instrument data at different NASA centers, the different data file formats and projection, and inadequate cyberinfrastructure. We recently initiated the Terra Data Fusion (TDF) Project, supported under NASA Grant Number NNX16AM07A, to tackle two long-standing problems: 1) How do we efficiently generate and deliver Terra data fusion products; 2) How do we facilitate the use of Terra data fusion products by the community in generating new products and knowledge through national computing facilities, and disseminate these new products and knowledge through national data sharing services? Blue Waters provides the computational resources needed to solve these problems.

The TDF project has transferred the Terra L1B data (~ 1 PB) from the three NASA data centers to ROGER and Blue Waters. This has been challenging owing to the inadequate infrastructure that NASA has for filling and delivering such large data orders. Our work with NASA over the past two years has resulted in a 2.5 increase in NASA's ability to deliver Terra instrument data to the community. Due to the NASA firewall and Globus endpoint issues, the data were first transferred to ROGER before being moved to the BW Nearline system at an unexpected low rate of less than 400 MBps. Nearly twice the size of our allocation on the BW online storage system, the data has to be staged on Nearline and then get processed on Scratch. The slow and frequent data transfers between Nearline and Scratch creates a bottle neck in data processing. We, hence, had been using ROGER as an additional platform for our code development and data processing.

Code development for fusion of Terra instrument data is progressing on schedule. Our so called "basic fusion" product is near completion, and testing on Blue Waters has shown that mission-scale processing requires only ~ 32,000 node hours, which is exciting as it points to our ability to derive other mission-scale products with low amounts of node hours. However, storage cost is high for the basic fusion files at ~2.5-3 PB even with a deflate lossless compression scheme. We have compiled a draft version of the basic fusion algorithm theoretical basis and data specification document temporally posted for download at:

https://1drv.ms/w/s!Agotov0_Ayi7jBBH80y0gJG36_q7

The data format for the basic fusion product is hdf5 with a file structure constructed to comply with Climate and Forecast(CF) conventions, which follows the netCDF-4 data model enabling NetCDF4 tools as well as HDF5 tools to access and explore file contents. All of the granule-level metadata for the entire mission complying perfectly with the NASA CMR format have been generated and stored on the BW nearline system and will be ingested into the NASA Earthdata system for a broad user community to search and access the basic fusion product. We also accomplished a beta version of the resampling and reprojection toolkit, which fuses the Terra data into common grids adopted by any Terra instruments or map projections chosen by users. In addition, we worked with NCSA's Advanced Visualization Lab to dynamically display and project the radiance imageries generated from one single BF granule onto a 3-D Earth as being orbited by Terra for all of the five Terra instruments. Our visualization approach not only helps us to validate and explore the BF data, but also has a profound educational influence among the broader scientific community and general public. An animation clip was created and posted on YouTube:

https://www.youtube.com/watch?v=C2uyjRGwwOs

This was also reported as one of NCSA's notable accomplishments for 2017:

http://www.ncsa.illinois.edu/news/story/ncsas_2017_year_in_review

We have been working on three scientific use cases that help sharpen and push the project forward: (1) monitoring climate change from Terra, (2) retrieving liquid water cloud drop size distribution, and (3) retrieving cirrus cloud ice crystal shapes. The first in a series of planned studies to monitor climate change from Terra was reported in last year's BW report and in Zhao *et al*. (2016).

Results of the second use case, where we fused the Terra MISR and MODIS data to determine biases in cloud drop size datasets that are in widespread use. Our approach builds upon the approach we showed in Liang et al. (2015). In the past year, we extended that study under NASA Grant Number NNX14AJ27G, to examine the underlying causes of the biases. This has become the research thesis of graduate student Dongwei Fu. This required additional code development for large scale analysis on BW. Our analysis has been completed for the month of January, and we are currently extending the analysis to include July. Dongwei's results for January are striking, painting a new view of the global distribution of cloud drop sizes that are now in line with spot measurements had from field campaigns. He has presented these results at several venues and he is currently writing up the results for a peer-reviewed publication and for his M.S. dissertation.

On the third use case, we are working closely with Prof. Ping Yang at Texas A&M on a specialized MISR and MODIS fusion dataset designed for retrieving ice cloud microphysical properties. This work is supported under NASA Grant Number NNX15AQ25G. The codes for this are now complete, and we have produced one year of data for Prof. Yang and his students for analysis. Data transfer between BW and Texas A&M using Globus worked very well. Early results from his group were presented at several meeting and conference venues (see below), showing an altitude (hence temperature) and regional dependence on ice crystal structure.

The TDF project has many big data science issues that are common throughout the sciences and that are part of larger discussion underway by the National Data Services consortium. As such, the TDF project has been identified as a use case under development for the National Data Services:

http://www.nationaldataservice.org/projects/
https://nationaldataservice.atlassian.net/wiki/spaces/NDSC/pages/4358159/Collaborative+Projects+Pilots

Blue Waters, the Terra data, and other satellite datasets are also being used in analysis supporting the meteorological studies in SE Asia as part of the Cloud and Aerosol Monsoonal Processes Philippines Experiment (CAMP2Ex) field campaign, with NASA assets being deployed from the Philippines starting in July 2018. This research is being carried out under NASA contract 80NSSC18K0144. Currently, postdoc

Yulan Hong and student Soumi Dutta have been examining MISR and MODIS data from Terra to ascertain cloud characteristics for the study region, with initial results shown at last week's CAMP2Ex Science Team meeting at Cal Tech.

We are also using BW and the Terra data to support the development and testing of the cloud detection algorithm for NASA's upcoming Multi-Angle Imager for Aerosol (MAIA) mission under JPL contract 1586704. I'm a Co-I on this mission, whose measurements will be combined with population health records to better understand the connections between aerosol pollutants and health problems such as adverse birth outcomes, cardiovascular and respiratory diseases, and premature deaths. Cloud detection issues will be a major source of error for this mission, as well as other research listed above using Terra data. Postdoc Yizhe Zhan is prototyping a new deep learning approach using TensorFlow on BW, with assistant from Aaron Saxton in NCSA. The project only recently begun using Terra data as our testbed for MAIA. Leveraging on the GPU nodes on BW, our preliminary results using a 4-layer CNN network show promising results with respect to MODIS operational cloud detection method. The CNN successfully learns the filters that were hand-engineered in the traditional algorithms, and the consistency reaches up to 90% after testing on independent MODIS images. Moreover, this CNN approach indicates a great potential in identifying clouds over traditional hard-to-identify areas (e.g., sun glint region). Thus, in the coming year, we plan to pursue our research on deep learning applications in identifying clouds from aerosol and snow/ice. For the latter, the accuracy of current operational cloud masks is as low as 81.2% (Ackerman et al. 2010).

In addition to the satellite data processing and analysis work described above, we also continued to work on advancing our 3-D radiative transfer models and research involving these new models on BW. We completed a manuscript on the design and verification of the first open source, publically available, benchmarked 3D monochromatic 3D Monte Carlo radiative transfer model for atmospheric science research. It has been accepted for publication in the Journal of Atmospheric Sciences, with an electronic early release version available for download (Jones and Di Girolamo 2018). We have secured a collaborative grant with colleagues at Technion – Israel Institute of Technology under the Binational Science Foundation (BSF grant 2016325) to use this model as a forward 3D radiative transfer model in solving new tomographic approaches in retrieving cloud properties from multi-angle measurements.

We are also working on finalizing a manuscript on the first spectrally integrating, Monte Carlo 3D radiative transfer model that includes both internal and external sources, and accounts for absorption, emission, and scattering by the atmosphere, clouds, and the surface. The accuracy of our RT model has been verified with extensive comparison to analytical solutions and results from the world's most advanced 1D Line-by-Line RT model. Our model is now ready to act as the first 3D broadband standard of accuracy for comparison by other RT models that make simplifying assumptions. It will be released for public use and development as a community model upon publication. This model will be used to compute full 3D radiative heating rates for the CAMP2Ex project noted above.

Publications and Presentations Associated this this Work

Jones, A.L., and L. Di Girolamo, 2018: Design and verification of a new monochromatic thermal emission component of the I3RC Community Monte Carlo Model. *J. Atmos. Sci. (in press; early online addition at* http://journals.ametsoc.org/doi/abs/10.1175/JAS-D-17-0251.1*)*

Di Girolamo, L., G. Zhao, J. Towns, S. Wang, Y. Liu, and K. Yang, 2017: The Terra Data Fusion Project. *2017 Blue Waters Annual Report*, University of Illinois Press, Urbana, IL

Di Girolamo, L., A. Bucholtz, J. Reid, S. Schmidt, G. Smith, and B. van Diedenhoven, 2018: CAMP2Ex radiation focus area overview. *CAMP2Ex Science Team Meeting*, Jan 23 – 25, Pasadena, CA

Fu, D., L. Di Girolamo, L. Liang, and G. Zhao, 2017: The observed behavior of the bias in MODIS-retrieved cloud drop effective radius through MISR-MODIS data fusion. *American Geophysical Union 2017 Fall Meeting*, December 10-15, New Orleans, LA.

Wang, Y., S. Hioki, P. Yang, L. Di Girolamo, and D. Fu, 2017: Seasonal bias of retrieved ice cloud optical properties based on MISR and MODIS measurements. *American Geophysical Union 2017 Fall Meeting*, December 10-15, New Orleans, LA.

Di Girolamo, L., et al., 2017: The Terra data fusion project: an update. *American Geophysical Union 2017 Fall Meeting*, December 10-15, New Orleans, LA.

Plan for Next Year:

Our work for the upcoming year extends much of the work described above. The Terra Data Fusion (TDF) project is fully supported under NASA Grant Number NNX16AM07A, with additional cloud product R&D supported under for other NASA-sponsored projects: NNX14AJ27G for the cloud drop effective radius of liquid water clouds, NNX15AQ25G for ice crystal roughness for cirrus clouds, 1586704 for the cloud detection for the MAIA mission, and 80NSSC18K0144 for CAMP2Ex. In all cases, allocation on Blue Waters within Di Girolamo's current Blue Waters Professorship allocation was defined. In addition, the OVCR has committed 2 PB of storage on Nearline for the TDF project, and the NCSA director was gracious to commit an additional 2 PB. For the upcoming year, we anticipate 130,000 node hours on Blue Waters for the processing and analysis of the Terra data over all these projects that total 20 mission-scale processing activities. This is based on our experience in current processing of the Terra data, where we estimate 2000 to 32,000 node hours per activity, depending on activity.

The spectrally integrating, atmospheric Monte Carlo 3D radiative transfer model will be made available to the public and published in a peer-reviewed journal. This model has been developed in an object-oriented style, meant to allow for further community development. Before public release, some work to improve the memory utilization and fine tune the performance of the model at scale on Blue Waters will be carried out. Optimization of this codes on Blue Waters is essential since the intent is to make them available to the broader community, some of whom may wish to carry out their research with these codes on Blue Waters. Along with its sister monochromatic version (Jones and Di Girolamo 2018), these models will be utilized to provide highly accurate standards of comparison for other radiative transfer models. This year, these models will be used as part of funded research associated with BSF grant 2016325 and NASA grant 80NSSC18K0144. Additional experiments, highlighting the bias in satellite products due to 3D effects will also be conducted. Based on our experience with these models over the past few years and our planned experiments, we anticipate approximately 80,000 node hours to bring this work to completion.

We therefore request 210,000 node hours for next year. We expect the usage break down by quarter to be the following:

Q1: 20% Q2: 30% Q3: 30% Q4: 20%

The storage requirement for the 3D radiative transfer modeling work is not anticipated to be large. Tables of scattering properties will need to be retained for each unique atmospheric domain, however total storage for those tables and the corresponding output should not exceed 50 TB. The model requires only two input files and produces one output file, so there is no anticipated taxing of the file system expected due to large numbers of files. The radiative transfer model is comprised mainly of logical operations to determine the fate of the bundle of light, i.e. comparisons of random numbers to cumulative distribution functions and simple arithmetic calculations to tally the contribution of each bundle as it travels through the domain. Memory usage will depend on domain size.

The storage requirement for the Terra work will be large. At the moment, we have stored ~0.75 PB of Terra data transferred from the NASA data centers on Nearline and on ROGER and another ~0.25 PB will be transferred from ROGER to Nearline soon and anticipated to be completed within the next 2 months. The compressed Basic Fusion files will take up an additional ~2.5-3 PB. The Basic Fusion files will replace all original Terra data, so there will be no need to keep the original Terra files once the Basic Fusion files are verified and complete. The various mission-scale products that will be derived from the Basic Fusion files are also projected to be about 1 PB compressed. Therefore, with proper data management, we anticipate that our current allocation of 4 PB of Nearline storage will be sufficient for this year.