



United States Department of Agriculture
Agricultural Research Service

USDA-ARS SCINet Newsletter: July 2020

Contents

- **How to Get Started**
- **SCINet Website Update**
- **SCINet User Tips**
- **SCINet Training Program**
- **Research Highlights**
- **Tech Update**
- **Contribute / Contact**

How to Get Started

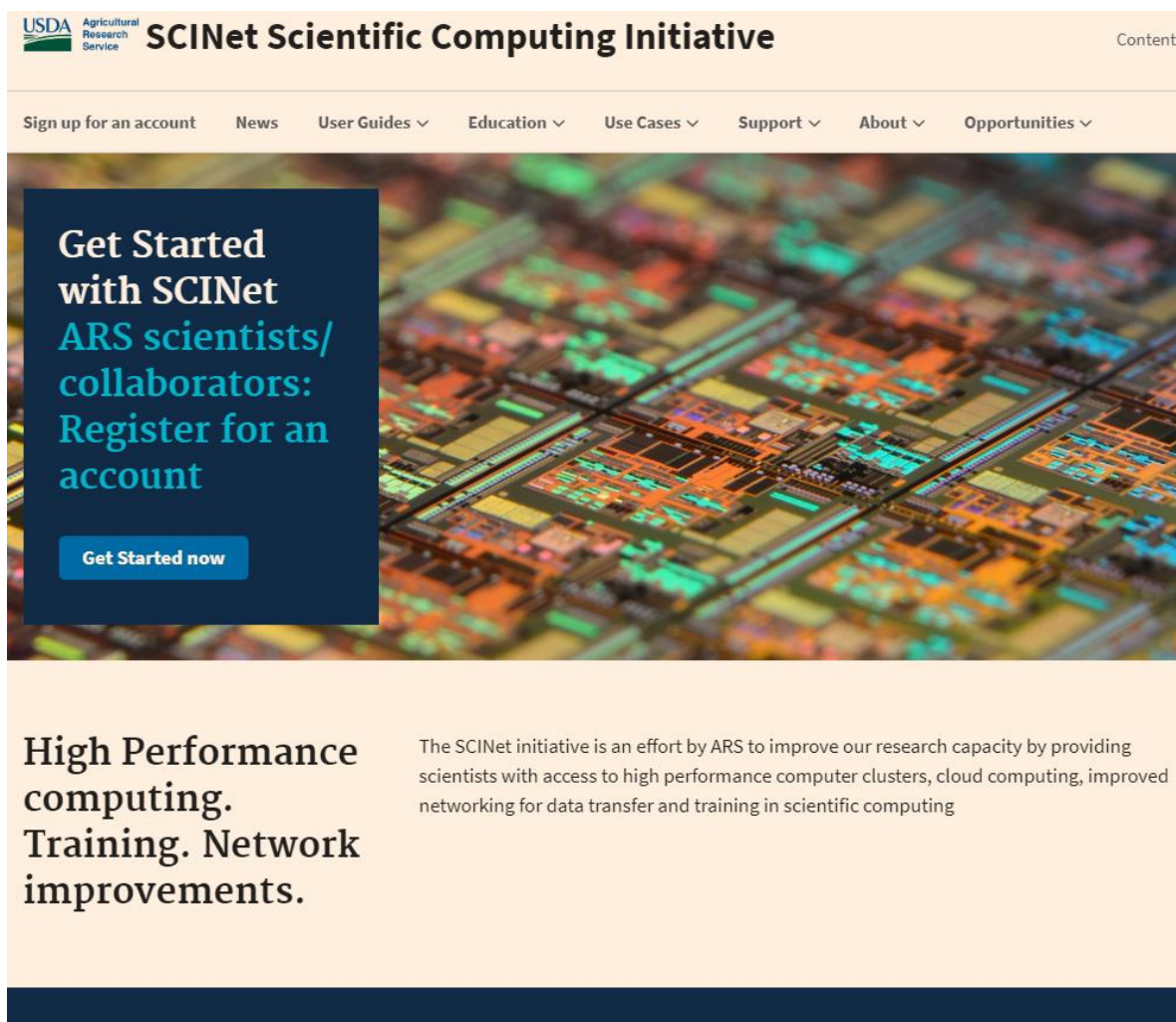


Simply [request a SCINet account](#) (eAuthentication required) to get started. Upon approval, you will receive instructions for logging into SCINet and accessing Basecamp.

Checkout the [new SCINet website](#) for more info on how SCINet can enable your research.

Read the [SCINet FAQs](#) covering general info, accounts/login, software, storage, data transfer, support/policy/O&M, parallel computing, and technical issues.

SCINet Website Update



USDA Agricultural Research Service **SCINet Scientific Computing Initiative** Contents

Sign up for an account News User Guides Education Use Cases Support About Opportunities

Get Started with SCINet
ARS scientists/
collaborators:
Register for an
account

Get Started now

High Performance computing. Training. Network improvements.

The SCINet initiative is an effort by ARS to improve our research capacity by providing scientists with access to high performance computer clusters, cloud computing, improved networking for data transfer and training in scientific computing

New content is constantly being added to the [SCINet website](#). Since the last quarterly newsletter we've added a [Ceres Job Script Generator](#) and new user guide for [accessing Ceres using JupyterHub](#).

Please send any website feedback to SCINet-Newsletter@usda.gov.

SCINet User Tips

Users can access SCINet's high-performance computing through the web instead of the command line.

- **JupyterHub:** A new and easy way to access SCINet HPC resources from your web browser, geared towards researchers who work in JupyterLab or RStudio. Go to <https://jupyterhub.scinet.usda.gov> and use your SCINet username and password to launch a JupyterLab session on the Ceres HPC. You can even upload data from your computer to Ceres with the click of a button. Learn more in the new [JupyterHub user guide](#).
- **Galaxy:** Access to the Ceres HPC from your web browser; geared toward bioinformatics researchers. Galaxy lets you easily upload and share your data, as well as provides access to many bioinformatics analysis tools. Go to <https://galaxy.scinet.science/> and use your SCINet username and password to launch a session. Learn more in the [Galaxy on SCINet user guide](#).

Do you have tips to share?

Email them to SCINet-Newsletter@usda.gov to be included in future newsletters.

SCINet Training Program

SCINet-funded Training

- **The Carpentries R and Python workshops:** Multi-day remote workshops covering Unix, git, and either R or Python from [The Carpentries](#) will be offered in July and August, details on the [Upcoming Events page](#). July workshops are full, but if you're interested in joining August workshops or other upcoming trainings, contact Kathy Yeater.
- **The Carpentries instructor training:** Training will be offered in September 2020 for up to 25 ARS researchers to become certified Carpentries workshop instructors. [This training](#) teaches participants The Carpentries pedagogy and results in the ability to lead Carpentries workshops in Unix, git, R, Python, and more. Contact SCINet-Newsletter@usda.gov for more information.
- **SCINet Geospatial Research Working Group Training Sessions:** The working group will be hosting multiple training and tutorial sessions on Zoom that cover accessing the Ceres HPC through SSH and JupyterHub, basic linux, parallel computing with Python and Dask, computational reproducibility techniques including Conda environments and containers, and a machine-learning tutorial on gradient boosting to predict NDVI dynamics. The sessions will run 1-3 hours each and will occur August 24 through September 3. Agenda and updates to be provided on the group's basecamp site. Contact Rowan Gaffney for access.
- **Coursera.org certified courses update:** The SCINet initiative is still in the process of purchasing licenses for ARS researchers to take Coursera courses with certification. Information about how to obtain a license is expected to be emailed and also posted on the Free Online Training page by fall 2020. Browse the [Coursera](#) website for courses you may be interested in or visit the [Free Online Training page](#) for suggested courses.

Free Online Computational Training (Self-paced)

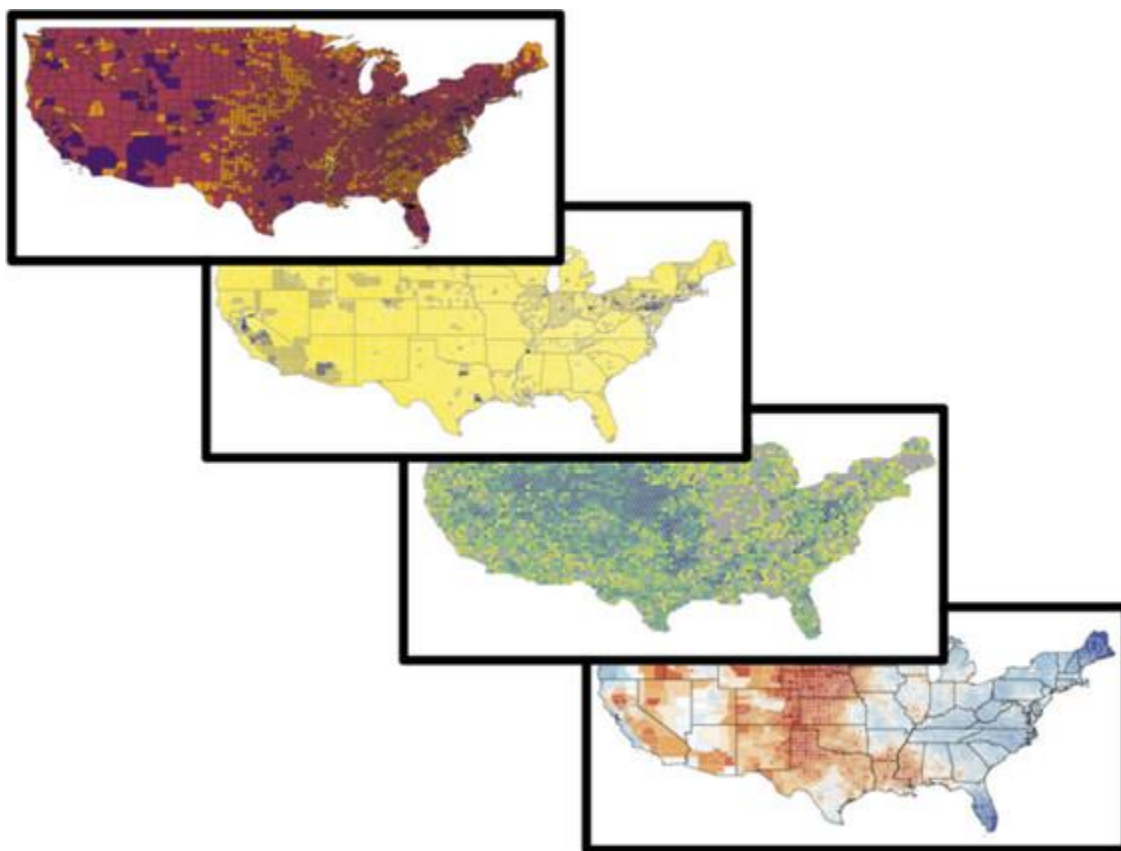
- Make use of your work-from-home time with computational training! A large list of free tutorials and courses has been compiled on the [Free Online Training page](#). Training topic areas include Python, R, SAS, and MATLAB programming; statistics; data science concepts; AI and machine learning; GIS; Google Earth Engine; Git and GitHub; reproducibility, productivity, and integration management tools; and bioinformatics and ecology domain learning. Know of additional free training opportunities? Send them to SCINet-Newsletter@usda.gov.

SCINet Online Science Tutorials

- Browse our growing set of SCINet science tutorials created by ARS scientists and the SCINet Virtual Research Support Core. Our [ARS Science Tutorials page](#) now includes Ceres Onboarding and Intro to Unix for new HPC users, two geospatial computing tutorials, a QTL Analysis tutorial for breeding, and machine learning training material.

Research Highlights

Geospatial Research Workflows in Agricultural Research: when we do (and when we don't) use the HPC



Geospatial research can require different computational needs, typically dependent on the spatial scale and resolution of the problem being addressed, number of data layers, and types of analyses being conducted. ARS scientists tackle a wide range of geospatial research questions that require different computational tools and workflows to do it. Here we highlight workflows and computational requirements for geospatial research from several ARS scientists as examples of the types of research problems that can be solved using local computers (workstations, servers) or those that require a high-performance computing cluster (HPC), such as Ceres, available from SCINet.

Dr. Alisa Coffin classifies land along Georgia's coastal plain suitable for biofuel production using ArcGIS Desktop; Dr. Sarah Goslee identifies drivers of continental agricultural diversity running GRASS/R scripts on local high-performance linux servers; Dr. Rowan Gaffney quantifies cattle grazing preferences using Python containers with the Ceres HPC; Dr. John Humphreys conducts spatial disease modeling across the continental US on Ceres.

GIS on a local computer

Dr. Alisa Coffin is a landscape ecologist at the Southeast Watershed Research Laboratory in Tifton, Georgia, who leverages ArcGIS to analyze geospatial data. ArcGIS is a proprietary geographic information system

developed and maintained by ESRI. She recently used ArcGIS Desktop to identify land along Georgia's coastal plain that would both be suitable for planting biomass fuel crops and would benefit conservation by reducing erosion rates (Coffin et al. 2016). Alisa used the Spatial Analyst module of ArcMap and the "Hydrology Toolkit" to analyze 6GB of raw data (croplands, 10-meter digital elevation models, soils maps, etc.), with intermediate steps producing 118GB of files. Memory capacity problems can arise and sometimes require Alisa to 'chunk' the geospatial data into smaller subsets. Interactive sessions of mapping, modeling, and analysis are ideal for ArcMap on a workstation, and researchers who do not write code may be more comfortable using a software package on their workstation rather than submitting scripts to the HPC.

Alisa saves history logs to document her workflows in ArcGIS. However, as a supervisor, she finds that this method of reproducibility is not ideal because she also needs to access the history logs of the people she supervises. While Alisa did not require the use of an HPC for this analysis, Jupyter notebooks available on the HPC may be a useful tool to improve reproducibility in her workflow. ArcGIS Enterprise 10.7 has a new capability of ArcGIS Notebooks which provide users with a Jupyter notebook environment hosted in the ArcGIS Enterprise portal.

GIS on local servers

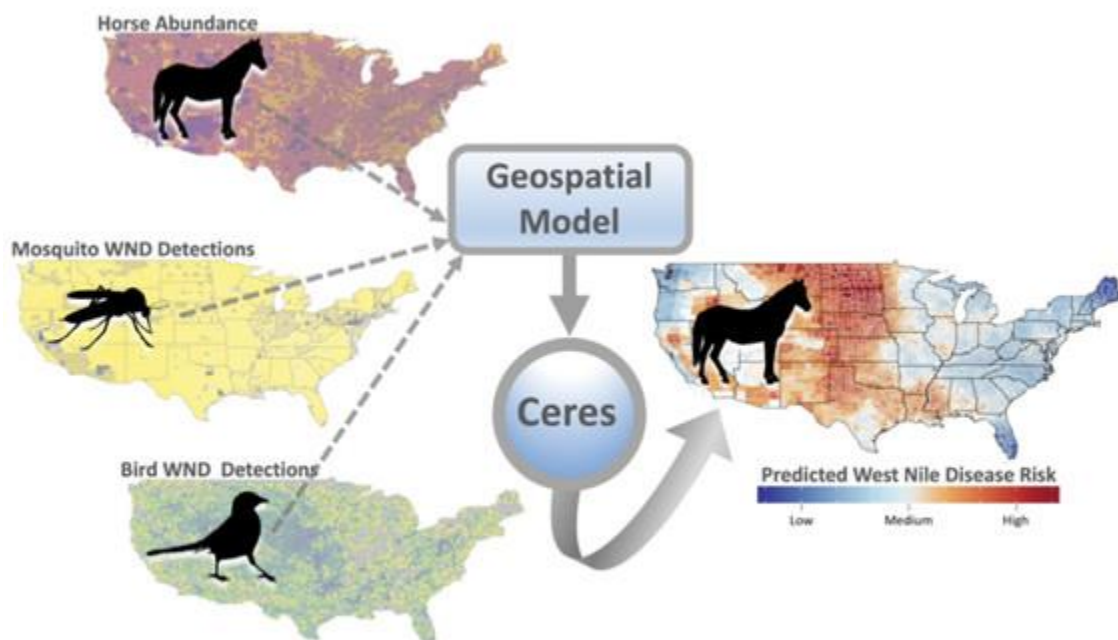
Dr. Sarah Goslee in the Pasture Systems and Watershed Management Research Unit at University Park, PA conducted a continental-scale analysis of drivers of agricultural diversity (Goslee, 2020). Sarah follows principles of reproducible research by scripting in the Open-source software [GRASS GIS](#) and [R](#), and then running these scripts on her local servers. Sarah has a 60-TB hard drive array and manually splits jobs between several high-performance linux servers. In this study, Sarah used a Random Forest model to rank the importance of geospatially resolved climate, soil, and irrigation data to crop diversity.

With over 24 climate, soil, and irrigation data layers being examined for the entire continental United States and aggregated to a common coarseness of 4km for 14 years (2001-2015), a major reason why Sarah has not translated her workflow to the SCINet HPC lies in the time and effort of learning and implementing a new workflow as well as uploading her TBs of data from her machine to the HPC.

Parallel Computing on the Ceres HPC

Dr. Rowan Gaffney in the Rangeland Resources and Systems Research Group in Fort Collins, CO is working with a team to quantify cattle grazing preferences using [hyperspectral imagery from NSF's National Ecological Observatory Network \(NEON\)](#) and GPS cattle collar location data from the LTAR Central Plains Experimental Range (CPER) site. From the NEON imagery, Rowan has developed a high resolution (1-meter) vegetation map over the CPER site (publication forthcoming), which is a major improvement in spatial resolution over traditional methods of estimating vegetation cover classes at the ranch-scale. For each year of the NEON data that he used to develop his vegetation map, there were approximately 26 NEON flights collecting hyperspectral imagery in 426 different spectral bands. Multiplying the number of spectral bands by the more than 200 million imagery pixels covering the CPER site resulted in over 90 billion observations and ~600 GBs in memory per data year to process.

This type of “big data” processing is well suited for the SCINet HPC, which can save a researcher countless hours of computing time versus processing on individual workstations or smaller lab servers. Rowan uploaded the NEON imagery data to the Ceres HPC on SCINet using [Globus](#). He processed the data using an array of Python tools, including xarray, scikit-learn, and dask to run multiple simultaneous computing jobs (parallel processing) on the HPC. To support reproducibility of these efforts, the analysis was conducted in a [singularity container](#), which was [built and archived on DockerHub](#). Additionally, by utilizing a SCINet HPC for his research, Rowan was able to get support from the [SCINet Virtual Research Support Core \(VRSC\)](#) to help install software, build containers, and optimize his research code.



[Spatial Disease Modeling on the Ceres HPC](#)

A fundamental need to meet USDA ARS’s Grand Challenge initiative is to improve agricultural production while reducing the impact of the emerging pests, pathogens, and invasive species that threaten US livestock. Pathogens such as West Nile Virus (WNV), Vesicular Stomatitis Virus (VSV), and others impair livestock health, deplete veterinary resources, and threaten agricultural trade. To better anticipate and prepare for future disease outbreaks caused by viruses, it is essential to model the virus-vector-host interactions and environmental factors that drive disease spread across geographic space and through time. Because disease models must provide high-resolution outputs across expansive geographic extents while simultaneously accounting for the correlations that exist in input variables in both the temporal and spatial dimensions, analyses are often too computationally demanding for traditional hardware and necessitate use of high-performance computing resources like those available through SCINet.

As part of the ARS Predictive Disease Ecology Grand Challenge Project led by Drs. Deb Peters and Luis Rodriguez, a spatiotemporal disease model was developed to forecast future West Nile Disease (WND) outbreaks in horses across the continental US. Postdoc Dr. John Humphreys led the analysis to predict the

distribution and timing of future WND outbreaks. The Centers for Disease Control and Prevention (CDC) records provided the count of veterinary-reported WND cases for horses between 2000 – 2018 aggregated by county. The team used the [USDA National Agricultural Statistics Service database to map horse populations](#), incorporated [CDC mosquito surveillance reports](#) to identify insect vector ranges, and analyzed more than 10 million [bird occurrence records from the Cornell Laboratory of Ornithology](#) to map the distributions of avian species known to host WNV. These datasets allowed the model to link the at-risk livestock population (horses) to times and locations with both the WNV reservoirs (birds) and the WNV vectors (mosquitos) that transmit the virus between those reservoirs and livestock. The research team applied a Bayesian hierarchical modeling framework to construct the model and specified that the prediction for any one location be dependent on the disease risk estimated for surrounding areas and past times (manuscript is in preparation).

After 24 hours of processing on a laptop (Intel Core Processor i9-8950HK, 8 Core, 2.9GHz), the model was only approximately 20% completed, with multiple model versions to be run. Conversely, running the model on the Ceres HPC completed in less than 10 hours, dramatically reducing the processing time and freeing up the laptop for other uses. Preparation for running the model on Ceres entailed [uploading the existing R script](#) and model input data (using [Globus](#)), creating a text file specifying the number of nodes, cores, and memory needed, and then submitting that file to the Ceres job scheduler and management system (Slurm). The team opted to double what was available on their laptop and requested two nodes each with 8 cores. Conveniently, the R package used to run the model (r-INLA) included native multithreading (OpenMP) to handle parallel processing and allowed HPC cores to run concurrently.

In summary, USDA ARS scientists conduct novel geospatial research within the SCINet Big Data paradigm and have numerous SCINet resources at their disposal. Increasing collaboration between researchers and multiple team-member groups can spread the value of using the SCINet HPC along themes of reproducibility and efficiency. SCINet efforts to improve file transfer time, the creation of a data commons, and the support of the VRSC are promising to increase HPC usage among geospatial researchers; these resources are available to other ARS researchers as well.

Contact one of the SCINet team members to find out how we can help you optimize and increase the efficiency of your geospatial research problem through the use of SCINet resources.

References:

Coffin, A.W., Strickland, T.C., Anderson, W.F., Lamb, M.C., Lowrance, R.R., Smith, C.M., 2016. Potential for Production of Perennial Biofuel Feedstocks in Conservation Buffers on the Coastal Plain of Georgia, USA. *Bioenergy Res.* 9, 587–600. <https://doi.org/10.1007/s12155-015-9700-4>

Goslee, S.C., 2020. Drivers of Agricultural Diversity in the Contiguous United States. *Front. Sustain. Food Syst.* 4, 1–12. <https://doi.org/10.3389/fsufs.2020.00075>

Do you use SCINet for your research?

Contact SCINet-Newsletter@usda.gov for a chance to be featured in the newsletter!

Tech Update

The new USDA-Azure [team](#) provides a forum to harness the collective knowledge of [Microsoft Azure](#) cloud, from enterprise admins to scientific computing users, web developers, database administrators, and more. Azure is an alternative cloud option to AWS for SCINet users. Members can share tips & experiences, mention caveats and limitations, post Azure-related announcements and USDA use cases, and seek advice on how to choose from & use the myriad [Azure cloud services](#) available. To request an invite to this team, [click here](#) or email SCINet-Newsletter@usda.gov.

Contribute / Contact

For questions about this newsletter, to contribute content, feedback on the SCINet website, or SCINet policy and development questions please email SCINet-Newsletter@usda.gov.

For technical assistance with your SCINet account, please email scinet_vrsc@usda.gov.

SCINet Leadership Team

Deb Peters, Acting Chief Science Information Officer

Stan Kosecki, Acting SCINet Project Manager

Adam Rivers, Science Advisory Committee (SAC) Chair

Brian Scheffler, Ex Officio

[SCINet Website | Comments](#)

Stay Connected with the USDA Agricultural Research Service
5601 Sunnyside Avenue, Beltsville, MD 20705

