CSDMS 2021: Listening to Community Feedback

Lynn McCready

Gregory E. Tucker

Tian Gan

CSDMS Integration Facility, University of Colorado Boulder

July 2021

Abstract

This report presents results from an online survey of members of the Community Surface Dynamics Modeling System (CSDMS) conducted in 2021. A total of 135 responses were received from community members. Demographics indicate the same lack of diversity that applies across the US geosciences. The survey indicates strong interest in CSDMS' community-building activities, and suggests that CSDMS has succeeded in lowering the barrier to code sharing and access. Continuing technical barriers relate in part to developing and debugging codes for modeling and model-data analysis, and to learning and using software created by colleagues. There is a strong need for cyber-learning opportunities, with desired training modes include multi-day in-person courses, and self-paced online materials. Interest is growing in CSDMS products such as Landlab, and services such as research software consulting. Collectively, the survey highlights continuing needs for community engagement on a variety of levels: more training opportunities; networking and interaction; technical support and assistance; barrier-bridging technologies; and proactive outreach to broaden access to and participation in the Earth-surface process community.

(pre-print; not peer reviewed)

Introduction

What is CSDMS?

The Community Surface Dynamics Modeling System (CSDMS) is a unique national facility that catalyzes new paradigms and practices to understand the Earth's surface—the ever-changing dynamic interface between lithosphere, hydrosphere, cryosphere, biosphere, atmosphere, and anthroposphere. It is a broad community of experts promoting the modeling of Earth surface processes by developing, supporting, and disseminating integrated software modules that predict the movement of fluids, and the flux (production, erosion, transport, and deposition) of sediment and solutes in landscapes and their sedimentary basins. The CSDMS vision is to streamline the processes of idea generation, hypothesis testing, and prediction by (1) supporting and disseminating community-generated software; (2) developing modular tools for efficient model construction, analysis, coupling, and hypothesis testing; and (3) providing training opportunities and resources.

14 Years of Community Service

Quantitative dynamic stratigraphy emerged as a new discipline in the early 1990s, with coastal and deltaic process-response models rapidly evolving from earlier stratigraphic models. The community soon expanded to include geomorphology and modeling of landscape development. These efforts were strongly supported by the International Association of Mathematical Geosciences, the International Association for Hydraulic Research, the U.S. Geological Survey, and NSF panels on stratigraphy, geomorphology, and marine geology and geophysics. At the request of these communities, a workshop was convened in 2002, at the University of Colorado, in Boulder, CO to discuss strategies for development of a community sediment model toolbox with the capability of sharing and coupling model components. A second workshop was convened at the University of Minnesota, in Minneapolis, MN to further discuss rationale/strategy and to develop grand challenge questions, which resulted in a Science Plan and Implementation Plan.

The CSDMS initiative was initially funded by the National Science Foundation in 2007 (**CSDMS 1.0**) to provide cyberinfrastructure for coupling and running community-generated numerical models representing diverse processes and scales across the Earth's surface. CSDMS' overarching goal remains to expand intellectual frontiers by facilitating exploration and modeling of Earth-surface dynamics, to improve understanding of and resilience to short-term hazards and extreme events, and to investigate longer-term landscape and seascape evolution.

Between 2007 and 2012, the general governance structure of the organization was implemented (Working and Focus Research Groups, Executive Committee, Steering Committee, and Interagency/Industry Advisory Panel), the basic cyberinfrastructure was developed (the CSDMS Modeling Tool, Basic Model Interface (BMI) precursor,

Standard Names, HPC for community, and Model and Education Repositories) and educational workshops, clinics, and short courses were offered to community members. During this period, the CSDMS Annual Meeting was also established, and the community grew from an initial 80 members to over 1,000 members in 2012.

Between 2013 and 2018 (**CSDMS 2.0**), efforts were focused on expanding the community and incorporating ecological and social dynamics modeling communities. The CSDMS cyberinfrastructure was expanded to include: the Web Modeling Tool (WMT); BMI 1.0; expanded repositories (Model and Education); and improved metadata.

Between 2018 and 2021 **(CSDMS 3.0)** the CSDMS Integration Facility (CIF) focused on further expanding uptake and developing tools for the Earth surface processes (ESP) modeling community to take advantage of the Earth-surface data revolution. The Python Modeling Toolkit (PyMT) was developed (succeeding the WMT), BMI 2.0 was released, and a Basic Data Interface (BDI) was developed, along with a series of six Data Components that use BDI. The Model Repository expanded to include 279 models, 122 tools, and 35 components, and Landlab was incorporated into the CSDMS Workbench. To further serve the community, a new Research Software Engineering as a Service (RSEaaS) program, an online Help Desk, and a cloud-hosted JupyterHub were launched. The EKT repository was overhauled to include a series of 24 Jupyter-based labs that can be run either locally or remotely using the CSDMS JupyterHub. A webinar series was launched, and numerous short courses and workshops were provided to the community. A pilot version of an Earth Surface Processes Modeling Institute (ESPIn), an 8-day intensive summer training opportunity for students and early-career scientists, was co-hosted by the CIF.

Need for Community Feedback

Since the inception of CSDMS in 2007, ESP science has advanced significantly, and the cyberinfrastructure landscape and supporting software technologies developed to achieve these advances have continued to evolve. Software products developed by the CIF have also evolved to keep pace with community needs. CSDMS continues to develop new technologies, such as the Basic Data Interface for Data Components. Some older CSDMS products, such as the CMT and WMT, have become less popular and have been superseded by newer products like the PyMT. Additionally, as of mid-2021, the CSDMS membership has grown to exceed 2,100 individuals, with increasingly broad and deep domain expertise. The success of CSDMS has also created some growing pains. For example, the working groups (WGs), which were originally envisioned as having 20 to 40 participants, now have hundreds of individuals; the membership of the largest of these, the Terrestrial working group, is now over 1,000. Organization, communication, and management of such large groups is increasingly challenging for the volunteer Group chairs. Because of these challenges, new opportunities and ever-changing community needs, CSDMS conducted a survey in early 2021 to provide members with an opportunity to shape the future of the

CSDMS community activities, computing products and services, and educational opportunities.

Design and Distribution of Survey

The community survey was created in the Qualtrics platform, and questions were developed with guidance from the CSDMS Executive and Steering Committees. The survey consisted of 19 questions (multiple choice, ranking scale, matrix, and open-ended) on community, computing, and educational needs and priorities. Data were collected anonymously due to the sensitive nature of several questions and to encourage honest responses. Initial requests for community participation were sent via email by the Working and Focus Research Group Chairs to their respective groups. These requests were followed by invitations in the community newsletter, direct appeal to the CSDMS mailing list, and by announcement on social media (Twitter). The survey was open for participation for a 30-day period from mid-February to mid-March 2021.

Response Rate

A total of 135 responses were received from community members, representing just under 10% of the 1,450 mailing list subscribers. An average of 20 minutes was needed to complete and submit the detailed survey. Only responses from completed surveys were counted (an additional 16 uncompleted surveys were not included).

Other Community Feedback

Since 2018, the CSDMS Integration Facility has been collecting community input in a variety of ways, including feedback from Group Meetings and Breakout Sessions at the Annual Meetings, surveys from several CSDMS workshops/clinics, survey results from the 2020 and 2021 ESPIn, and the recommendations of the CSDMS Executive and Steering Committees. These additional sources of feedback are also considered in the Discussion and Conclusions Section.

Community Demographics and Interaction with CSDMS

Current Community Demographics

As of July, 2021, the CSDMS community comprised 2,101 members (Figure 1) from academic (majority of membership), government agency and industry sectors. Members hail from 75 countries, with the top five countries being: USA 56%, China and UK each ~5%, India and Canada each ~3%. Members by region/continent are: Europe 19%, North America (and Central America) 59%, South America 3%, Africa 2%, Asia

15%, Oceania 2%. The career level of members ranges from graduate students through to late-career professionals.

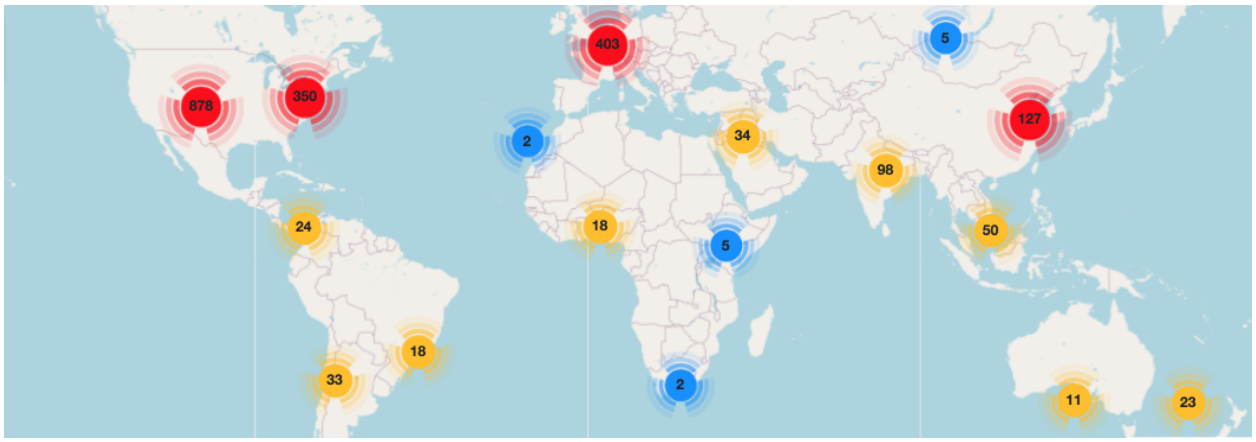

Figure 1. CSDMS Membership Location

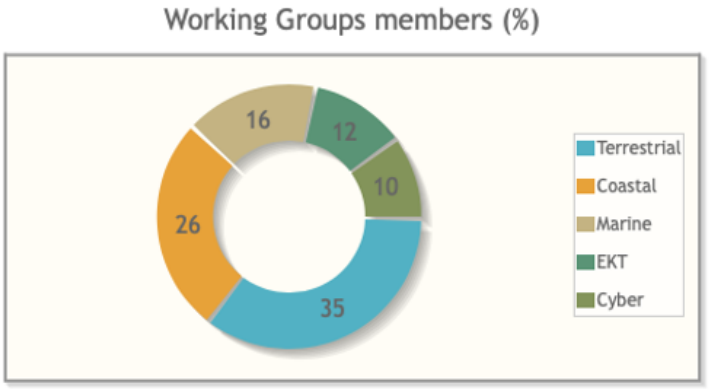

Figure 2. Working Group Membership

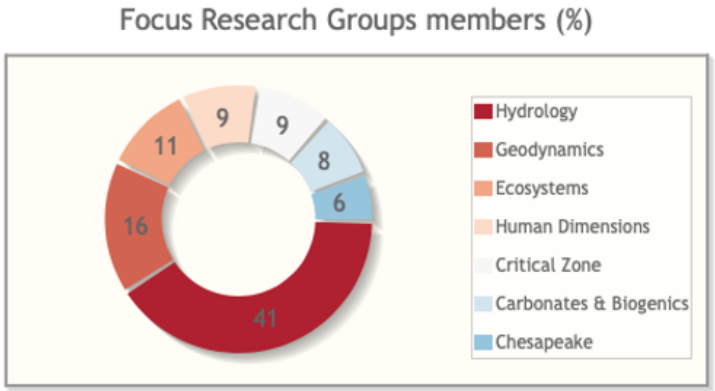

Figure 3. Focus Research Group Membership

CSDMS currently has five Working Groups (Terrestrial, Coastal, Marine, EKT and Cyber) and seven Focus Research Groups (Hydrology, Geodynamics, Chesapeake, Carbonates and Biogenics, Critical Zone, Ecosystem Dynamics and Human Dimensions). Percent of member participation per group are illustrated in Figures 2 and 3.

Survey Demographics

The majority of respondents were affiliated with the academic sector (81%); this percentage is similar to that of the full CSDMS membership. Government agencies were represented by 16% of responses, and other affiliations by 3%. Survey respondents were asked (in an open-ended format rather than selecting from a predefined list) to identify their science domain and primary subfield(s). Among these, Geomorphology was the most frequently mentioned, followed by 2) Hydrology, 3) Geophysics/Tectonics/Geodynamics, 4) Ecology/BioGeo, 5) Sedimentology, Marine, and Cryospheric science, 6) Geochemistry, Human Dimensions, and Natural Hazards (Table 1). Engineering, management, and/or applied science were mentioned in 16% of responses. The domain responses cover the domains associated with all 12 Science Priority Questions highlighted in the NRC's 2020 *Earth in Time* decadal survey report, as well as several of those in the NRC's 2015 *Sea Change* report.

Table 1. Respondent Science Domains and Primary Subfields

Identified Domains	*AGU Sections*	*NSF Programs*
Geomorphology	Earth and Planetary Surface Processes (EPSP), Global Environmental Change (GEC)	EAR/GLD
Hydrology	Hydrology, GEC, EPSP	EAR/HS
Geophysics	Nonlinear Geophysics, Study of Earth's Deep Interior, Seismology, Geo/Paleo/Electromag	EAR/Geophys
Tectonics/Geodynamics	Near-surface Geophysics, Seismology, Tectonophysics, Volc/Geochem/Petrology	EAR/Tectonics
BioGeo/GeoBio & Geochem	Biogeosciences, GEC, GeoHealth	Bio/DEB
Sedimentology	EPSP	EAR/SGP, OCE/MGG
Marine	Ocean Sciences	OCE/MGG
Cryosphere	Cryosphere Sciences	OPP/ANS
Human Dimensions, Planning & Policy, Economics	Science and Society, GEC	SBE Directorate
Natural Hazards	Natural Hazards	EAR/OCE/OPP
Environmental Engineering	GEC, Science and Society	ENG Directorate
Applied Mathematics	All	MPS Directorate
Critical Zone Science, Pedology	EPSP, GEC	EAR

Approximately 55% of respondents were students and early career individuals, followed by 23% mid-career and 22% late career (Figure 4).

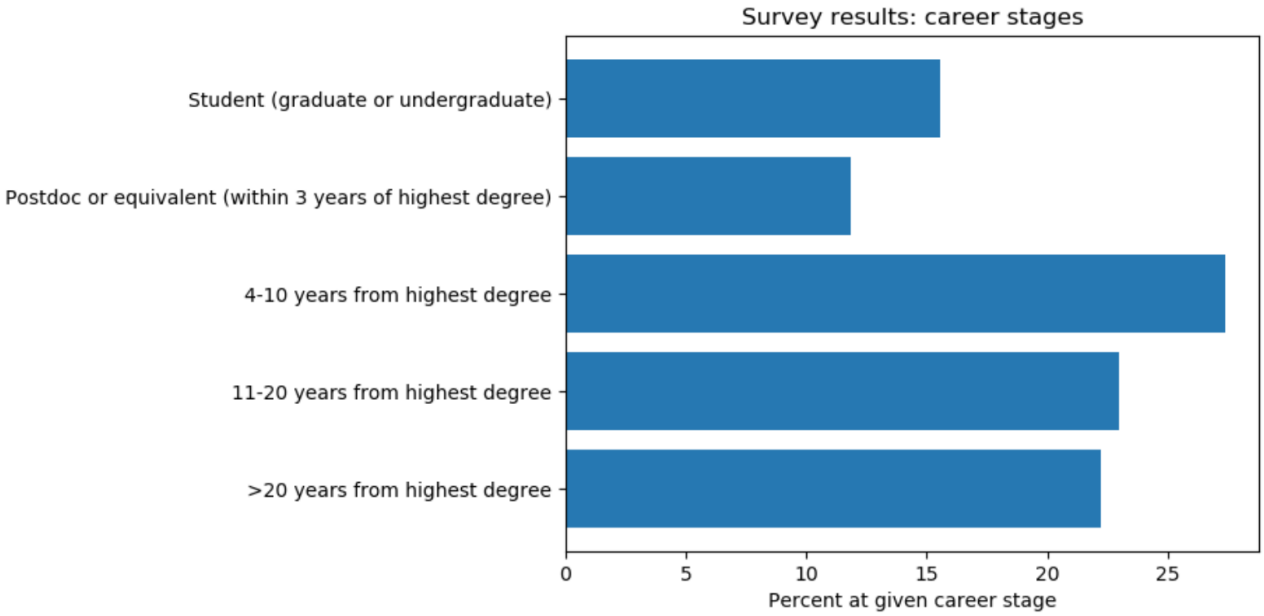

Figure 4. Career stages of survey respondents.

The ethnicity/race of respondents is represented in Figure 5. Approximately 8% of respondents preferred not to report an identity/race. Despite direct appeal to the NABG and SACNAS, no responses were received from geoscientists identifying as Black or Native American.

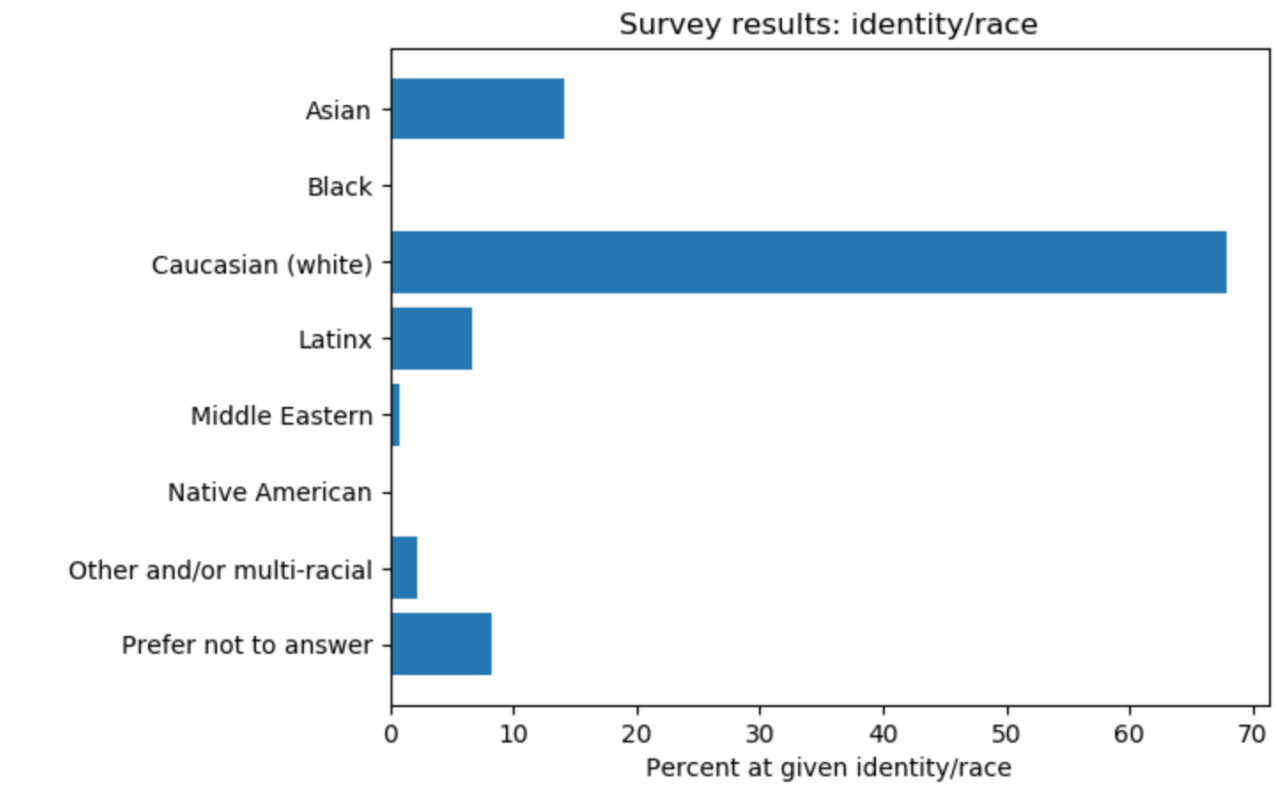

Figure 5. Identity/Race of respondents.

CSDMS Interaction and Usage

The CSDMS web portal is the primary interface with the community. All CSDMS products and services, the model repository, EKT repository, group pages, jobs board, and events are accessed through the portal. Since 2018, there have been an average of ~500 web portal visits per day. The portal currently contains over 50,000 web pages and 5,559 documents. The CSDMS web portal is built on the MediaWiki platform, which enables community members to interact with and contribute content. Survey respondents indicated that the most utilized aspects of the web portal were 1) Model Repository (75% indicating at least one use; 49% indicating occasional or frequent use), 2) Landlab (67% / 32%), 3) BMI (59% / 19%), and 4) Events (50% / 34%) (Figure 6). The least utilized were other pages about products like the BMI Builder/Tester, Standard Names Registry, Dakotathon, Babelizer, and CSDMS services (Research Software Engineering as a Service, proposal support).

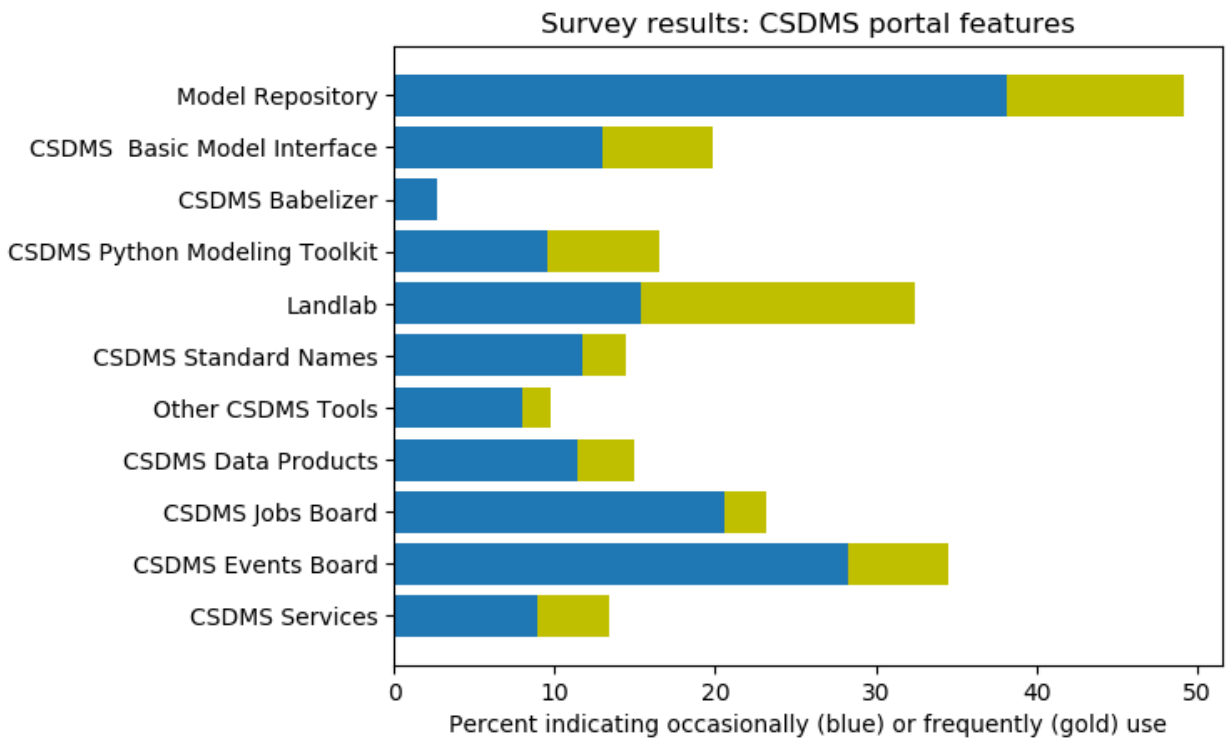

Figure 6. Usage of web portal features.

In addition to the portal, CSDMS facilitates member interactions and network building through a variety of efforts, including the Working and Focus Research Group activities, meetings, workshops, and training events. When asked how important the community-building aspects of CSDMS have been to advance their science, 81% of respondents indicated that it was moderately to extremely important; an additional 17% of respondents indicated that it was slightly important, and only 2% indicated it was not important.

The CSDMS Annual Meeting has been a cornerstone for member engagement for over a decade and has been held in an in-person/on-site format. Recently, due to travel restrictions imposed due to the global pandemic, community members have increased experience with virtual meetings. When asked what meeting format they preferred, respondents ranked a combination of in-person and online format (e.g., plenary presentations open to virtual participation and all other aspects of the meeting, including clinics and breakout sessions open to only on-site attendees) highest. Although many comments noted that in-person meetings were preferable, they acknowledged that funding challenges, timing/work challenges, family responsibilities, carbon footprint, and diversity/equity/inclusion considerations make a hybrid virtual/onsite meeting with broader participation the most beneficial format for the community. Fully virtual meeting formats were indicated to be the least effective/preferred.

Computing Practices and Needs

Languages

Survey results indicated that usage of the Python programming language is increasing and now exceeds Matlab usage, which is in line with other community feedback collected over the past 3 years (Figure 7). Although Julia was listed as the least utilized language, several respondents indicated a desire to learn the language, and based on other feedback there is an increasing desire within the community to test out the language. The TIOBE Index for July 2021 indicated that Julia was ranked 35th of the top 50 programming languages and that downloads increased 87% in 2020 to more than 24 million (Python ranks as #3).

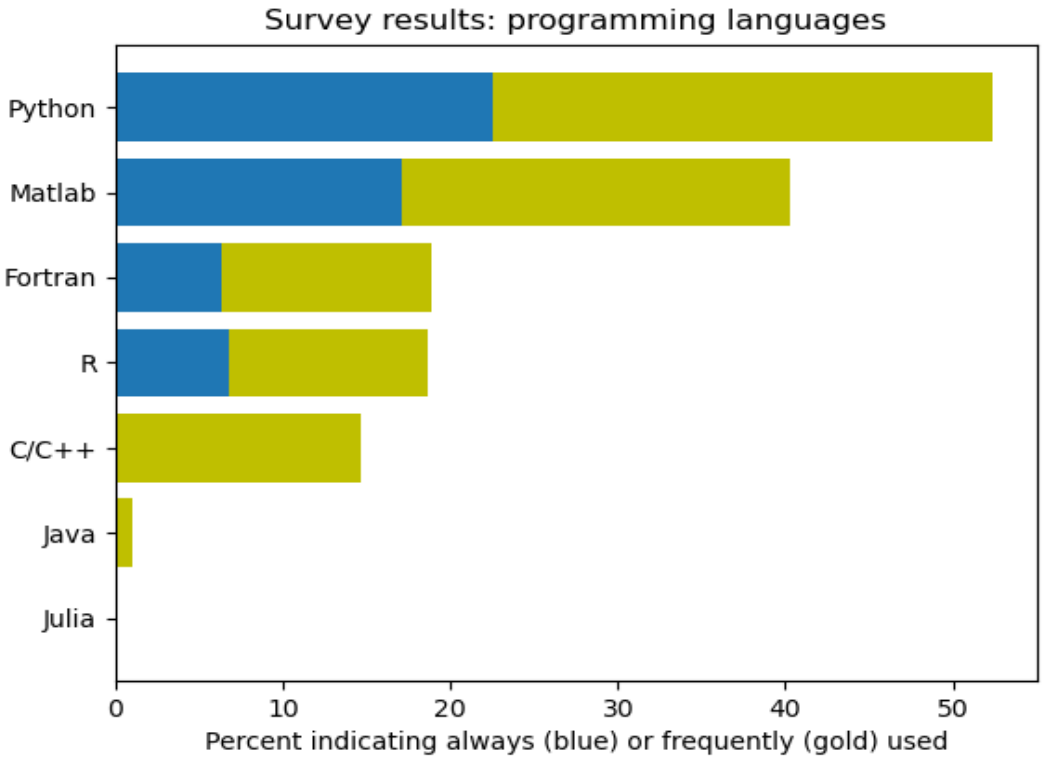

Figure 7. Programming Language Usage

Open Source Software Engagement

Figure 8 illustrates the survey respondent levels of engagement with open source software and models. Many respondents indicated that they use end-user open source software, but only 12% indicated that they contribute code or documentation, review

code, fix issues, or maintain code. 9% reported making feature requests and/or bug reports. Community feedback collected since 2018 has consistently indicated a need for training to increase coding skills and language proficiency.

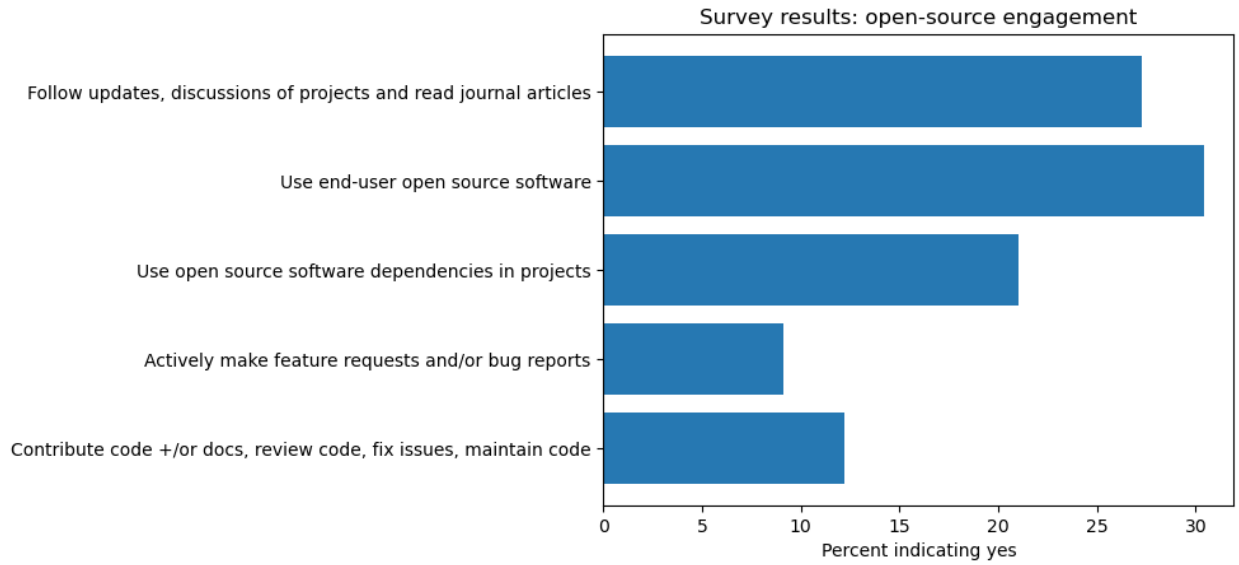

Figure 8. Open-source software engagement

Models/Codes Most Frequently Utilized by Respondents

The respondents' top 5 most used models/codes listed by rank are 1) Landlab, 2) Delft3D, 3) Matlab, 4) QGIS and 5) R. A complete list of responses can be found in Appendix A. A total of 256 codes, packages, data types, and related were listed by respondents. Of these, Landlab was mentioned 23 times, representing nearly 10% of all items. The Delft3D hydrodynamic model was mentioned 12 times. Collectively, hydrodynamic models, including 2d and 3d surface-water flow models, storm-surge models, ocean circulation models, computational fluid dynamics libraries, and similar, were mentioned at least 30 times. Geographic Information Systems and other packages for analysis of geospatial data were mentioned 25 times.

Most Utilized Computing Resources

An overwhelming majority of responses (Figure 9) indicated that the most utilized hardware resource was a desktop/laptop (nearly 99%), followed by a high-performance computing (HPC) facility/resource (46% utilized sometimes or often), and Jupyter Notebooks hosted remotely on servers (38% utilized sometimes or often). Cloud resources were ranked as the least utilized resource (35% utilized sometimes or often).

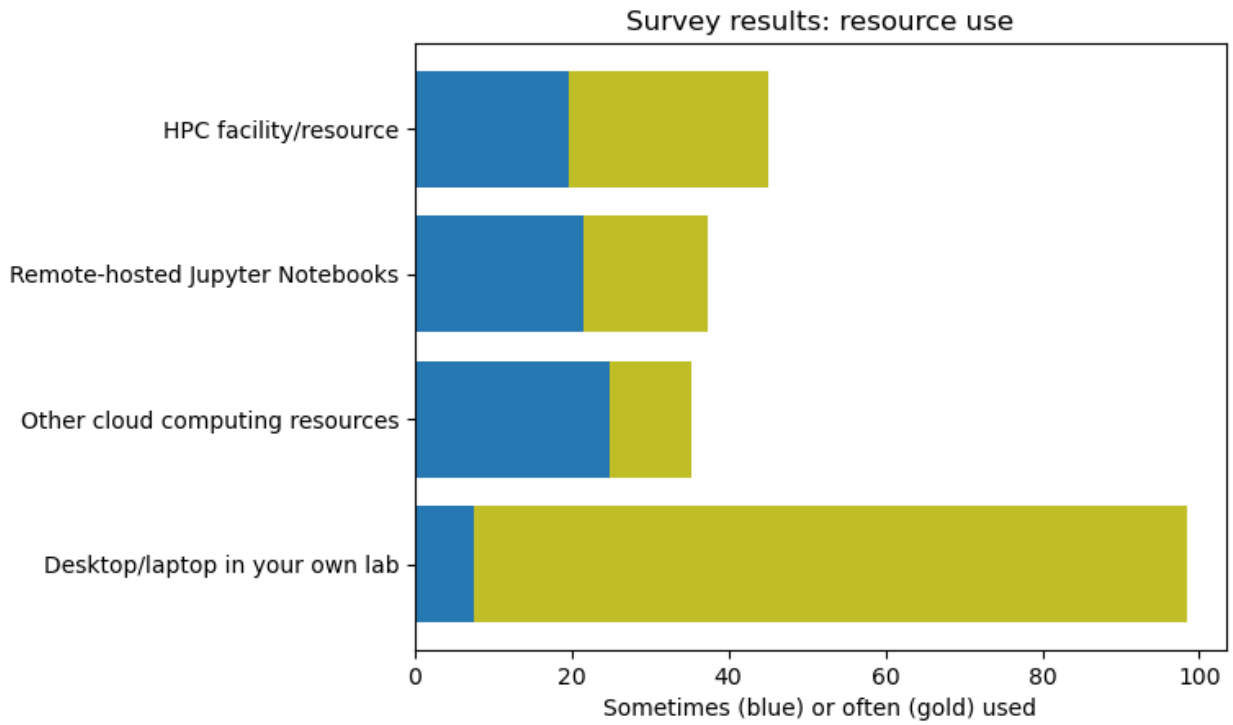

Figure 9. Computing resources usage by survey participants.

Computing Technique Familiarity

Survey participants were asked to indicate their familiarity with and use of: unit testing, workflow tools, reproducibility, benchmarking, continuous integration, and FAIR principles (Findable, Accessible, Interoperable, Reusable). It should be noted that the majority of responses, 65%, indicated that they were either “not at all familiar” or “somewhat familiar” with the techniques listed whereas, only 35% indicated that they were either “familiar and use occasionally” or “familiar and use frequently”. The least familiar/used techniques (Figure 10) were continuous integration (80%), followed by unit testing (73%) and workflow tools (71%). Reproducibility was the most familiar/used technique, with 53% of respondents indicating they are familiar with and use the technique occasionally or frequently. FAIR data and software techniques were utilized by 47% of respondents, followed by benchmarking at 35%.

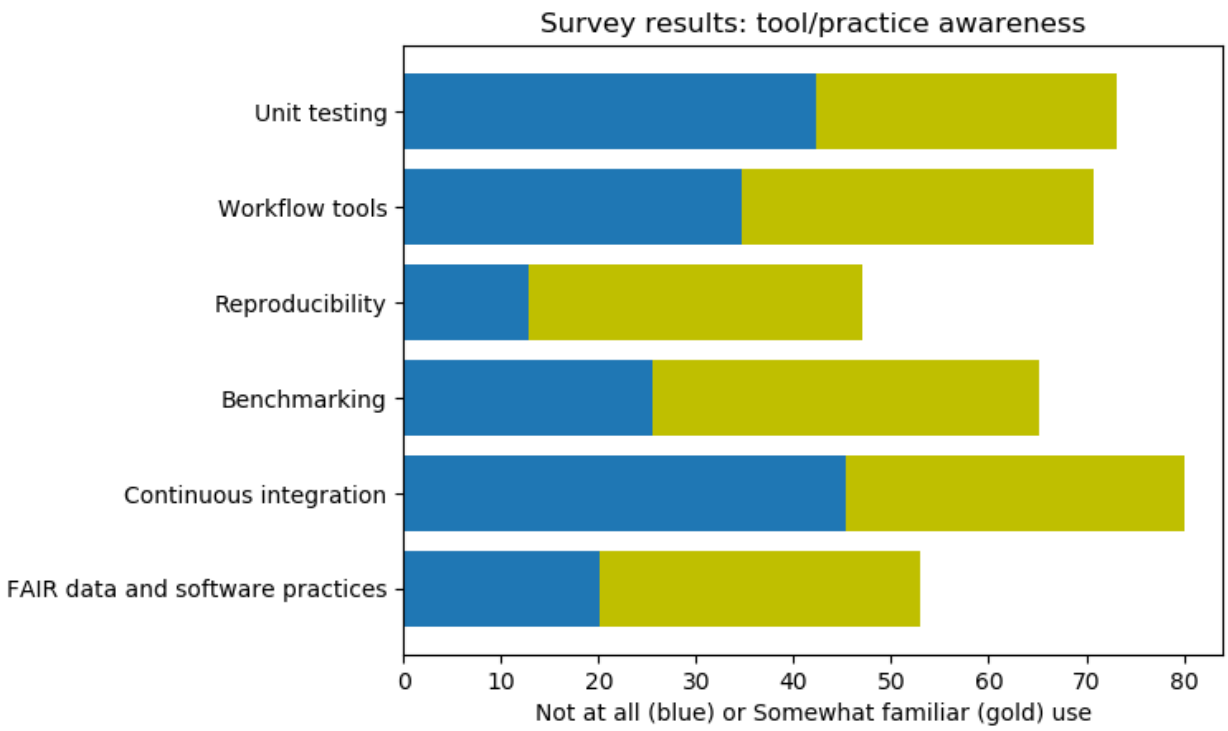

Figure 10. Familiarity and usage of tools/practices.

Barriers to Modeling

Survey participants were asked the following question: “In the coming decade, ‘science questions will increasingly require advancements in high performance computing, improved modeling capabilities, enhanced data curation and standardization and robust cyberinfrastructure that link together observations across many types of records’ (Earth in Time, 2020). What are the biggest barriers to modeling for you and your colleagues”? Writing, testing, and debugging code for a model was cited as a frequent barrier by 60% of respondents (Figure 11), followed by learning how to use a model (45%), and lack of adequate documentation (40%). The least cited barriers were accessing models and HPC resources. Interestingly, two community resources provided by CSDMS are the Model Repository and HPC Blanca.

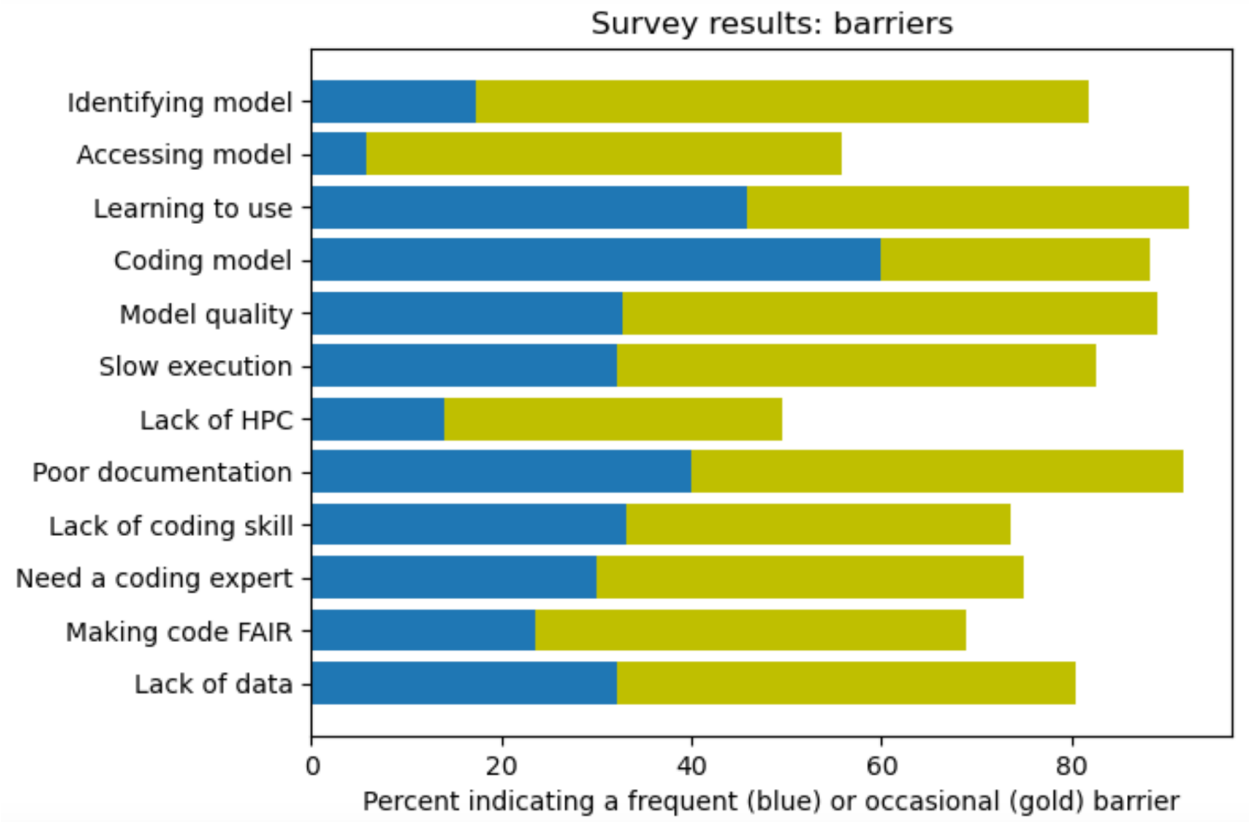

Figure 11. Barriers to Modeling in the Next Decade.

Improvements

When asked, “what improvements to cyberinfrastructure are needed by the Earth Surface Processes Modeling Community to increase open source/FAIR software development and use”, respondents indicated that training and improved recognition/credit for making models FAIR were the most needed strategies to increase open source and FAIR software development and use within the ESP modeling community (Figure 12).

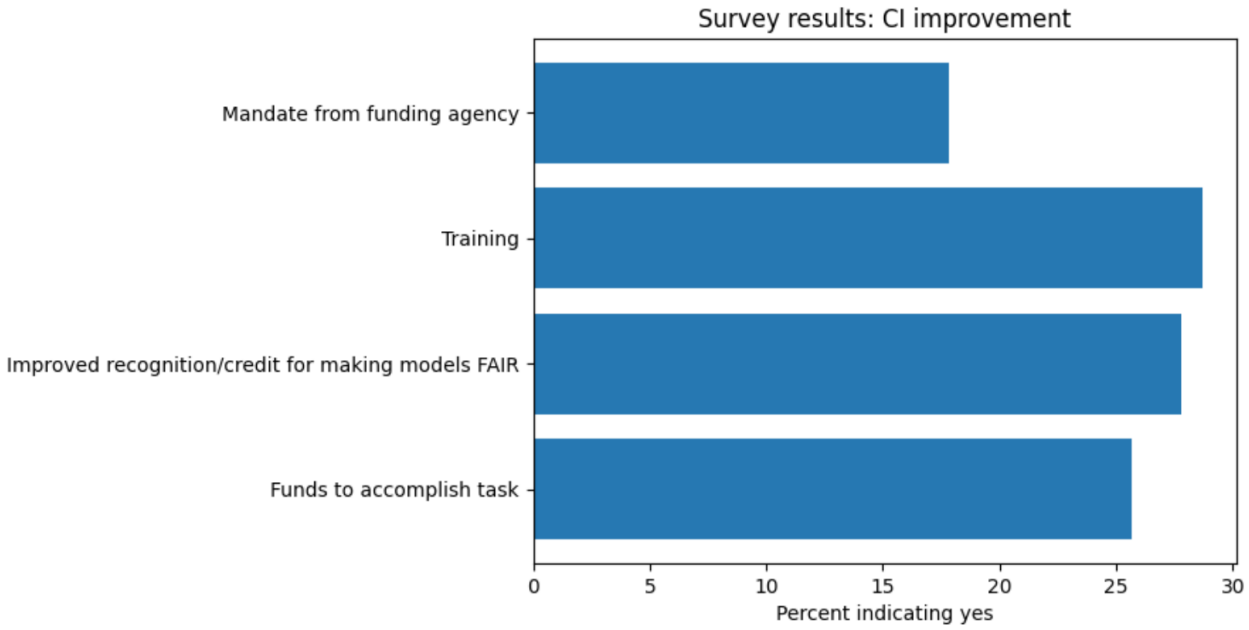

Figure 12. Cyberinfrastructure needed to increase open source/FAIR software development and usage.

Educational and Training Needs

Cyber-Skills Training Needed

Respondents cited model analysis methods (e.g., nondimensionalization, sensitivity analysis), scientific best practices, creating mathematical models, scientific computing tools, and numerical methods as topics with the greatest need for cyber-skills training. Training on web programming and development, image processing/analysis, and data logger/microcontroller programming were listed as the least needed (Figure 13).

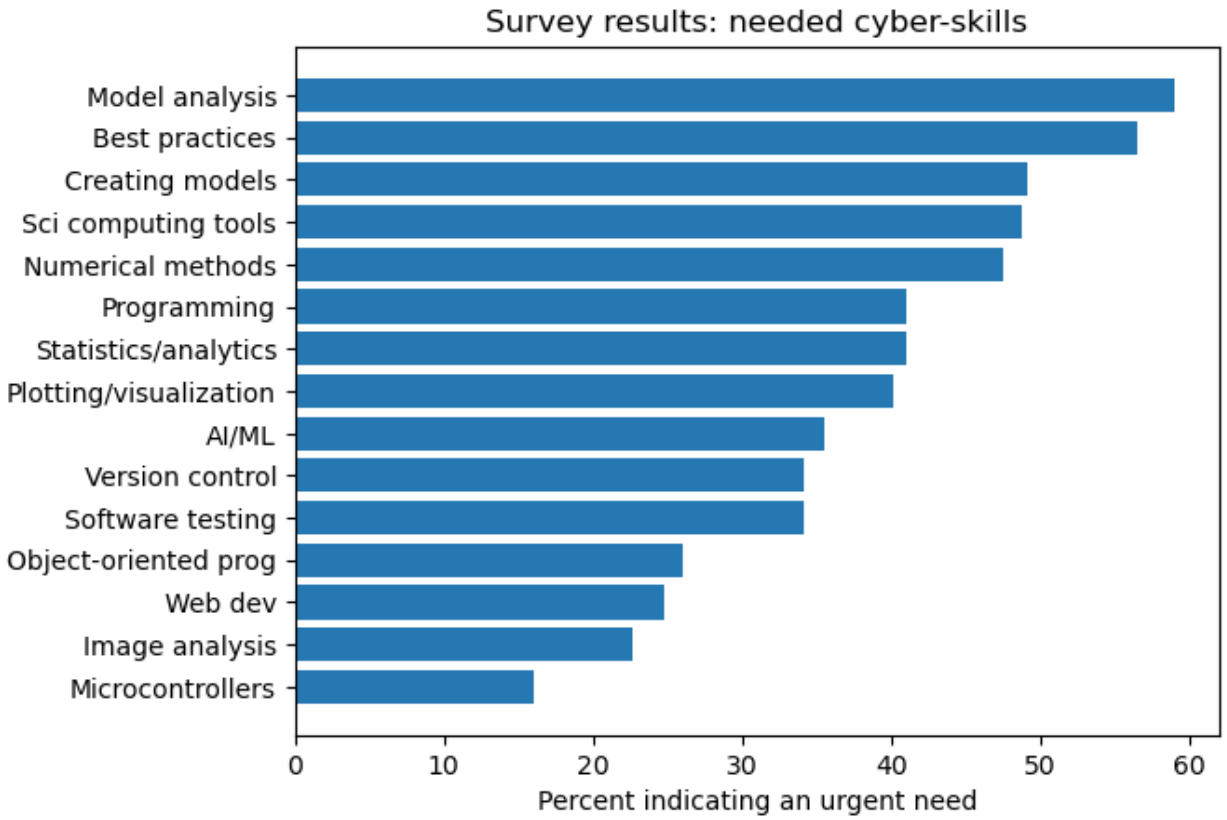

Figure 13. Urgently needed cyber-skills training.

Learning Experience Effectiveness

Respondents were asked to indicate the effectiveness of specific types of learning experience. Extended onsite training (such the summer institute run by the former National Center for Earth-Surface Dynamics, or the pilot 10-day Earth Surface Processes Institute that CSDMS ran in 2020 and 2021) drew the highest number of "effective learning experience" ratings by respondents (52% "effective", 32% "somewhat effective"). This was closely followed by online asynchronous training that is self-paced (47% "effective", 46% "somewhat effective"), and half-day to 2-day modeling workshops provided during a larger conference (e.g., AGU, EGU, GSA, ESA) (46% "effective", 37% "somewhat effective"). Figure 14 combines the "most effective" and "somewhat effective" categories. Online, self-paced asynchronous training emerges as the top rated method when combining categories. Overall, extended not-for-credit and for-credit courses were chosen as least effective.

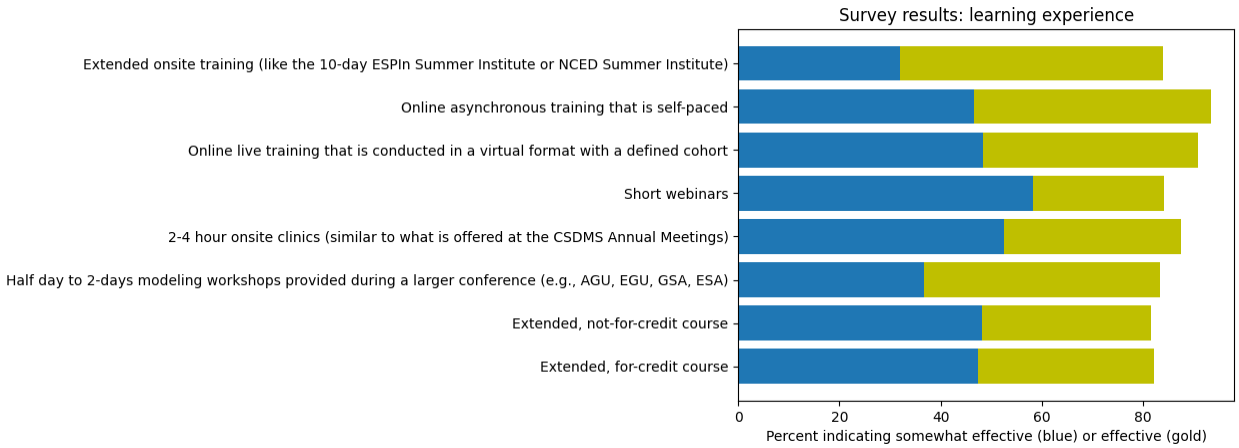

Figure 14. Learning experience effectiveness.

Future Resource Needs and Emerging Topics of Interest

ESP Community Needs

When respondents were asked to indicate the value of specific resources that could be developed and/or made available to the community in the next 10 years, “Research Software Engineering as a Service for community members to assist with model development, model coupling, componentization and model sustainability” was cited by 88% of respondents as a “critical” (42%) or “valued” (46%) community need (Figure 15). Data components (small programs that access data sets) have recently been developed by the CIF utilizing the Basic Data Interface. CSDMS currently offers six data components for community usage, and 72% of survey responses indicated that more data components would be a valued or critical community need. A Jupyter server for community use (teaching and collaboration) and a free community HPC resource were ranked highly by respondents (68% and 69% “critical” or “valued”, respectively). A parallel capability for BMI has been frequently cited by the Geodynamics Focus Research Group as a desired future capability to assist with coupling tectonic and surface processes, and 63% of survey responses indicated that this is a valued or critical need. The lowest-ranked valued or critical need was adding more BMI-enabled models.

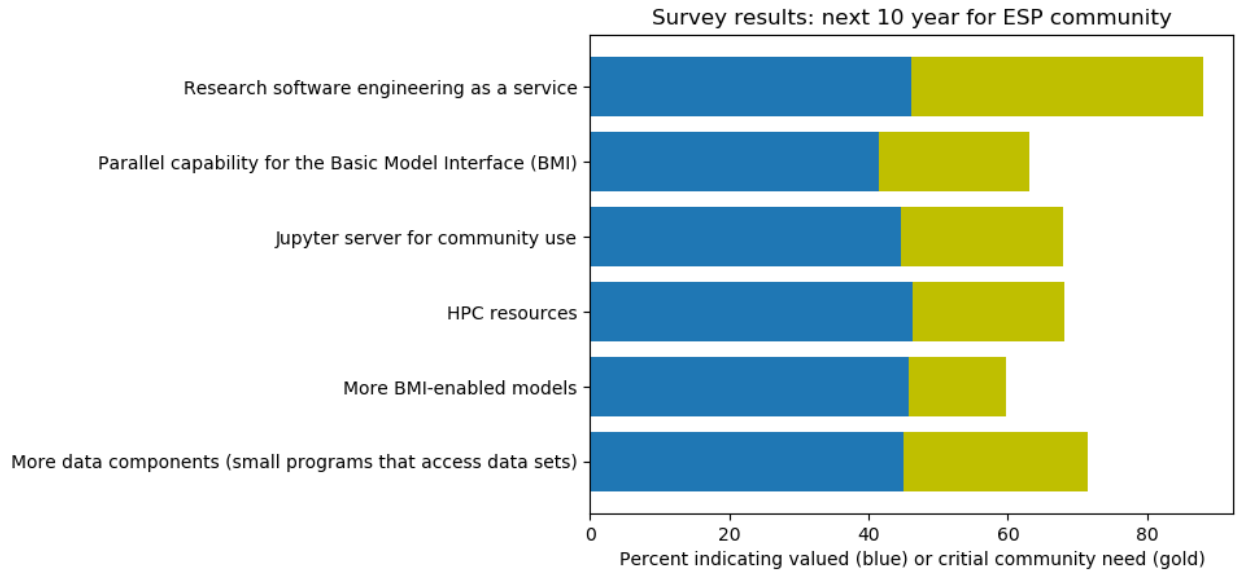

Figure 15. Earth Surface Processes Modeling Community needs in next decade.

Emerging Topics of Interest

In an open-ended question, respondents were asked what new/emerging ESP modeling techniques, strategies, and resources they were most excited about. Based on responses, there is considerable demand for the following:

- Modular modeling systems
- Coupling/integration across systems traditionally studied in different disciplines, and including coupling of agent-based and continuum models
- Artificial intelligence and machine learning applications/tools
- Student and classroom resources

Additionally, there is some demand for the following:

- “Social coding” events / clubs / challenges
- Parallelization
- Tectonics and surface processes
- Up/down scaling
- Julia
- Model-based hypothesis testing
- Landscape evolution modeling

The full list of responses can be found in Appendix B.

Other Community Feedback

Based on community feedback collected between 2018 and 2021 in breakout sessions at the CSDMS Annual Meetings, post-workshop surveys, Executive Committee and Steering Committee recommendations, and ESPIn participant feedback the following common themes/needs were identified:

Community:

- Improved search capability for the model repository
- Facilitation of cross-disciplinary interaction, and a general need to catalyze collaborations (e.g., focused brainstorming sessions on science topics)
- Hybrid format for annual meetings to broaden participation (with virtual participants able to participate in a meaningful way), providing time for informal meetings, and continuing to feature diversity in speakers
- Actively seek out allies at minority-serving institutions and community colleges to recruit into CSDMS membership
- Hackathons, coding camps, and facilitated mentoring

Computing:

- Model coupling/integration across systems traditionally in different disciplines (e.g., tectonic and surface processes, human dimensions and biophysical systems, biological/chemical and physical processes)
- Increase data components, especially with large, long-term efforts like CZOs, NEON, etc.
- Expand Research Software Engineering as a Service, include weekly “office hours”
- Need for model intercomparison information
- Modeling workflows/techniques to cite as a standard
- Establish guidelines/tools for best practices for participatory modeling
- Create ontology for human dimensions models and biophysical models to communicate variables across model types
- BMI with parallel computing capabilities
- Continuous Integration tools
- Expand documentation for repository models and Workbench tools

Education:

- Expand CSDMS EKT repository (online, asynchronous training) with more lab notebooks, mini-seminars on basic computing skills, FAIR principles, and domain modeling exercises/examples (should include some materials at the undergraduate level)
- Make EKT repository searchable
- Extended, on-site training like the ESPIn summer school

- Training on using JupyterHub for faculty/instructors at the beginning of each semester or annually
- Continue to work with existing programs, e.g. RECCS, RECESS, SOARS, to host and mentor undergraduate summer interns

Discussion and Conclusions

Community Demographics

The community served by CSDMS spans a range of career stages, with a relatively even distribution: 27% students and postdocs, 27% early-mid career (within 4-10 years of highest degree), 23% later-mid career (within 11-20 years of highest degree), and 22% later career (>20 years since highest degree). Most of the current community is associated with academia (81%), with most of the rest representing government agencies (16%). Participation from the private sector is limited (3%), which may indicate an opportunity for increased engagement.

Members' identification by racial or ethnic group rings the same alarm bells that have been ringing across the geosciences for some time. Members identifying as Latinx constituted only 7% of respondents, and none of the respondents identified as Black. This finding, while not surprising given what is known about representation in contemporary geoscience, underscores the need for CSDMS to re-double its efforts to engage members of traditionally underrepresented groups in geoscience.

A Lower Code-Accessibility Barrier

When CSDMS was founded, very few modeling codes were openly shared or available to the community at large. The survey results suggest that this barrier has been substantially reduced. Among the technical barriers listed in the survey, "accessing models for download" ranked the lowest, with only 6% of respondents flagging it as a frequent barrier. The fact that the CSDMS Model Repository ranks as the most frequently used resource on the CSDMS web portal suggests that the Repository is at least partly responsible for the increased level of accessibility.

Connecting Across Communities

One of CSDMS' important roles is fostering connections across its constituent research communities. That 81% of respondents ranked the community-building aspects of CSDMS as "moderately" (20%), "very" (40%), or "extremely" (21%) important in advancing their science suggests that CSDMS has been succeeding in this mission. An additional, albeit indirect, indicator is that a relatively large fraction of respondents (34%) reported using the Events Board on the web portal either occasionally or frequently.

One implication of this finding is that there is continuing demand for community-building activities like the all-hands meetings. The finding that the community sees great value in the Annual Meetings is consistent with member reports of new collaborations arising from chance encounters at the Annual Meeting. For example, a recent paper by Barnes et al. (2020) noted in its acknowledgments that "This collaboration resulted from a serendipitous meeting at the Community Surface Dynamics Modeling System (CSDMS) annual meeting, which [the first author] attended on a CSDMS travel grant." Another recent paper arose from a collaboration that began with a chance meeting at the coffee station at a CSDMS Annual Meeting (Auad et al., 2018).

One downside of all-hands meetings in their traditional in-person format, however, is that they require travel, which costs extra time and money (as well as generating a carbon footprint). The need for travel can disproportionately disadvantage those community members who lack the resources, such as academics working at non-R1 colleges and universities, who may find it even harder to obtain travel support than their R1 colleagues. A travel requirement can also disadvantage parents of young children. Remote online meetings can provide an alternative that eliminates travel costs and democratizes accessibility, but at the cost of losing in-person human interaction. When asked about preferred meeting format, a plurality (47%) selected a hybrid in-person and online mode that provides live broadcast for remote attendees. About 1/3 indicated preference for in-person meetings, and 22% preferred online/virtual meetings. Comments by respondents highlighted the dynamic tension between the cost and complexity of travel on the one hand ("with young kids and scarce funding, it is hard"), and the difficulties of an all-virtual format on the other ("the virtual experience is empty for me. No feelings, no real exchanges"). These results imply a need to test out hybrid-format meetings. To be successful, a hybrid format meeting coordination team would need to be intentional about making the remote option work well, with careful attention to audiovisual connections and a clear, easy pathway for interpersonal interaction between on-site and remote participants.

Demand for Landlab

Landlab is a Python-language programming library and component-based modeling framework (Hobley et al., 2017; Barnhart et al., 2020). It was developed to serve as a "build your own model" toolkit that complements CSDMS' growing collection of componentized legacy codes. Where the componentized legacy codes provided through the Python Modeling Tool (PyMT) tend to be complete, stand-alone numerical models in their own right, Landlab was designed to enable researchers to build and experiment with new models relatively quickly. Landlab accomplishes this by providing building blocks for models, such as gridding and input/output functions, together with relatively granular *process components*. Unlike many PyMT components, Landlab components are not stand-alone codes. Rather, they are elements that perform one particular type of task or calculation, such as implementing a numerical solution to a 2D diffusion equation. Landlab components can be imported into a Python program and combined together, along with a grid and layers of data ("fields"), to create an

integrated numerical model. As of mid-2021, Landlab featured in over 40 published journal articles, with applications ranging from tectonic geomorphology to ecosystems.

The survey data indicate growing adoption of Landlab. When asked to list the top three "models or other software [that] are you most involved with or use the most", Landlab was by far the most commonly listed named product, appearing in 23 out of 147 responses (16%). Landlab was also the second highest scoring item among features of the CSDMS web portal (behind the Model Repository). This finding implies a need for continued development and support of Landlab, as well as to encourage more community-wide contributions to the code-review process and to the library itself.

Computational Practices and Needs

The Python programming language has become quite popular in the research communities served by CSDMS. This provides evidence that the Integration Facility's choice of Python as a primary "hub" language continues to be appropriate. In addition to Python, C, C++, and Fortran also continue to be widely used, which suggests value in continuing to provide language-bridging tools such as Babelizer and BMI templates for codes written in these languages. The survey results also show that Matlab continues to be popular in the CSDMS community. This finding implies a need to add Matlab (or its open-source close cousin, Octave) to the menu of CSDMS-supported languages. Julia currently appears to have few users in the CSDMS community, but there seems to be enough curiosity about its potential to warrant providing opportunities to learn more about this relatively new language and its capabilities, through venues such as meeting clinics or webinars.

More than a third of respondents indicated occasional or frequent use of cloud computing in general, and remote-hosted Jupyter servers in particular. The Project Jupyter family of products is relatively new; the name Jupyter was first applied to IPython notebooks in 2015, and the first stable release of JupyterLab was in 2018). The fact that so many respondents report using Jupyter products suggests rapid growth and adoption, and provides motivation for CSDMS' use of Jupyter notebooks for tutorials and teaching, and its recent release of a cloud-hosted JupyterHub server, as well as the recent addition of a CSDMS-equipped Jupyter server to the CUAHSI HydroShare hub server.

Training Needs and Modes

Survey results point toward a need for expanded training opportunities in various aspects of research software and cyberinfrastructure. Among potential themes for enhanced training opportunities, the top four identified by respondents included topics related to programming as well as topics related to numerical modeling and analysis methods. In the first category, 49% identified "tools such as numeric and scientific libraries" as urgently needed, and 57% saw an urgent need for training in scientific programming "best practices". In the second category, 59% of respondents indicated an urgent need for training in "model analysis methods (e.g., nondimensionalization,

sensitivity analysis, optimization)", and 48% reported an urgent need for training in numerical methods. Interestingly, these somewhat advanced topics scored higher than introductory computer programming (which nonetheless was flagged as an urgent need by 41%). For each of the listed skills, apart from "specific programs, packages or libraries", training was flagged as an urgent need by at least 20% of respondents. Overall, the response to this question suggests that universities have not been adequately fulfilling the research community's need for training in and around scientific computing in the geosciences and related areas. There remains a major need for community facilities like CSDMS to help fill this gap.

These findings are echoed in the responses to the "technique familiarity" question (#9), which indicate that many in the community are unaware of, and have little experience with, cyber tools and practices that could increase the efficiency and effectiveness of their computational research. For example, 42% of respondents reported having no familiarity at all with unit testing—a technique that can greatly increase the reliability of research software, and lower the risk of mistakes in the science that derives from it—while only 7% reported frequent use. Similarly, the relatively low percentage who reported contributing to open-source software (9% reported making feature requests and bug reports; 12% reporting contributing to code, documentation, or reviewing) might reflect unfamiliarity with community open-source practices (though it could also mean that the community sees little incentive in such engagement).

When asked about their preferred mode of training, the two most popular options overall were asynchronous self-paced online training (47% "effective" and 46% "somewhat effective"), and extended onsite training (52% as "effective" and 32% "somewhat effective"). Although all of the suggested modes were rated more frequently as "effective" than "not effective", somewhat less enthusiasm was expressed for extended formal courses or short webinars. The perceived value of asynchronous online training is interesting because it is the only one of the suggested modes that is easily scalable to a large audience. Given the interest in both live and asynchronous training opportunities, one potential solution would be to use both approaches in tandem, by deploying materials developed for live sessions as online materials that could be followed asynchronously for self-paced learning.

Technical Barriers

One of the most interesting survey results was the identification of "barriers to modeling". The biggest barrier respondents identified was writing, testing, and debugging code for a model (60% flagged this as "frequently a barrier"; only 12% said it was not a barrier). This response indicates that many community researchers are writing or adapting their own models and analysis codes, and that the task is complex and time-consuming. The majority of respondents also reported encountering barriers in learning to use models (93% occasional or frequent barrier), and in lack of adequate documentation (92% occasional or frequent).

These results indicate that the work involved in creating, modifying, and/or learning to use numerical models constitutes a significant barrier to progress in research. Part of

the barrier lies in programming and code development. This is a barrier that could potentially be lowered with open-source libraries (e.g., packages such as Landlab, which provide computational building blocks), and through greater access to technical assistance (e.g., help desks, forums, or research software engineer consulting programs). In addition, the combination of “lack of adequate documentation” and “learning how to use a model” suggests that researchers often face a steep learning curve when they wish to use an existing community model. Reducing the “time to proficiency” for community models might be accomplished by combining better documentation with short tutorial examples that could be run remotely without local installation.

Cyber Products and Services

When asked about the perceived value of various cyber products and services over the next ten years, respondents expressed particular enthusiasm for research software engineering as a service: 42% identified this as “a critical community need” and 46% indicated it as “valued”. This suggests a community need for direct personal support and assistance with computational work, and underscores the finding that developing and/or learning modeling software presents a significant community challenge. Responses to the open-ended question about CSDMS products and services and community needs indicate continued enthusiasm for modular modeling systems, like Landlab and the Python Modeling Toolkit, that can facilitate research across domains and traditional disciplinary bounds, while lowering technical barriers.

Summary & Conclusions

The landscape of scientific computing is evolving rapidly. In addition to domain-driven innovations from facilities like CSDMS, new developments in computing technology across the sciences, such as interactive notebooks, cloud computing, and collaborative code-development platforms, are opening new possibilities for research. In order to take advantage of these, geoscientists need to know about them, and they need an efficient way to learn the necessary cyber skills. To assess the community's cyber interests, knowledge, and needs, a survey was conducted of CSDMS members in early 2021. A total of 135 responses were received from community members, equating to just under 10% of the 1,450 mailing list subscribers.

In terms of community demographics, survey results indicate the same lack of diversity that applies across the US geosciences, with Black and Latinx scientists underrepresented relative to their fraction of the population. This finding, while not surprising, underscores the need for proactive engagement of historically underrepresented groups.

The survey indicates strong interest in CSDMS' community-building activities, such as all-hands meetings. Results also suggest that CSDMS has succeeded in lowering barriers to code sharing and access. Respondents report that today's technical barriers relate, among other things, to developing and debugging codes for modeling and

model-data analysis, and to learning and using software created by colleagues. The survey findings also indicate a strong need for cyber-learning opportunities, on topics ranging from scientific programming best practices to advanced model analysis techniques. Desired training modes include multi-day in-person courses, and self-paced online materials.

Python has become the most popular programming language in the community, which aligns with CSDMS' choice of Python as a *lingua franca* in its software tools. Among CSDMS products and services, Landlab is clearly growing in popularity. Respondents also expressed strong interest in Research Software Engineering as a Service; the finding that interest in this program exceeds current usage suggests a need for more effective communication about this and related resources.

Collectively, the 2021 CSDMS survey highlights continuing needs for community engagement on a variety of levels: more training opportunities, networking and interaction, technical support and assistance, barrier-bridging technologies, and proactive outreach to broaden access to and participation in the Earth-surface process community. The results motivate further work to address these needs, through a combination of community engagement and modern cyberinfrastructure, using a dynamic, adaptive, and metric-driven approach aimed at building a sustainable community cyber-ecosystem for the mid-2020s and beyond.

Acknowledgments

CSDMS is supported by the US National Science Foundation (1831623).

References

Auad, G., Blythe, J., Coffman, K., & Fath, B. D. (2018). A dynamic management framework for socio-ecological system stewardship: A case study for the United States Bureau of Ocean Energy Management. *Journal of Environmental Management*, *225*, 32-45.

Barnes, R., Callaghan, K. L., & Wickert, A. D. (2020). Computing water flow through complex landscapes–Part 2: Finding hierarchies in depressions and morphological segmentations. *Earth Surface Dynamics*, *8*(2), 431-445.

National Academies of Sciences, Engineering, and Medicine. (2020). A Vision for NSF Earth Sciences 2020-2030: Earth in Time. National Academies Press.

Appendix A

Complete list of models respondents indicated they use most frequently (sorted by frequency of mentions):

Code Name	Number of mentions
Landlab	23
Delft3D	12
MatLab	8
Qgis	8
R	7
ArcGIS	5
FLAC3d	5
TopoToolBox	5
Google Earth Engine	4
RivGraph	4
SWAN	4
telemac	4
ADCIRC	3
ASPECT	3
WRF	3
XBeach	3
ANUGA	2

CEM	2
COAWST	2
InSAR	2
OpenFOAM	2
Python	2
ROMS	2
SiStER	2
SWAT	2
SWIMM	2
3Dec	1
Aquatox	1
ARIES	1
Artificial Neural Networks	1
Badlands	1
Bottom Boundary Layer model	1
CarboCat	1
Caesar-LisFlood	1
CEM2D	1
CESM	1
ChesROMS-ECB	1
CHILD	1
Classification Decision Tree	1
Climate data operators	1

CLM-ml v0	1
Cloudcompare	1
CoOMSOL	1
Corebreakout	1
Cormix	1
CROCO	1
CryoGrid	1
DeaLII	1
DeltaRMC	1
dendra.science	1
DHSVM	1
DiscoverFramework	1
Distributed Hydrological Model	1
Dorado	1
DSAS	1
DSS (agricultural model)	1
Dual SPHysics	1
E3SM	1
Ecopath/Ecosim	1
EFDC	1
EGRET	1
ESMF	1
FABM	1

FEWCalc	1
Food Web Model	1
GAMIT/GLOBK	1
gdal	1
GeoCLAW	1
Geonet	1
GEOPANDAS	1
GEO SX	1
GETM	1
Glide	1
GMT	1
GOSPL	1
GRASS	1
Groundwater Tutor	1
Hec-Ras	1
HYCOM	1
HydroTrend	1
HYDRUS suite	1
ILAMB	1
iRIC	1
iSALE	1
ISCE	1
ISSM	1

JAGS	1
Jupyter Lab	1
K-Means	1
Keras	1
LIGGGHTS	1
Litholog	1
Lobyte 3D	1
LSD TopoTools	1
Mapinfor	1
MeanderPy	1
MIKE	1
MITGCM	1
MODFLOW	1
Multiphysics	1
National Hydrologic Model	1
Network Extraction from Bathymetry	1
Numpy	1
Pecube	1
Permafrost Modeling Toolbox	1
PETSc	1
PFLOTRAN	1
Photoscan	1
PISM	1

Pix4D	1
PyLith	1
Pymc3	1
pymt	1
PyReef	1
QTQt	1
R-Spatial	1
R-Tideverse	1
RAFEM	1
RAIDER	1
SCHISM	1
SciPy	1
Serac	1
SMC	1
SMS	1
SNAC	1
SPSS	1
SRH-2D	1
SSR	1
Stan Hamiltonian Monte Carlo System	1
Stanford NatCap	1
Striplog	1
Stromatobyte	1

Tensorflow	1
Terrainbento	1
TISC	1
TMB	1
TopoModel	1
TRAIN	1
Ttlem	1
UA	1
UCODE	1
Visjet	1
Visual Plumes	1
WASP	1
WaveWatch III	1
WBMsed	1
Weka	1
Whiteboxtools	1
OTHER	
agent-based models	2
Python repositories/packages	2

CSDMS Workbench	1
reactive transport and hydrologic codes	1
own software	1
analytical models	1
cluster analysis	1
GIS & image processing	1
hydrology	1
tsunami, wave propagation and seismological models	1
in house, proprietary and commercial	1
	256

Appendix B

What new/emerging Earth Surface Processes modeling techniques, strategies, resources are you most excited about?
Community
I really like initiatives such as the 30 day map challenge on social media. It gets programmers from all walks of life to produce, and importantly share interesting and novel techniques.
This really is not new, but getting more people to do social coding (or partner coding?) via shared, versioned, git-enabled, Jupyter notebooks.
Computing
Model interoperability
Model interoperability (such as LandLab or other packages built to work with a variety of Earth surface models) and parallelizability (whether on an HPC, GPU, or just locally)
Parallel Computing
Parallel processing
Parallelizability
Parallel global scale landscape and stratigraphic forward modelling software
Model Coupling/ Integration
Coupling between models using interfaces/apis like LandLab and the BMI.

Integration of coastal and hydrology models
Coupling discrete particle modeling and lattice Boltzmann modeling for analyzing sediment transport.
Easier coupling of models, and generally the increasing push towards modularity.
Modular systems on heterogeneous ressources
Model coupling and integration
Combining multiple models to understand systems. Replacing parts of models with real data (eg. real ET in a hydrological model).
Truely cross-disciplinary integrated modeling
Coupling physical-based models with HD/Eco/Agent-based models
Making models interdisciplinary - combining landscape evolution with human/animal interaction, historical context, or other non-geology factors
Cross-disciplinary/interdisciplinary couplings (climate plus landscapes, weathering plus erosion, ecology plus landscapes)
Integrative modeling that includes human and economic components and are relatable to the people who need to implement solutions to earth resource problems.
Integrating networks and agent based models
Earth system models: linking physical biological and other domains. And networks and topology
Coupling physical-based models with chemical models
The ability to merge physical and chemical fluctuation/evolution in models
Coupled tectonic and surface processes models
Fully coupled tectonic / surface processes models
Integration of tectonics and surface processes.
Machine Learning/AI
AI habitat modeling

Machine Learning
Modelling aided with Artificial Intelligence, Modelling aided with Artificial Intelligence, Model downscaling and data assimilation techniques
Machine learning methods for hypothesis testing and data-driven parameterization development.
Machine learning models explanation.
Parallel processing, machine learning and image analysis
Different machine learning /MCDM models used in natural hazard susceptibility/vulnerability/risk modeling, landscape evolution models, etc.
Application of AI in long term evolution of beaches and barrier islands!
Neural networks, machine learning, Bayesian methods
Mixing data (AI) and models
Big data + AI. The amount of global/large scale datasets and analyses popping up is overwhelming, and uncertainty quantification is starting to become more prominent (or perhaps, the fact that it's usually poorly done is beginning to become more painful). However, it's really exciting to work in a time where we can play with global data that respects local features (e.g. pixels don't need to be 1 km resolution anymore to do analysis thanks to improving compute and AI techniques).
Datasets/components/data assimilation
Gridded meteorology datasets, global DTM datasets
Data assimilation techniques
Data--model integration
Integrating models with global data sets like EarthEngine and upcoming NASA SWOT.
Am most excited about Digital Globe (World View) data in terms of future accessibility. Right now it is mainly cost-prohibitive to the average scientist but I envision it to be more freely available.
Julia
I'm intrigued by Julia but have yet to use it.

Julia
Scaling
Model downscaling
Methods for upscaling processes to longer time periods
Model Comparison/analysis/testing
Comparison of physical and numerical models
Model analysis, multi-model testing, hypothesis testing with models.
Ability to constrain model equations and parameter space based on calibration with natural topography - and utilization of the constrained parameters to better understand processes in (similar) natural environments and their spatial and temporal scales
Semantic Web
High Res Images and ESP Model linkages
Linkages between high-res images and ESP models, specifically those related to soil processes and impacts to ET and surface runoff
EKT
Online training
More training
I want to make sure students can run and edit codes from anywhere with minimal roadblocks to installation
Csdms labs for classroom use. Creating easy to use modules or activities to help teach computational geosciences at undergrad level. Machine learning for geoscience applications.
jupyter notebooks
Common libraries and toolkits! I teach at an undergrad-only institution and in my research and theirs, I don't have time to teach them how to troubleshoot or code properly. Having toolkits like Landlab (with more components) is so vital for my research in terms of available time and student expertise.

Science
Finite Volume Method
sediment transport/deposition models
novel parameterizations of fluvio-glacial erosion
decision-making
developing new computational methods (outside of mainstream community)
Network and graph-based approaches
Bayesian methods
agent-based modeling
Python embedded land escape evolution model
dendra.science
digital twins
Miscellaneous
Also hopeful that high speed, large bandwidth internet access will be made available to more of our citizenry. There is great untapped potential in rural and inner-city regions of the U.S. that can be utilized by the earth sciences community if we have the vision and leadership to capitalize. Those communities can greatly benefit from the tools and expertise that can be made available through virtual platforms.
Still get the most out of developing my own strat forward models and running as experimental tools - seems to me building your own experiments is still the best route to improve understanding
Science topics/themes/areas mentioned
global scale landscape and stratigraphic forward modelling

Integration of coastal and hydrology models
sediment transport
landscape evolution with human/animal interaction
climate plus landscapes, weathering plus erosion, ecology plus landscapes
human and economic components
tectonics and surface processes
habitat
natural hazard susceptibility/vulnerability/risk modeling
landscape evolution models
beaches and barrier islands
topography
soil, ET, runoff
fluvial-glacial
landscape evolution