NSF Workshop Report on a Scientific
Software Innovation Institute (S2I2)
for Environmental Observatories

NSF Workshop on Creating Scientific Software Innovation Institutes for Sustained Cyberinfrastructure Achievement and Excellence

Workshop Report and Recommendations

A National Science Foundation Workshop

October 29, 2010

Report from the NSF-funded workshop held October 4-5, 2010 in
Arlington Virginia to address a Scientific Software Innovation Institute
(S2I2) for Environmental Observatories

This material is based upon work supported by the National Science Foundation under grant number 1049273. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Workshop Organizing Committee

Stanley C. Ahalt, Ph.D., PI
Director, RENCI
Professor, UNC Computer Science Department
Chair, Coalition for Academic Scientific Computation (CASC)
alaht@renci.org
919-445-9642

Barbara Minsker, Ph.D., Co-PI
Professor and WSC Community Representative
Department of Civil and Environmental Engineering and NCSA
University of Illinois Urbana-Champaign
minsker@illinois.edu
217-265-5293

Ray Idaszak, Co-PI
Director, Collaborative Environments, RENCI
University of North Carolina at Chapel Hill
rayi@renci.org
919-445-9671

Dave Tarboton, Sc.D.
Professor and CUAHSI Community Representative
Utah Water Research Laboratory Civil and Environmental Engineering Department,
Utah State University
mailto:dtarb@usu.edu
435-797-3172

Michael Tiemann
Vice President, Open Source Affairs, Red Hat
tiemann@redhat.com
919-754-4222

James Brunt
Chief Information Officer
Long Term Ecological Research Network Office (LTER)
University of New Mexico
jbrunt@lternet.edu
505-277-2534

Robert Tawa
Director of Computing
National Ecological Observatory Network, Inc. (NEON)
rtawa@neoninc.org
720-746-4850

Peter Backlund, Ph.D.
Director of Research Relations and co-lead of cyberobservatories.net
National Center for Atmospheric Research (NCAR)
backlund@ucar.edu
303-497-1103

Mark Williams
Principal Investigator, Niwot Ridge LTER program
Co-PI, Boulder Creek Critical Zones Observatory
Chair, Integrated Data Management System for the Critical Zones Observatory program
University of Colorado at Boulder
Institute of Arctic and Alpine Research
markw@culter.colorado.edu
303-492-6387

Mark Parsons
Program Manager, National Snow and Ice Data Center
Co-PI, Cooperative Arctic Data and Information Service (CADIS) that oversees the
NSF-funded Arctic Observing Network (AON)
University of Colorado at Boulder
parsonsm@nsidc.org
303-492-6199

1.0 Overview

As part of its Cyberinfrastructure for the 21st Century Science and Engineering (CF21) vision, the National Science Foundation is exploring the creation of three to six Scientific Software Innovation Institutes (S2I2). To assist with informing the NSF about what should go into an RFP for an S2I2, the NSF issued a Dear Colleague Letter: Scientific Software Innovation Institutes – Call for Exploratory Workshop Proposals (NSF 10-050).

RENCI[1] and NCSA[2] hosted the workshop that explored a Scientific Software Innovation Institute for Environmental Observatories. The workshop was the seventh funded in a series of workshops hosted by organizations across the U.S. on a variety of subjects designed to inform the NSF S2I2 RFP process. The workshop was held October 4-5, 2010, adjacent to the NSF at the NCSA ACCESS Center in Arlington, Virginia. Representatives from nine NSF-funded environmental observatories and observatory-related cyberinfrastructure projects attended: AON[3], CUAHSI CHyMP[4], CUAHSI HIS[5], CZOs[6], HydoNexrad[7], LTER[8], NEON[9], OOI[10], and WSC[11]. These environmental observatories span five NSF directorates: Geoscience [GEO]; Engineering [ENG]; Social, Behavioral, and Economic Sciences [SBE]; Biological Sciences [BIO]; and Polar Programs [OPP]. Vendor representatives were Red Hat, Microsoft Research, IBM, Environmental Systems Research Institute (ESRI), and Danish Hydraulic Institute (DHI). Other organizations represented included the National Center for Atmospheric Research (NCAR), Unidata, California Institute for Telecommunications and Information Technology (Calit2), San Diego Supercomputer Center (SDSC), National Snow and Ice Data Center (NSIDC), Stroud Water Research Center, and several universities where cyberinfrastructure research and development for environmental science is conducted. Four sites participated remotely via Access Grid videoconferencing. Reed Beaman, Alan Blatecky, Christy Geraci, Bruce Hamilton, Nancy Huntley, Cliff Jacobs, Peter McCartney, Manish Parashar and Kevin Thompson attended from the NSF. In total there were 49 people who attended in person including nine from the NSF, and an additional four remote sites participated in the workshop via videoconferencing.

The workshop provided an open forum for the attendees to explore realization of the NSF CF21 vision and how to yield robust, sustainable products and methods that respond to the growing cyberinfrastructure needs of the environmental observatory communities. It addressed community architectures driving effective governance models for coordination of sustainable community practices, including integration of community workflows, interoperability of resources, and virtual organization principles that are expected to drive new science at the interstices of these communities. The workshop provided the opportunity for open discussion among leaders at each of the environmental observatories, cyberinfrastructure experts,

[1] Renaissance Computing Institute; http://www.renci.org/

[2] National Center for Supercomputing Applications; http://www.ncsa.illinois.edu/

[3] Arctic Observing Network; http://www.aoncadis.org/

[4] Community Hydrologic Modeling Platform; http://www.cuahsi.org/chymp.html

[5] CUAHSI Hydrologic Information System; http://his.cuahsi.org/

[6] Critical Zone Observatories; http://criticalzone.org/index.html

[7] Hydro Next-generation Radar providing radar-rainfall data for use in hydrology, hydrometeorology, and water resources; http://hydro-nexrad.net/

[8] Long-Term Ecological Research; http://www.lternet.edu/

[9] National Ecological Observatory Network; http://www.neoninc.org/

[10] Ocean Observing Initiative; http://www.oceanleadership.org/

[11] Water Sustainability and Climate Program; http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=503452

and industry experts on the topic of what constitutes a successful S2I2 institute. The workshop report herein distills the key organizational precepts and insights from the workshop and provides a summary of critical elements of an environmental observatory S2I2 that are respectfully recommended to the NSF as it crafts the S2I2 RFP and subsequent proposal evaluation process.

2.0 Workshop Summary

The workshop was awarded by the NSF on September 20, 2010, and held Monday, October 4 and Tuesday, October 5, 2010, in Arlington, Virginia, to facilitate and promote attendance by NSF personnel. The workshop organizing committee announced the workshop at http://www.renci.org/s2i2workshop and assembled a representative spectrum of attendees as described in Section 1.0. To facilitate workshop functions, a workshop wiki was set up at https://s2i2eonworkshop.renci.org/ (login = environmental or ad\environmental , password = observatories!). The workshop wiki includes all the plenary, panel, and breakout presentations described in this summary. The organization of the workshop included morning plenary discussions on Monday followed by a panel involving representatives from the nine NSF-funded environmental observatory and related cyberinfrastructure projects and programs. On Monday afternoon, six breakout sessions were held on topics corresponding one-to-one with the required elements in the NSF Dear Colleague Letter: Scientific Software Innovation Institutes (S2I2) – Call for Exploratory Workshop Proposals (NSF 10-050). Tuesday morning featured a special guest speaker from Red Hat followed by reports from each of the Monday afternoon breakout sessions. Open discussion and report planning occurred during a working lunch into early afternoon, when the workshop adjourned. Attending members of the organizing committee convened after the workshop to generate report-writing assignments, timeline, and due dates.

The workshop commenced on October 4 with PI Stan Ahalt's plenary discussion titled “Software for Science: An Institute Vision for Environmental Observatories.” The presentation provided a framework for workshop discussion, objectives and deliverables. Dr. Ahalt emphasized the NSF CF21 vision with a challenge to the attendees to identify requirements associated with how an S2I2 institute with a cross-cutting cyberinfrastructure could uniquely drive and sustain new science not otherwise possible through other NSF vehicles. Potential frameworks for governance and interoperability were addressed. A special emphasis was placed on cyberinfrastructure software for data, recognizing that there are a preponderance of data usability issues and associated software considerations across the environmental observatory communities.

Co-PI Barbara Minsker presented the next plenary presentation titled "Cross-cutting Environmental Observatory Community Challenges That Could Benefit From an S2I2: Summary of the FEON Vision." The presentation provided an historical perspective of planning efforts in recent years to coordinate, or federate, environmental observatories with a shared cyberinfrastructure. Dr. Minsker noted that the enthusiasm of the communities has always been very high, but what has been lacking is an appropriate NSF funding vehicle to elicit the required "activation energy" from the community. One example offered was a 2008 workshop titled "Cyberinfrastructure for Environmental Observation Networks (CEON)" that put forth the vision of a Federation of Environmental Observatories, or FEON. There was consensus from workshop attendees that interest remains very high in coordinating environmental observatory communities' research and education efforts through a shared cyberinfrastructure.

Cliff Jacobs, NSF Program Director in the Division of Atmospheric Sciences, presented a one-hour plenary session titled "Software Infrastructure for Sustained Innovation (SI^2)." The session served to inform attendees of the NSF Cyberinfrastructure for 21st Century (CF21) vision and how this relates to the goals of the SI^2 program, the need for Scientific Software Innovation Institutes, and questions the workshop could help answer. The presentation emphasized that software should be treated as infrastructure and that it is an integral part of the investment in CF21. Other discussion included rethinking the software challenge, elements of SI^2, and specific goals and output from the workshop in influencing the future directions of SI^2.

The Monday morning session concluded with a panel discussion titled "Individual Community Introspectives on the Role of a Cross-cutting S2I2." Representatives from each of nine NSF-funded Environmental Observatories and related cyberinfrastructure projects each presented a brief introduction to their community's efforts. Each addressed what software needs and services could be best met by a cross-cutting software institute and the corresponding challenges and opportunities. The presenters were Dave Gallaher for AON, Dave Tarboton for CUAHSI HIS, Rick Hooper for CUAHSI CHyMP, Anthony Aufdenkampe for CZO, Mohan Rammamurthy for Unidata and HydroNexrad, James Brunt for LTER, Bob Tawa for NEON, John Orcutt and Matt Arrott for OOI, and Barbara Minsker for WSC. Co-PI Ray Idaszak, master of ceremonies for the workshop, served as panel moderator. Emphasis was placed on defining requirements and services rather than specific solutions. At the conclusion of the nine presentations, the panel assembled to answer questions from the other workshop attendees.

Monday afternoon was dedicated entirely to breakout sessions addressing each of the six sets of example issues included in the NSF Dear Colleague Letter. The breakout sessions were held three at a time concurrently in three rooms with the first three sets of issues addressed in the early afternoon and the second three sets of issues addressed in the late afternoon. Breakout sessions attracted between nine and seventeen attendees, including one moderator and one note taker. The workshop then adjourned until Tuesday with the moderators and note takers working after hours to consolidate the raw notes into slides to be presented Tuesday morning. The workshop recommendations to the NSF from the breakout sessions are documented in Section 3.0 herein.

Tuesday morning highlighted the workshop's invited special guest speaker: Gunnar Hellekson, Chief Technology Strategist for Red Hat. In his professional role, Mr. Hellekson assists Red Hat's U.S. Public Sector group, where he works with systems integrators and government agencies to encourage the use of open source software in government. Additionally his background includes being a co-chair of Open Source for America and a select member of the Commission on Government Technology Opportunity in the 21st Century (GTO-21). His background uniquely qualified him to present an in-depth view of the open source model as successfully applied to the public sector. Mr. Hellekson provided effective examples demonstrating how open source mechanics have been successfully used in efforts beyond software and that embody an entire community of practice. In the Q&A session following his talk, Mr. Hellekson discussed how open source mechanics might be applied to operationalizing a cross-cutting cyberinfrastructure among environmental observatories, emphasizing the view of treating software as a process, rather than a product, in this endeavor.

The workshop continued Tuesday into early afternoon with presentations by a designee from each of the six breakout sessions. Each presentation summarized key findings and each presentation was followed by an all-hands discussion and refinement of the findings. These findings, combined with observations from all of the workshop discussions, were used to derive the recommendations given in Section 3.0. The workshop adjourned Tuesday afternoon with an all-hands wrap up that discussed where the group had reached consensus and where it had not, followed by a closed session among members of the workshop organizing committee to generate writing assignments, a timeline, and due dates.

An initial draft of the workshop report was created by October 22, 2010, announced, and made available to workshop attendees on the workshop wiki for open review and comment. The open review period lasted through October 27, with daily new versions posted to the wiki integrating community comments. The final draft was then reviewed for typographical errors and grammar, and respectfully submitted to the NSF.

3.0 Workshop Key Organizational Precepts, Conclusions, and Recommendations

This section discusses the key organizational precepts, conclusions, and recommendations from the workshop and summarizes critical elements of environmental observatory software and cyberinfrastructure that should be part of an NSF Scientific Software Innovation Institutes (S2I2) RFP. Sections 3.1 to 3.6 correspond to the six areas described in the NSF Dear Colleague Letter: Scientific Software Innovation Institutes (S2I2) – Call for Exploratory Workshop Proposals (NSF 10-050). The subject matter for each section was derived from the entire workshop, including plenary and panel presentations, breakout sessions, and open discussions. ***Explicit recommendations to inform the NSF S2I2 RFP and subsequent proposal review process are indicated in BOLD ITALICS.***

3.1 How will significant challenges faced by the environmental observatory community benefit from an S2I2 in terms of scientific innovation as well as productivity?

- **Software Ecosystem for New Science –** A primary tenet of the workshop was that data, software for data, and application software are inextricably linked. Many common issues faced by NSF-sponsored observatories in offering data and software services in support of grand challenge science would benefit from common software research and development coordinated by an institute. An S2I2 for environmental observatories would provide a critically necessary software ecosystem that enhances and accelerates new data-driven science within communities while enabling new science between communities. An S2I2 will facilitate sharing of data elements via associated interoperability of software for data. There was an assertion that significant new grand challenge science and innovation will happen at the interfaces or interstices between communities. The need for an environmental S2I2 was emphasized as a vital necessity to marshal the software assets enabled by NSF funding and that it be sustained for future scientific explorations not otherwise possible.

- **Synthesis –** An environmental observatory S2I2 would serve as a synthesis center enabling multi-disciplinary collaboration on critical grand challenge science such as mitigation and adaptation to climate change, global hypoxia, and ecosystem and water sustainability. In addition, support for cross-disciplinary working groups to collaborate in manipulating heterogeneous data from multiple sources into new scientific methods and knowledge is critical for progress on these challenges. That is, while *data is the idiom of environmental grand challenge science, software is its modality*. ***It is critically important that the NSF recognize the importance of an S2I2 serving the role of a true cross-cutting multi-functional institute in expanding synthesis within and across communities through data and software interoperability, support for synthesis working groups, and sharing of best practices.***

- **Software for Data –** Each of the environmental observatories recognized inherent data challenges within their respective communities. The potential benefits of an S2I2 to the environmental observatory communities were largely oriented around data services, including improved data handling and standards for interoperability, ease of use, and collaboration through data. A particular need was discussed for an infrastructure to enable community members to easily create and share custom data products that integrate diverse data types at multiple scales and are more easily used for analysis and modeling, and estimating the state of environmental systems across space and time (including error bounds). A recent memorandum from the Executive Office of the President regarding U.S. Science and Technology Priorities for the FY2012 Budget reinforces the priority of sustainability and interoperability of data:

> "Agencies...should develop and sustain their datasets to better document Federal science, technology, and innovation investments and to make these data open to the public in accessible, useful formats. Agencies should develop and regularly update their sharing policies for research performers and create incentives for sharing data publicly in interoperable formats to ensure maximum value..."[12]

There was discussion on the need for an S2I2 to provide software as infrastructure to facilitate these requirements both within and between communities, and on the importance of clearly differentiating the role of an S2I2 versus the roles of other NSF data-centric programs such as DataNet. ***An environmental observatory S2I2 should connect to DataNets and other data-generating initiatives and provide and sustain a software ecosystem that promotes an interoperable suite of customizable data products that are closely coupled with corresponding analytic and synthesis activities.***

- **Productivity and the Transformation of Science –** Productivity improvements and the transformation of science would result from more data of higher quality and greater usability from more sources. Currently, finding and transforming available environmental data for use in scientific studies and modeling can require months of effort for every research project, while readily available and reliable data products from an environmental observatory S2I2 could reduce this time to minutes. The need for increased productivity and enabling research in transformative science is a definite prerogative of an S2I2.

3.2 What are potential key attributes of an S2I2 that would benefit the environmental observatory communities? What are the appropriate S2I2 organizational, personnel and management structures, as well as operational processes for this community?

- **Interoperability –** Interoperability was especially significant in that some workshop attendees described interoperability *as the grand challenge* including how to bridge across the various communities. There was considerable discussion, but not consensus, on the approach. Some proposed a full-blown institute providing SaaS (Software as a Service) relying on resource providers where environmental observatory services could run primarily on academic or commercial resources not owned by the community, and where the institute would coordinate data management. Others wanted a "lighter" solution with a focus on resource coupling, resource pooling and standards. The recommendation that distilled out of this discussion is that there are a variety of possible approaches to implementing an S2I2, whether for environmental observatories or another focus. ***The NSF should view the S2I2 RFP as an opportunity to enable and***

[12] Orszag, Peter R., Holdren, John P. Science and Technology Priorities for the FY 2012 Budget. Memorandum For the Heads of Executive Departments and Agencies, Executive Office of the President, Washington D.C., July 21, 2010, pg. 2. Retrieved October 14, 2010 from http://www.whitehouse.gov/sites/default/files/omb/assets/memoranda_2010/m10-30.pdf

encourage the community of responders to be creative and persuasive in their responses. To that end, the S2I2 RFP language should not be overly prescriptive. The S2I2 RFP should require responders to assert clearly the community/science needs and cyberinfrastructure grand challenges they are addressing and provide reasonable validation that the need actually exists broadly among the community(ies).

- **Software Sustainability –** ***A clearly identified issue with sustainable software is that support for grant-developed software is typically nonexistent at the end of a grant. This is a key issue for the S2I2 to address. S2I2 responders should clearly state their method for encouraging and supporting grant awardees in transitioning successful software into sustainable operational infrastructure, and if possible, provide substantive examples demonstrating that the proposed method has been successfully utilized in the past in achieving sustainability.***

- **Support for Entire Software Lifecycle –** To enable sustainability after a project, an S2I2 should work with developers during a project. Example activities could include promoting best practices, providing test harnesses, standardizing documentation guidelines, and delivering software engineering education. ***S2I2s should support the entire lifecycle of software in supporting production science and indicate how S2I2s will accomplish the support and transition the software throughout the software lifecycle.***

- **Support for Entire Project Lifecycle –** ***Proposal responders should be required to very clearly walk an example project through their proposed institute from innovation to sunset, including metrics and education elements.*** To assist S2I2 reviewers, one idea is for proposers to document an example project within their S2I2 as a trial balloon moving through the S2I2 over time to elicit the attributes of the mechanics and flow of the proposed S2I2. Exactly how is sustainability achieved? Outline who is working on code and projects over time and differentiate between S2I2 staff and non-S2I2 members of the community. In describing their S2I2, respondents should elaborate their use of popular or common approaches. For example, many workshop attendees gleaned from Red Hat's talk that success in *open source* encompasses much more than choosing the right license and making the source code openly available under that license. ***Proposal responders should be required to state explicitly how incorporation of a proposed concept, such as open source, whether original or borrowed from industry, will lead to sustainability, innovation, and new science.***

- **Ratings –** There was much discussion about how an S2I2 would provide some sort of rating or assessment of aspects of software, projects, developers, and possibly even whole communities. Each rating scenario discussed included levels for the entity to graduate through as a function of capability and maturity. Various analogies were made to current assessment models. For example a *Software Readiness Level* was proposed that might be based on the *Technology Readiness Level* used by government agencies. The Carnegie Mellon University *Capability Maturity Model* was discussed to assess software project processes. To assess whole communities, the idea of a *Cyberinfrastructure Maturity Model* was discussed as loosely based on the *Open Group Service Integration Maturity Model*. The Apache Software Foundation model was discussed. In each example, the benefit was to level set entities of a similar nature in a common assessment framework to assess capability and maturity and thus inform the communities according to stated goals and objectives. ***The consensus was that the S2I2 should continually provide substantive assessment approaches of software, projects, developers, and communities and continually make this information readily available.***

- **Standards –** Workshop attendees discussed that standards, while providing benefit, are currently overwhelming because of the number of extant standards, and the challenges in terms of identifying which of hundreds to use, how to use them effectively, and why. ***An S2I2 should provide a substantial benefit not only in terms of collaborating with existing standards-setting organizations to create new standards, but in cataloging and identifying which extant standards to use, and providing associated software promoting interoperability across communities.*** The analogy given was for an S2I2 to serve as a version of an "Underwriters Laboratory" for vetting standards.

- **Governance Representation –** An environmental S2I2 should implement a governance structure that is representative of the community across multiple dimensions including, for example, observatory type, codes, nature of research processes, education, analytics, and information technology. There should be diverse, multi-sector participation with comprehensive expertise in the represented areas. ***The S2I2 governance should be driven by and responsive to the community that the software is intended to serve.***

- **Community Architecture –** The consensus was that an ideal S2I2 approach would not be an "*uber*-institute" with a one-size-fits-all technical solution, but rather an effective distributed network of shared interoperable, extensible, and modular infrastructure; policy; governance; and community and virtual organization best practices. ***An S2I2 should implement a community architecture which clearly states the processes and policies by which:***
    - ***user requests are accommodated,***
    - ***priorities are determined,***
    - ***development and/or development direction is provided,***
    - ***community direction is assimilated,***
    - ***codes are matured,***
    - ***certifications are granted,***
    - ***governance is established and evolved,***
    - ***various committee members are appointed or elected,***
    - ***special cross-cutting projects, workgroups and task forces are instantiated,***
    - ***effective incentive and reward structures are continually refined,***
    - ***self-sustainment is achieved, and***
    - ***other processes and policies are developed and implemented as needed.***

  Workshop attendees asserted that it is important to involve end users in most if not all of these activities.

- **Institute versus Large Project –** Given that the S2I2 is likely to be created as an institute, there was discussion regarding ***what differentiates an institute from a large NSF-funded project. The consensus was that the S2I2 RFP should require responders to explicitly make this distinction and clearly identify how their proposed effort will function as an institute. Furthermore, it should answer the question of how only an institute can uniquely service the proposed effort versus funding it as a large project.***

- **Software Infrastructure –** ***There was consensus that an S2I2, as an institute, should preclude serving the sole role as a primary developer of major software.*** There were examples discussed during the workshop that suggested this approach falls short of the intended vision and scope of an S2I2. ***The recommendation is for the S2I2 RFP to preclude the building of monolithic***

"one-size-fits-all" software, but instead to allow the creation of software infrastructure efforts like Software as a Service (SaaS) environments in response to community requirements as appropriate. The aim is to emphasize the sustainability and interoperability of software and not the building of new, large monolithic software.

- **Expanded Computation and Analysis –** Much environmental observatory computation and analyses are confined to desktop PCs or small workgroup servers. An ongoing significant challenge to the environmental observatories is that they often lack some or all of the required staff, knowledge, tools, time, infrastructure, or institutional support to migrate to modern HPC architectures, grid, and/or cloud environments. Correspondingly, environmental observatories are not taking full advantage of many larger more sophisticated visualization and analysis tools associated with modern large-scale distributed computing. ***An environmental observatory S2I2 will be integral to realizing the linking of environmental model and data products to emerging computational and analysis environments. An S2I2 should assist with transitioning software from desktops to modern HPC architectures, grid and cloud environments.***

- **Permanence –** Another issue that was discussed was institutional permanence, noting that environmental observatories as instantiated last longer than 10 years. Accordingly, there was a discussion on how to attract and retain graduate students to participate in the S2I2 for the long-term when only five years of S2I2 funding were secure. An idea put forth was for S2I2s to employ "programming fellows" (described also in Section 3.3) whose role would be to support various aspects of code sustainability beyond the more transient nature of grad students. ***The longer-term "NCAR" sustained funding model was discussed as a possible model for an S2I2, and the workshop recommendation was that the NSF either needs to do long-term support for S2I2s, or require a plan from each for how they will transition away from funding after 10 years. Given the NSF's currently stated intent of the 5+5 year approach, the S2I2 RFP should preclude responders requiring long-term funding beyond 5+5 years to be successful and require a clear sustainability plan.***

3.3 What expertise and capabilities should an S2I2 provide and how should it interface and interact with science communities? What education and outreach functionalities are meaningful in an S2I2?

- **An Intellectual Hub for Synthesis –** One mechanism that the workshop attendees identified for how an environmental observatories S2I2 should interface and interact with science communities was serving as an intellectual hub for synthesis. As an example, organizations like the National Evolutionary Synthesis Center host synthesis activities; one such activity described as "an intense event at which a group of programmers with different backgrounds and skills collaborate hands-on and face-to-face to develop working code that is of utility to the community as a whole. The mix of people...include domain experts and computer-savvy end-users."[13] During synthesis workshops, workshop attendees would promote intense, hands-on multi-disciplinary coding and/or operational sessions that endeavor to push the boundaries of broader data and resource incorporation and analysis towards identifying the elements of transformative science and research. ***An S2I2 RFP should require responders to state specifically how they will enable transformative new science.***

[13] National Evolutionary Synthesis Center Hackathon CFP; Retrieved October 14, 2010 from http://gmod.org/wiki/GMOD_Evo_Hackathon_Open_Call

- **Education and Outreach –** ***Workshop attendees agreed that an S2I2 needs to play a strong role in education and outreach as it engages with science communities to train computationally competent scientists and engineers. Education should target students, the PIs and managers of environmental observatories, and all other end-users. Outreach should extend beyond the environmental observatory communities via linking with environmental practice and other research and data provider communities, including other Federal agencies. Outreach should promote software that is modular, customizable, and extensible to enable interoperability. Software reuse and repurposing of applications and methods should be taught. S2I2 outreach should also facilitate software adoption among communities and must have end users involved in software development from the onset.***

- **Expanded Communication –** ***An S2I2 should facilitate expanded communication among and between the user communities, particularly including the social sciences. Communication should also be facilitated between domain and developer communities, including domain and computer science terminology and concepts.*** This could include a common vocabulary for education and outreach supporting data storage, data ingestion, data visualization, and so forth and also concepts like domain-specific languages, frameworks, and language-oriented programming. The idea of a "distinguished fellows" program was put forth where scientists in residence would facilitate inter-community communication and collaboration to ensure effective coupling and flow of information across communities.

- **Open Source Mechanics –** ***On the development side, open source should be explicitly supported and promoted.*** Models can be borrowed from successful open source examples including Apache Software Foundation and Microsoft Outercurve. ***An S2I2 should provide expertise to ensure long-term sustainability of the software efforts it facilitates, including perhaps some type of certification that qualifies software for interoperability and sustainability. An S2I2 should provide support for the creation of software foundations that build upon successful open source practices.***

- **Programming Fellows –** A concern that was expressed was how to get around the issue of graduate students leaving before software is fully developed from the S2I2 perspective of interoperability and sustainability. One potential solution was for the S2I2 to create "programming fellows" who support community codes long-term with sustainable practices including open source and good software engineering. Such software might be transitioned from “program to product” if the S2I2 supported the original developers of software with promise to work directly with programming fellows in residence at the S2I2.

- **Student Support –** ***An S2I2 must support students.*** One idea suggested was to bridge gaps by looking at where students could eventually be employed upon graduating and then facilitate the positioning of students for these destinations while supporting S2I2 community codes. Target destinations for students could include EPA, USGS, NOAA and NASA, for example. The S2I2 could offer internships for students including opportunities for domain students to learn computer science, computational/cyber-science and engineering, and software engineering topics. The S2I2 could create computational curriculum modules and implement a train-the-trainers model.

- **Boot camps –** ***There was consensus among workshop attendees that an S2I2 should provide a "boot camp" to educate and train end-users in all facets of the environmental observatory***

cyberinfrastructure ecosystem. Topic areas could be domain oriented, computer science oriented, or other. For example, there is a large gap between professional developers and academics. How does an S2I2 educate the community to transform the scientific software development community? Boot camps would be one approach. The "boot camp" connotation is meant to evoke that it is low cost, group-oriented, intense but reasonable to traverse for a diverse audience, and of sufficient duration to address a variety of issues and achieve a broad range of objectives.

- **Computational Science Education –** A critically significant educational challenge in computational science was emphasized at the workshop. Computer scientists in the U.S. are not being adequately educated in emerging architectures such as heterogeneous many core systems. Further, domain scientists are not being educated in good programming practices, software and data lifecycles, computational data analysis techniques such as data mining, and sustainability methods such as key aspects of software engineering best practices. Computer and domain scientists are not being adequately educated in team approaches to computational science involving computer scientists, domain scientists, and numerical analysts responsible for best algorithm selection. The problem is exacerbated by the absence of these effective and knowledgeable computationally-oriented teams to address refactoring existing scientific code or creating new scientific code so that scientific codes can be readily updated and sustained, new capabilities can be easily added, new platforms can be supported, results can be repeated, new communities of developers can be added, and elements can be reused over long periods of time. A recent article in the October 2010 issue of Nature[14] reinforces the fact that most science code - while functional and often of high value - is not written according to computer science best practices and therefore is not sustainable, reusable, and often is not repeatable. ***Resolving the computational science education problem needs to go beyond training current computer and domain scientists and reach all the way into the undergraduate and graduate classrooms. Given the enormity of the problem, the recommendation is that the S2I2s should begin to address this issue in order for the U.S. to maintain international computational science leadership.***

3.4 What are the critical linkages between an S2I2 and other components of a community cyberinfrastructure (i.e., software tools, databases, instruments, etc.)? What is the unique role of an S2I2 in the broader cyberinfrastructure ecosystem (e.g., TeraGrid/XD, DataNet, MREFC, etc.)?

- **Linkages and Coordination –** Currently there is no overall coordination among the environmental observatories. While some of the environmental observatories individually interact with other NSF efforts like TeraGrid, DataNet, MREFCs, and others, it follows that there is no overall coordination between environmental observatories and these projects. Additionally, there is no overall coordination between environmental observatories and other federal agencies such as EPA, USGS, NOAA, USDA, NASA, or Army Corps of Engineers, although some individual projects (e.g., CUAHSI Hydrologic Information System) have partnerships with agencies. The coordination that is desired is to promote greater software and data interoperability in this broader cyberinfrastructure ecosystem. Interoperability is a grand challenge for the

[14] Merali, Zeeya. Computational science: ...Error ...why scientific programming does not compute. NATURE, October 14, 2010, Vol. 467, pgs. 775 - 777. Retrieved October 14, 2010 from http://www.nature.com/news/2010/101013/full/467775a.html

environmental observatories. Subsumed in this grand challenge are the management of current standards; development of data-driven tools and infrastructure exploiting standards and high-performance computing resources; promotion and adoption of good software engineering principles; addressing all aspects of the software and data lifecycle; effective governance, community architecture, and open source models; and effective education, training, outreach, and workforce development. Given that applications in the environmental observatories read and write data, enabling new science is a direct function of the usability of this data and ergo the software that works with this data. Environmental observatory S2I2 critical linkages will rely on collaboration through data, and this implies recognizing the relationship between sustainability of data collections (including real-time sensor feeds) and sustainability of the software tools that work with these. ***An S2I2 RFP should require proposers to show how they will sustain software in a highly dynamic cyber-ecosystem.***

- **Value of Industry Partnerships –** Several vendors were represented at the workshop including Red Hat, Microsoft Research, IBM, ESRI, and DHI. It was clear to attendees that industry partners are essential to the success of an S2I2 and should be required in an S2I2 RFP. Red Hat's Chief Technology Strategist provided a 1.5-hour presentation and discussion on Open Source Mechanics that informed attendees of many facets and dimensions of open source that attendees had not known. Microsoft Research discussed an environmental informatics initiative and also an educational initiative with project attributes that would be difficult for a community to replicate on its own and virtually impossible to replicate and sustain across communities. Another angle on industry partnerships was the idea of an S2I2 utilizing professional commercial software developers and/or development firms in coordinating and servicing certain community development requirements. An S2I2 can uniquely and critically serve the role of effectively assimilating and disseminating the value of industry partnerships across communities. ***An S2I2 RFP should require industry partnerships and clearly describe how the value provided by these partnerships will be assimilated and disseminated to across communities. Furthermore, letters of industry participation should be required over letters of support.***

- **SSE, SSI, S2I2 Interaction –** ***The current SI2 RFP for SSE and SSI efforts stipulate that there should be interaction between SSEs, SSIs, and S2I2s. The workshop attendees agreed with this objective, with the caveat that this occur only if there is a mutually beneficial relationship and that the S2I2 also be required to interact with other existing cyberinfrastructure projects pertinent to their community, of which there are many in the environmental observatory space outside of the SSE and SSI programs.*** If the SSE or SSI focus or activity doesn't fit with a given S2I2, it should not be required to interoperate with that particular S2I2. In this way the set of S2I2 institutes may provide coverage for all SSEs and SSIs without any one S2I2 being required to provide unilateral SSE or SSI coverage. There was also discussion of expected interaction among S2I2s. For example, if there is an environmental observatory S2I2 and a security-themed S2I2, then the environmental S2I2 should leverage the security offerings of the security S2I2.

- **Multiple S2I2s –** Another comment was the endorsement of the NSF supporting multiple S2I2s versus a single S2I2. The challenge of software infrastructure and sustainability is too complex to bank on just one S2I2 institute and approach being correct or adequate. NSF should employ multiple S2I2 institutes servicing multiple disparate community dimensions.

- **Education, Training, and Workforce Development –** ***As elaborated in Section 3.3, education, training, and workforce development to support and sustain internationally competitive***

computational science in the U.S. is an enormous challenge and a critical need for the environmental observatory communities. From the perspective of critical linkages, the recommendation is that if the S2I2 effort is not the NSF vehicle responsible for addressing this, then it is paramount that the S2I2s link to the NSF efforts that are responsible for addressing workforce development in a very significant way.

- **Providing Catalogs –** ***There was strong consensus among workshop attendees that an S2I2 institute serving environmental observatories should perform the role of cataloging various software and data elements. This should include discovery, a measure of code and data quality, who is using it and how (e.g. usage models and user groups), and who is sustaining it and how. The cataloging of software and data should facilitate discoverability and interoperability across communities. What is the QA process for software and data, how is it characterized, and how will these functions be sustained over time? The cataloging function does not imply that the S2I2 itself should necessarily host the software or data, but rather that it serves the role of cataloging what the communities may themselves host. The catalog should offer users the assurance that the cataloged elements are supported and will continue to be there over time. S2I2 proposers should elaborate their plan here. S2I2s should also catalog people, that is, identify the location of experts on a particular topic or technology. The S2I2 should promote good software engineering practices and promote software for data interoperability as a function of technology, policy, and community practice, as stated earlier.*** There was an interesting idea that the NSF could use the S2I2 institutes to catalog and promote discoverability of all NSF projects and programs beyond what is provided by the current NSF website. For example, a user search across *all* NSF projects and programs via an S2I2 could yield more tangible assets (data, software, models, publications, etc.) for the purposes of incorporating and/or interoperating with that project or program's technology as appropriate. This cataloging, discoverability, and interoperability theme of an S2I2 can span many areas and communities in a high-value manner not provided for by current NSF efforts.

- **Service and Resource Provider –** An S2I2 could serve the role as a provider of storage for catalogs, cloud computing services provider, build-and-test environment provider, code hardening services provider, provider of services to make data more robust, and so forth. Some combination of these services and resources may be relevant to all S2I2s. ***An S2I2 RFP should state clearly what services and resources it would make available to communities.***

3.5 What are meaningful metrics, evaluation mechanisms and governance structures for an S2I2? What are appropriate approaches to sustainability of the S2I2?

- **Adoption and Effectiveness –** ***A primary metric identified at the workshop was adoption. Here adoption could refer to adoption of codes; best practices in software engineering; open source models; outreach, education, and training models; incentives and rewards; and standards. The adoption should be within and between communities, and could also include industry adoption. Industry adoption could include the aforementioned aspects along with intellectual property considerations, commercialization, patents, and spin-offs. Adoption should be measured as a dimension of effectiveness. There was discussion that new metrics should be able to evolve with the S2I2 based on experience and success, and that there should be a plan for adapting metrics to a changing environment.***

- **Impact –** Tied to adoption is the assessing of impact. Achieving interoperability is a key facet of an environmental observatory S2I2 on multiple levels; therefore evaluating interoperability is tantamount to success. Many of the metrics discussed were geared towards critical S2I2 outcomes such as achieving new science. An S2I2 should be able to measure how it uniquely enables the solving of more complex and/or more significant research problems than possible without an S2I2. ***The measurement of new science resulting from the S2I2 is critical, for example number of published papers as a result of having expanded cross-cutting access to the multi-community resources that an S2I2 enables. An important evaluation metric is to acknowledge, measure, and assess the impact of the S2I2's role in enabling its partners and users to produce quality software.***

- **Evaluation –** Evaluation mechanisms should clearly measure against the metrics and ultimately distill out from multiple vantage points what is working within an S2I2 cyberinfrastructure. For example, ***how does an S2I2 evaluate that it is choosing the right projects and activities to benefit its users while achieving its sustainability and transformative new science objectives?*** Third party firms could be employed in evaluation. ***Involving end-users in the evaluation process is also critically important. With respect to governance, metrics should be designed to allow the course of the S2I2 to self-correct as it is measured and assessed. Education and training metrics should enable evaluation of both quantity and effectiveness.***

- **Community Ownership –** An S2I2 will provide many high-value critical capabilities, functions and services to communities, however it is also important for the communities to express some ownership of the S2I2. How are the communities demonstrating that it is important for the S2I2 to continue its role and to continue to sustain? What are communities contributing back to assure an S2I2's sustenance? ***An S2I2 RFP should require responders to describe mechanisms and metrics of community ownership of an S2I2 that contribute to its sustainability.***

- **Governance Structure Effectiveness –** ***Proposers should outline their governance structure and carefully explain how the governance will be effective for addressing the community needs they are serving, ensuring that the software ecosystem is requirements driven rather than technology driven. The governance structure needs to allow for models where representatives from the various environmental observatories can be included in the institute, with technical personnel serving on the development team, as appropriate, and end users determining the overall direction and requirements that the software ecosystem will address. It should describe how governance enables the proposed software ecosystem to be extensible and responsive to the community. For environmental observatories, the community subsumes both the environmental observatories' researchers and staff and their many and varied users. The governance plan should indicate how end users will be included, and it should discuss how prioritization will occur when workload exceeds available resources.***

- **Opportunity and Innovation –** When a center or institute is in the position of providing resources and/or services, end-users are expected to take reasonable advantage of these offerings and correspondingly advance their research to new levels of achievement. Certain researchers are able to build on early successes, reapply for more services based on early successes, and see their successes snowball if they can continue building on prior individual or cumulative successes. On the one hand, this is beneficial for the researcher and the science represented, possibly benefitting the entire community. On the other hand, a new researcher not in a position to take advantage of early access to center or institute resources and thereby build on early successes may be less able to compete for support of a new program when there are finite resources to allocate, yet may have superior potential with a novel approach. How can an S2I2 make certain and be able to measure that it is always giving opportunity to new higher-risk researchers who are possible engines of innovation? ***An S2I2 RFP should require responders to address how they will offer new higher-risk researchers the opportunity to partake of an S2I2's offerings in balance with supporting more accomplished, but lower-risk researchers. An S2I2 should emphasize serving a wide gamut of multidisciplinary users and not just a small niche of users and types.***

- **Approaches to Sustainability –** Taking into consideration the tremendous transformative value an environmental observatories S2I2 will provide its community as described herein, it is reasonable to conclude that the value will not diminish over time. Specifically, the S2I2 can be thought of as an engine of continual transformative new science where the cylinders are coordinated interoperability enabled by the S2I2 across communities, the success and efficiency of the design is based on open source mechanics, and the air/fuel mixture is a relative never-ending abundance of data combined with finite funding. The proposed software ecosystem combined with effective governance and community architecture needs to achieve sustainable fuel. The need for this engine has been commensurate with the existence of environmental observatories since their instantiation, and the absence of this engine has been the heretofore absence of continual, community cross-cutting, transformative new science. Community members, as owners and drivers, need to own and direct the sustenance of the S2I2 engine. The engine metaphor is intended to exemplify how all the various parts - including owners, drivers, and the NSF - must work together to achieve sustainability that drives continual transformative new science, and that is key to any approach to sustainability.

- **Expanding the Scope of Sustainability –** S2I2 proposers should state if their institute will contribute to the sustainability of other NSF initiatives. For example, will the proposed S2I2 institute contribute to the sustainability of TeraGrid, OOI, DataNet, etc., or perhaps even go beyond the NSF to include projects or programs within USGS, USDA, Army Corps of Engineers, EPA, and other organizations.

3.6 How would an S2I2 impact the science and engineering in the environmental observatory community and its practices, capabilities and productivity?

There was consensus among workshop attendees that an S2I2 for environmental observatories would greatly benefit their science and engineering and associated practices, capabilities and productivity by creating a cross-cutting cyberinfrastructure that enables interoperability, collaboration, and knowledge sharing at multiple levels.

- **Increased Value and Usability –** A key outcome is that the value and usability of existing data and software assets would be increased and consequently science would be transformed. An S2I2 for the environmental observatory communities was determined to be critical to enabling new science and engineering that could only be accomplished by a scientific software institute and not via traditional large project-oriented grants.

- **Coordination –** There currently is no effort to coordinate among or integrate environmental observatories. No one environmental observatory can "own" this role as the practices and capabilities of any one are not repurposable enough to encompass the others, and new science is not being realized for this reason. An environmental observatory S2I2 would enhance and accelerate new science within these communities through access to, and interoperability with, other environmental observatory assets. Furthermore new science would also be generated between communities that would otherwise not be possible. New science between communities would result from the combination of interoperability of resources at multiple levels combined with coordinated collaborative activities like synthesis workshops and working groups.

- **Community Impact –** Part of understanding the impact of the S2I2 requires understanding who the customers are and assessing what the impact of a successful S2I2 would be to them. The workshop attendees used this model to assess S2I2 impact on producers and consumers of environmental observatory resources as customers. Examples of impact included providing: use cases and user outreach guidance; software and data that will be put in front of many users in order to increase its usability; innovation; help in generalizing software and making it more robust; help in creating software for multi-observatory services; and incentives for researchers to contribute data and software.

- **Transformation of Science –** Part of an S2I2-enabled transformation of science would result from more scientific software - of higher quality and greater usability - from more community sources with more data interoperability tools and requiring fewer transformations.

- **New Levels of Achievement –** Whether individually or as a community, an S2I2 should enable new levels of achievement not otherwise possible. For example, a National Water Model in part requires standards, effective cataloging of computational models and data, effective use of standards, software engineering best practices to allow models to be discoverable in a cataloging system and interoperable, the coordination of data and analysis across many communities, and more. In effect, achievement requires most of the offerings and benefits described in this document. In the National Water Model example, both individual researchers and the community as a whole will reach new levels of achievement, and many more achievements would be made possible by an S2I2.

- **Strategic Use of Standards –** There would be more strategic use of existing standards across multiple communities, providing incentives for greater industry and other Federal agency participation.

- **Sustainability through Open Source –** A more informed use of the open source model would enhance development and sustainability of efforts that transcend students, postdocs, faculty and staff of environmental observatories.

- **Improved Quality –** An S2I2 for the environmental observatory community will introduce software engineering best practices, assessment frameworks, refined open source models, strategic use of standards, synthesis workshops, superior analysis and interoperability tools, and more to its community. Quality will be improved across several community dimensions.

Appendix A
Workshop Participants

Stan	Ahalt	RENCI, U of North Carolina at Chapel Hill
Dan	Ames	Idaho State U
Amy	Apon	U of Arkansas
Matt	Arrott	UCSD, Calit2
Anthony	Aufdenkampe	Stroud Water Research Center
Peter	Backlund	NCAR
James	Brunt	U of New Mexico
Randy	Butler	NCSA, U of Illinois
Scott	Collins	U of New Mexico
Judy	Cushing	Evergreen State College
Charlie	Dow	Stroud Water Research Center
Dave	Gallaher	NSIDC, U of Colorado
Ken	Galluppi	RENCI, U of North Carolina at Chapel Hill
Corinna	Gries	U of Wisconsin
Gunnar	Hellekson	Red Hat
Rick	Hooper	CUAHSI
Jeff	Horsburgh	Utah State U
Ray	Idaszak	RENCI, U of North Carolina at Chapel Hill
Marcelo	Lago	DHI
Kerstin	Lehnert	Columbia U
Yong	Liu	NCSA, U of Illinois
Alex	Mahalov	Arizona State U
John	McGee	RENCI, U of North Carolina at Chapel Hill
Mike	McGuire	U of Maryland
Anna	Michalak	U of Michigan
Barbara	Minsker	U of Illinois, NCSA
Reagan	Moore	U of North Carolina at Chapel Hill, RENCI
Larry	Murdoch	Clemson U
Marian	Muste	U of Iowa
John	Orcutt	Scripps Institution of Oceanography, UCSD
Mark	Parsons	NSIDC, U of Colorado
Beth	Plale	Indiana U
Arcot	Rajasakar	U of North Carolina at Chapel Hill, RENCI
Mohan	Ramamurthy	Unidata, UCAR
Charles	Schmitt	RENCI, U of North Carolina at Chapel Hill
Scott	Sinno	NASA NCCS
Sheila	Steffenson	ESRI
Allison	Steiner	U of Michigan
Dave	Tarboton	Utah State U
Robert	Tawa	NEON
Claire	Welty	U of Maryland
Mark	Williams	U of Colorado
Yan	Xu	Microsoft
Jeff	Zais	IBM
Ilya	Zaslavsky	SDSC, UCSD

Appendix B
NSF Workshop Participants

Reed	Beaman	NSF
Alan	Blatecky	NSF
Christy	Geraci	NSF
Bruce	Hamilton	NSF
Nancy	Huntly	NSF
Cliff	Jacobs	NSF
Peter	McCartney	NSF
Manish	Parashar	NSF
Kevin	Thompson	NSF

Acknowledgements

This workshop was funded by the National Science Foundation through grant number 1049273. The Organizing Committee wishes to express deep appreciation to the National Science Foundation for their support of this workshop. The committee also especially thanks all the attendees for actively participating in the workshop discussions and panel sessions, leading breakout sessions, taking notes, and contributing to this workshop report. The interest in pursuing a coordinated effort among all NSF-funded environmental observatory communities and related cyberinfrastructure projects is now stronger than ever as a result of this workshop. The consensus built, insights gained, and recommendations provided will be of great value to the NSF, our respective communities, relevant industry partners, and other government agencies.