LEAP

LEAP RESPONSE TO OVERNIGHT QUESTIONS FROM NSF SITE VISIT TEAM

Wednesday, September 23, 2020

We appreciate the NSF and Site Visit Team's time and thoughtful engagement with LEAP's proposed design and execution. Thank you for the opportunity to continue this discussion.

How Would a Budget Reduction Of 30% in Each Of The First Two Years Of Center's Life Impact Your Plans? Specifically, Which Activities Would You Keep And Which Would You Postpone/ Reduce?

A 30% budget reduction would cut approximately $1M in direct funds in Years 1 and 2. In order to best accommodate this budget cut, LEAP would both delay and extend the roll-out of multiple programs across all core functional areas. LEAP would reduce the number of "Proof of Concept" research projects, bringing more of these online beginning in Year 3. We would initially focus on making progress with projects for which we can see a clear path to Level 3 (ESM Implementation) by Year 3. In other words, in the face of a budget cut, we would first close the loop to full integration in CESM.

Specifically, LEAP would accommodate this budget reduction through eliminating or reducing the following budget items in Years 1 and 2.

Items to delay to start in Year 3 or eliminate:

- Eliminate the Capital Equipment line-item (currently $100k in Year 1). This was initially earmarked for cloud computing credits, although recent institutional commitments from Google and Microsoft mean we no longer need to request these funds from the NSF.
- Delay Hack-athon to Year 3. This program is vitally important for LEAP's broader impacts, although can be rolled-out in Year 3 in order to make expeditious progress on convergent research.
- Eliminate participation in AI4Earth in Year 2.
- Delay Journalist Workshops to begin in Year 3.
- Delay programming with American Museum of Natural History to begin in Year 3.
- Delay the "Lab-to-School" Summer Institute and Parent Trainingto begin in Year 3.
- Given the reduced scope for education and Knowledge Transfer in Year 1 and 2, we will reduce administrative staff by eliminating the Knowledge Transfer Program Manager position for these years (saving salary, fringe, and indirect cost). Responsibilities would be reallocated to the Assistant Director for Inclusive Education.

Reduced levels for Years 1 and 2:

- Reduce cost for a DEI consultant by 50%.
- Reduce graduate students by three, leading to a total of six students in Year 1 and seven students in Year 2. This will save stipends, fringe, indirect cost, and tuition.
- Reduce postdocs by two, leading to six postdocs in Years 1 and 2.
- Reduce post-baccalaureate BridgeToPhD students by one, leading to two in Year 1 and 2.
- Reduce cloud engineers by one, leading to one supporting LEAPangeo in Year 1 and 2.
- Reduce Publication Costs to $10k in Years 1 and 2.
- Send half as many trainees to NCAR for summer rotations, we could reduce costs for travel and housing. Since we now live in a more virtual world and the research scope would be reduced, a lesser physical presence at NCAR could be accommodated.

As some of the items above are currently covered from Institutional support, their elimination would allow for costs from the NSF Budget to shift to Institutional support.

As indicated in the table below, a reduced scope for year 1 and 2 would allow us to maintain a focus on research, but to still have significant impact on education, broadening participation and knowledge transfer.

LEAP Initiative	Proposed Percentage	Revised Percentage
Research	51%	60%
Education	19%	14%
Broadening Participation	9%	8%
Knowledge Transfer	10%	9%
Management & Operations	11%	9%

How Many Committees Will Each Member of the LEAP Executive Committee Be On? This Would Be Easiest Viewed As A Matrix Of People In The Rows And Meetings/ Committees/ Working Groups In The Columns. We Can Combine With The Earlier Question On Moving Parts - Number Of Meetings Etc

Monthly Meetings for LEAP Executive Committee

LEAP executive Committee	Committee Total	Executive Comm.	Convergence Subcomm.	Knowledge Transfer Subcomm.	Research Working Groups	Executive Coaching	Center Director Mentorship
Gentine	5						
McKinley	4						
Abernathey	3						
Burbano	3						
Cogburn	3						
Lawrence	3						
Revkin	2						
Schmidt	2						
Vondrick	3						
Zanna	3						
Zheng	3						

In Year 1, all of the Executive Committee will participate in the Management Institute (1.5 hour per month). In Year 1 and 2, Gentine will participate as a Provost Leadership Fellow (2 hours per month).

All members of the Executive Committee participate in 3 meetings each year to review progress and determine needed strategic shifts.

- March: Director's Council (4 hours)
- May: External Advisory Board (1 day)
- August: Annual Meeting (3 day)

In addition, there are annual commitments for some of the Executive Committee.

- January: Funding Review Panel (1 Convergence Subcommittee member, rotating)
- March and November: Corporate Working Group (Burbano)

- August: Set research working groups (Zanna and Vondrick)
- October: Climate Justice Leadership Board (Cogburn)

For Each Of The 5 Members of the Leadership Team, What Do You See Over The Next 5 Years Will Be The Fraction Of Your Effort Devoted To LEAP-related Research, LEAP-related Administration, and Teaching?

Distribution of effort for the 5 PIs and Co-PIs as in the proposal is indicated in the second column of the table below. In light of our experience preparing for the site review and the review committee's comments, we would like to increase the effort for Abernathey, Cogburn and Zheng to 1.5 mo, as indicated in the third column. The distribution across Research and Administration for this effort is indicated in columns 4 and 5.

Teaching is a core activity of university faculty. We are delighted that Columbia views LEAP's proposed courses as fully consistent with its educational mission. For the PIs and Co-PIs, our LEAP-related teaching will be covered by our 9-month salary as teaching faculty. Several courses in the Certificate Program (proposal section C1) are already established. New courses are *Climate Projection Challenges* (proposal section C3), the Research Seminars (proposal section C4), and the *Design Studio* (proposal section C8). These courses are well-aligned with the needs of our home Departments. As mentioned elsewhere, the past six months of working virtually has also given us new confidence in the ability to offer classes to students across the partner universities.

Leadership	Proposed Effort	Revised Effort	Research	Administration	Teaching
Gentine	3		1	2	Institutional
McKinley	2		1	1	Institutional
Abernathey	1	1.5	1	0.5	Institutional
Cogburn	1	1.5	0.5	1	
Zheng	1	1.5	0.75	0.75	Institutional

Your Proposal Was Written About 8 Months Ago. Since Then We Have All Experienced Changes in Our Everyday Life As Well As Changes In How Science And Education Are Conducted. If You Were To Write This Proposal Now, Briefly Describe What Would You Change?

While the socially-distant world we have inhabited for the last eight months has tested our ability to collaborate virtually, it has also dramatically increased our confidence in our ability to do so. We have achieved marked success with much greater dependence on Zoom, Slack and Github than we could have envisioned eight months ago. In addition, a heightened awareness of systemic racism has spurred increased attention and investment at all LEAP institutions, and we can leverage this energy to further advance our DEI strategies.

We are newly aware that we can effectively engage with a much broader range of potential partners to pursue our research, education, knowledge transfer, and DEI objectives. For example, LEAPangeo is already beginning to explore new collaborations with Historically Black Colleges and Universities to enhance data science curricula. Now that we can see how more physically distant partners could fully engage with LEAP, more such efforts would have been reflected in the proposal.

With the benefit of 2020 hindsight, we would have requested NSF funds to support distanced collaboration, education and meetings through new technologies. We would equip conference rooms at LEAP's dedicated headquarters to connect with state-of-the-art existing set-ups at NCAR, UC Irvine and Minnesota. NYU's expansion to its two global campuses means that its Washington Square campus is already retrofitted with an impressive suite of remote infrastructure. Such technologies include larger computer and projector screens, digital whiteboards, video that allows multiple views of a conference room, and other equipment that facilitates virtual meetings. We could also leverage Cogburn's leading expertise in innovative use of virtual reality for teaching and learning, and could have requested funds for the required equipment.

We do believe that LEAP's physical headquarters remains critical to supporting our team, and particularly to incorporate early career scientists in our Center. The faculty, postdocs, students, administrators, and Storytellers in Residence that *do* use the space will benefit tremendously from proximity to one another.

Across our universities and professional communities,multiple new and vigorous initiatives aim to enhance URM representation in STEM which should bolster LEAP's DEI strategies. Just by way of example, at Columbia:

- The Provost's Office has initiated an anti-racism institute for the Dean's Council (covering all units at the University), which is being designed and led by Cogburn, along with colleagues at the Columbia School of Social Work.
- McKinley, Bell, and Kingslake are members of a new DEI taskforce at Lamont Doherty Earth Observatory, chartered by Interim Director Maureen Raymo, who is also LEAP Sr. Personnel.
- McKinley chairs the diversity committee for Earth and Environmental Sciences, and in this capacity has joined a new "super-group" of the chairs of diversity committees of the science departments within Columbia Arts & Sciences. Several of the departmental committees have just been formed.

There Is Already Strong Collaboration Between Geoscientists And AI (e.g. At NOAA). How Does The Creation Of a LEAP STC Lead To Synergies That Aren't Going To Occur On Their Own As Climate Science And Data Science Naturally Converge?

Summary points:

- AI strategies at organizations like NOAA and NASA emphasize the importance of AI research but don't provide the focused structure needed **to fully integrate ML and Climate research via the creation of a new discipline**. For this, an STC is needed.
- LEAP will **accelerate and coordinate** the convergence between data science and climate science with sustained effort over 5 to 10 years, which can't be achieved without a deliberate targeted effort.
- LEAP's **synergistic approach** will drive advances in both data science and climate science.
- **More reliable projections on decadal-century timescales are only possible through a coordinated effort:** new machine-learning physics guided algorithms, improvement of Earth System models with uncertainty representation and reduction of structural model uncertainty (going beyond reducing initial condition uncertainty for weather forecasts)
- **Education of climate data scientists** that will become the next generation leaders in Earth System Model development and analysis.

There are indeed strong collaborations emerging between geoscientists and data scientists. While climate science and data science convergence may occur naturally, uncoordinated efforts which don't have buy-in from BOTH communities are less likely to come to fruition and may not spawn a sustained new research discipline (climate data science). **LEAP will accelerate this convergence through research and education.**

Research: LEAP will specifically focus on

- Climate and Earth System projections on decadal-century timescales, rather numerical weather forecasts. Improving the reliability of ESM projections requires a systematic collaborative effort between climate and data scientists over an extended period of time. We anticipate, based on prior experience implementing conventional parameterizations, that implementation of ML-based parameterizations into ESMs will encounter unforeseen challenges. Again, the sustained effort afforded by an STC will provide the platform for success.

- Structural uncertainty and model parameterizations, rather than a focus on initial condition uncertainty or a focus on improving computational speed through emulation. LEAP work will result in new more accurate parameterizations to be implemented in Earth System Models.
- The Earth System, as a whole, including the ocean, atmosphere, land, and cryosphere systems, including the carbon cycle, rather than on one component (e.g., the atmosphere). Research and application of ML methods to this broad diversity of problems will de facto challenge the data science field. The diversity of Earth system problems that will be addressed will require innovative and synergistic approaches that span data rich to data moderate to data poor regimes. The expansive and coordinated research in LEAP will systematically drive ML discipline forward.

Education: LEAP will

- Produce, through education and training, a new class of scientist (climate data scientist) with comprehensive understanding of both climate and advanced data science. Climate data scientists will leave a lasting legacy on Earth System Model development and analysis and will be attractive to businesses interested in either addressing the climate change problem through modified business practices or mitigating business risk associated with climate change.
- Create institutional knowledge in ML-based parameterization development, necessary to progress the fields of climate science and data science forward. The advances in ML will also advance other fields (engineering, biology, etc).
- Generate advances in geoscience through process-understanding of the Earth System which will be implemented in our teaching curriculum.

The Presentation Showed How The Current Understanding Of Physics Can Supplement ML To Develop Better Models. But How Can The Results From These Physics-informed Models Advance Our Understanding Of The Underlying Physics?

Center culture: Advancing our understanding of the underlying physics is a key objective of LEAP, and we designed the center to integrate data science into multiple levels of the scientific discovery process. The founding team, which includes both climate scientists and data scientists, will establish a center culture of advancing our knowledge and the representation of physical processes in the Earth System through the tight integration between climate science and data science. As mentioned in the proposal and 10-page response, our vision, our work, our education curriculum, and proposal review system will ensure that data scientists and climate scientists closely work together to solve problems, while training the next generation of climate data scientists by cultivating the investigation of physics and machine learning.

Overall research strategy: This strategy will create bidirectional transfer between the fields, forging opportunities for scientific discovery in two directions. Firstly, in a top down fashion, results from physics informed models will allow us to leverage domain knowledge to ask questions from the data. For example, causal inference enables our domain knowledge to be sketched as structured causal models in order to identify cause-effect relations, as demonstrated in a recent paper which learned the causal relationships in the Lorenz system (an idealized model of atmospheric convection) from timeseries data. Additionally, equation discovery along with domain knowledge allows discovering interpretable forms of hypothesized relationships. Secondly, in a bottom up fashion, the center will leverage data science to discover gaps in our existing knowledge. For example, physics-guided deep learning parameterizations will enable discovery of new relationships between quantities, prompting and motivating focused investigation into the underlying physical and biological mechanisms. Additionally, manipulating the latent representation of generative models will uncover properties that are missed by typical Earth system models In both cases, we believe data science will become an indispensable tool for scientific discovery on the Earth climate.

We envision physics-guided ML as part of an iterative process -- where feedback between the data and the physics is critical to driving advances -- and the center is structured to maximize the frequency of this feedback loop, with research led by co-Directors Vondrick and Zanna, a data scientist and a climate dynamicist, respectively.

Proof-of-concepts examples: We believe we are well-equipped to leverage the data science paradigm and advance our understanding of physics in this way. In our work, we have already shown how physics-informed models advance our understanding of physics. For example, physics-guided deep learning parameterizations, which have 1000s of parameters, can be interrogated to uncover relationships between different variables in a complex system. Work by Zanna (Bolton & Zanna, 2019 and Zanna & Bolton, 2020) and work by Gentine & Pritchard (e.g., Brenowitz et al., In revision, 2020), have shown that analyzing different features from the neural networks can enhance our understanding of how turbulent flows in the ocean and atmosphere operate. The interrogation of the deep learning algorithms will include "sensitivity analysis", given that neural networks are differentiable (similar to an adjoint analysis, which Zanna has expertise on) but also the analysis of feature maps, heat maps, etc to understand which relationships deep learning algorithms have uncovered. Transfer learning algorithms will also provide valuable information on which features are general to the system behavior (and are transportable) and which ones are specific to a given regime. Moreover, using domain knowledge (scientific insight) for processes such as ocean turbulence, atmospheric convection, etc, we will discover equations for complex processes from data, which can be noisy. Domain knowledge will inform which regularizations to use to balance simplicity with reconstruction of fidelity. For example, knowledge from non-dimensionalization provides a prior over which terms are important relative to other ones. The expression can themselves be interrogated because they are symbolic and easily interpretable. For example, Zanna + Bolton 2020, have shown that an ocean eddy parameterizations expression discovered by the machine learning algorithms, can be related to shearing and straining of the fluids, thereby enhancing our understanding of the turbulent nature of ocean processes and their interaction with the large-flow.

How Will LEAP Ensure That It Advances The State-of-the-art In Data Science And Machine-learning, Not Just In Climate Science, And Ensure That The Data Science Community Is Made Aware Of These Advancements?

Climate science drives technology innovation: Climate modeling poses new challenges that require fundamental advancements in data science methodology. As outlined in the proposal, machine learning *must* evolve along several directions to a) respect physical constraints, b) reliably extrapolate, c) robustly generalize, and d) provide interpretable and causal explanations. The infrastructure patterns in LEAPangeo will require innovations in data science more broadly, in areas such as data storage and interactive data-proximate computing. Historically, geosciences have a track record of fueling technology advancements in mathematical and computational scientists. We believe today's climate change challenges will be no different, and will synergistically require that data science Ph.D. students innovate novel methods in order to solve them. We envision LEAP catalyzing similar breakthroughs in machine learning fundamentals, due to the rich and complex nature of the climate problem.

Center structure requires co-development: In order to drive advances along these challenges, we have designed the center to co-develop data science with climate science. The center establishes two Co-Directors of research -- one who specializes in geoscience and one who specializes in data science -- who will work closely together to direct the research program. We designed the proposal review process, which is managed by the co-directors, to ensure that funded projects are innovative in both fields. The LEAP SP contains 28 climate scientists and 14 data scientists.

Research dissemination to data science community: As the LEAP team contains researchers for both disciplines, we are well-equipped to disseminate our findings across both fields. As new machine learning methods are developed by the team, we will publish them in the leading conferences and journals, e.g. NeurIPS, ICLR, and ICML. LEAP will also leverage its close relationship with the Data Science Institute, to which many SP are associated, to give talks. While the focus of LEAP will be machine learning for climate science, publishing them in diverse venues will allow the methodology to transfer to other application areas as well. The data science faculty will also include examples from their research into their data science courses. Several of the SP on the team also frequently organize conferences and workshops in machine learning and data science, which LEAP will leverage to disseminate our advances.

How Would You Improve Discoverability Of Assets Located In LEAPangeo To Support Fair Principles? How Would This Support Both the Unique Search Perspectives Of Climate Scientists And Data-scientists?

This is an excellent question, highlighting an important issue, which did not receive sufficient attention in our presentation and proposal. LEAPangeo is essentially a cloud-based data facility for analysis-ready climate data, analysis code, and machine-learning models. It is therefore essential that we follow best practices for making all LEAPangeo assets--data, code, and models--Findable, Accessible, Interoperable and Reusable (FAIR).

We first attempt to define the "unique search perspectives of climate scientists and data-scientists." We assert that climate scientists will primarily be searching for data based on a specific geoscience domain (e.g. *ocean sea-surface temperature*) or physical process (e.g. *ocean mesoscale turbulence*). Indeed, this is how data is organized and presented in common geophysical data repositories (e.g. NASA's PO.DAAC). Data scientists, in contrast, will often search for data associated with a particular machine learning problem (e.g. supervised learning, natural language processing), format (e.g. image, tabular), or license (e.g. creative commons, public domain). This is exemplified on popular ML online resources such as Kaggle. In order to achieve convergence between disciplines, LEAPangeo must expose the same data through both paradigms. It must also embody the principles of *linked data*, by exposing the relationships between different datasets and between data and code assets. These relationships will be displayed in an interactive, searchable form on the LEAP website. A great inspiration for what the LEAP website could be is the Radiant Earth ML Hub, which connects machine learning experts with geospatial imagery data, focused on agricultural applications and remote sensing.

We firmly believe that FAIR principles are best served by adopting established standards and best practices from the community. LEAPangeo will draw upon standards and best practices from EarthCube, Earth System Information Partners (ESIP), the Open Geospatial Consortium (OGC), the Research Data Alliance (RDA), and Schema.org. In the table below, we describe how FAIR principle will be implemented for data and code respectively.

FAIR Data

Findable	All data in the LEAPangeo data catalog will be cataloged using the Spatiotemporal Asset Catalog standard (STAC). STAC is a JSON metadata specification that serves as a foundation for both browse and search applications. A rich ecosystem of open-source tools built around STAC will make it easy to deploy a modern web front-end to the data library, as well as facilitating crawling of the data by search engines for maximum discoverability.
Accessible	Accessibility is central to the concept of LEAPangeo. Through our industry partnerships (Google and Microsoft), we will share datasets in cloud object storage. Data stored in cloud object storage (e.g. Google Cloud Storage) is maximally accessible on a global scale. (Netflix, for example, uses the same technology to store and distribute video data.) However, object storage requires new technological paradigms and software tools to enable accessibility, particularly for complex data. In the Pangeo project, we have pioneered the concept of *analysis-ready, cloud-optimized* (ARCO) data for climate datasets. ARCO data formats, such as Zarr, enable both quick looks at the data and metadata as well as facilitating massive-scale distributed processing.
Interoperable	Interoperability in data is achieved by using standard data, metadata, and

	catalog formats. All formats used in Pangeo will use openly licensed specifications (e.g. NetCDF, Zarr, STAC) which have implementations in many different scientific programming languages (e.g. Python, Julia, R).
Reusable	Datasets published in LEAPangeo will adopt Creative Commons licenses. These licenses give clear rules for reuse (commercial and non-commercial) and attribution, enabling downstream data consumers to derive new value from the datasets.

FAIR Code

Findable	Following industry best practices, we will use GitHub as our primary code collaboration platform. GitHub already has advanced features for code search, based on language and other tags, and is well indexed by major search engines. (We note that GitHub is owned by Microsoft, a LEAP partner; we will thus have access to all enterprise-level features of the platform.) However, GitHub is not an archival repository. Therefore, we will automate the linkage between GitHub and Zenodo to generate Zenodo archives (with DOI) for all code releases. The LEAP website will provide a browseable and searchable front-end for discovering all these code assets.
Accessible	GitHub and Zenodo together will provide full accessibility. Furthermore, we will strive to publish Binders for all LEAP code releases. Binder is a cloud-based technology which enables on-demand, interactive reproducibility of data science notebooks, drastically lowering the barrier for interested parties to actually run and modify the published code.
Interoperable	To facilitate interoperability, all LEAP code releases will strive to use only open-source dependencies. Environment specifications, required by binder, will be published with code, further enhancing interoperability by enabling the re-creation of the environment in which the code was run.
Reusable	All Pangeo code releases will employ OSI-approved open-source licenses, facilitating reuse in other research projects or commercial applications.

A third possible asset to share is the fully-trained ML models. Right now, the standards for publishing ML models with FAIR practices are far less well established than data and code. The Radiant Earth ML Hub is a leading example in the geospatial field. openml.org is one example of a GitHub-like platform for ML models and pipelines. There are also open frameworks for sharing ML models, such as ONXY, which allows for cross-language sharing of trained neural networks. We will monitor this area closely and adopt best practices as they emerge.

One strategy we will use to increase the reach and visibility of LEAPangeo datasets is to contribute datasets and leaderboards to establish “challenges” at computer science / data science conferences (e.g. NeurIPS, ICML, ICLR, CVPR) that focus on climate change processes. Such challenges (e.g. Kaggle) can attract thousands of participants and greatly increase the exposure of data scientists to climate problems. However, designing effective challenge problems requires care and interdisciplinary expertise--precisely the sort of expertise LEAPangeo will cultivate.

How, Concretely, Will LEAP Serve As An Exemplar For Other Subject Domains To Be Able To Emulate? How Will Those Other Interested Groups Be Facilitated In Creating Their Version Of Leap Through Processes Leap Pioneered?

Through the proposal process, the LEAP team identified and pioneered five *convergence strategies* as we laid out in our proposal and presented during the site visit. We believe all five strategies can be readily transferred to other *data-intensive* subject domains.

In the following, we rephrase our convergence strategies in more general terms, which can be broadly applicable to any data-intensive subject domain.

- ***CS1:*** *Leverage an open cloud computing platform* engaging a broader interdisciplinary community*;*
- ***CS2:*** *Gather a connected core group* of leading innovators and early adopters of machine learning in the domain science, and strongly motivated collaborators from data science*.*
- ***CS3:*** *Partner with a discipline-level community legacy* to mobilize broad interest and amplify potential impact.
- ***CS4:*** *Create alignment between research, education and DEI* to drive synergy and extend reach.
- ***CS5:*** *Establish a bidirectional communication pathway between academia, industry, and the public, forging mutual stakeholdership.*

Here are a number of ways that LEAP's proposed activities can help facilitate initiatives by other interested groups.

1. LEAP will broadly disseminate our best practices from LEAP's processes via publication, presentations, public communications and social media. In particular, since the data science community can serve as the unifying force for similar initiatives, LEAP will proactively disseminate LEAP's approach to academic leaders in data science by giving presentations at workshops for academic leaders such as the Academic Data Science Alliance and American Statistical Association's Caucus of Academic Representatives.

2. LEAP will make all products from our processes open-source, facilitating adoption by other interested groups. Again, since the data science community can serve as the unifying force for similar initiatives, our proposed dissemination efforts, as detailed in our response to question 8, hold the potential to drive adoption of LEAPs' processes in other subject domains.

3. LEAP will open our events such as lab-to-school parent conference, convergence luncheons, translate-a-thons, to interested groups, who can attend as participants and/or observers.

4. LEAP will extend an invitation to attend our annual meeting to interested mid-career scientists from other subject domains who aspire to create their version of LEAP.

   Our annual meeting touches upon all aspects of LEAP, providing a holistic demonstration of LEAP's processes, in terms of building and managing interdisciplinary centers, and driving convergent research, education, DEI and KT efforts.

How Could The Partnership Program Be Expanded To Incorporate Organizations' Interest In LEAP Outputs (Some Industry Have Strong Interests In The Data Climate Outputs, But There Are Other Intermediate Outputs/ Assets That Companies Could Be Interested In) Beyond Just Climate-specific Datasets And Simulation Results?

We appreciate the review team highlighting that LEAP has the potential to provide value to the private sector not only through climate-science-related outputs, but also through intermediate outputs such as algorithms and information on best practices and advances regarding physics-guided-AI technology, which has many potential applications in the private sector. Though not discussed explicitly during the site visit, an additional output of relevance to companies may be best practices gleaned from LEAP's DEI learnings, particularly given the increasing push to incorporate anti-racism practices into DEI in private sector organizations.

Our planned corporate engagement *processes* for bi-directional knowledge transfer with the private sector well-position us to gather critical feedback on what is most relevant on each of these topical domains. The change we can easily incorporate into our planned processes here would be to purposefully add the physics-guided-AI technology, and DEI, knowledge domains (in addition to the climate-data-knowledge domain). These areas may be incorporated into our feedback loop when gathering information on, and disseminating information about, the LEAP outputs of relevance to the private sector.

Our proposal indicated LEAP's intent to make all the LEAP outputs--data, models, and code--accessible via the LEAPangeo platform. These LEAP outputs will be made readily accessible under our current planned scope, and our response to Q9 indicates how we would make them discoverable by the relevant communities. Additionally, there are a few ways that we could expand the corporate engagement and partnership plan, as we outline below, to further incorporate organizations' interest in other LEAP outputs. If we obtain the award, we would welcome a conversation with the NSF about whether to implement these changes, given that they would involve a slight increase in scope:

1. Our previously described "Climate Information Portal" (the web interface directed at external stakeholders) could be re-conceptualized (and re-named) the "Learning the Earth with Artificial Intelligence and Physics (LEAP) Information Portal". In addition to providing the climate-related knowledge outputs on this portal, we could also aggregate, distill, and disseminate for a corporate audience physics-guided-AI-related knowledge and software outputs, including white papers on the technology and best practices learned from implementing the technology. If DEI-related best practices gleaned from LEAP are of relevance to the private sector, we could also aggregate, distill, and disseminate these in an executive-friendly manner and make these outputs available on our portal. This could include preparing a modification of the Inclusive Science Module and translating empirical reports/findings for corporate audiences.

2. We would be happy to explore the addition of a corporate partner who would represent companies less interested in the climate-related, and more interested in the physics-guided-AI-related, outputs emerging from LEAP. Companies in the autonomous self-driving car industry, whose business models need physically-guided AI to make self-driving safe and reliable would likely be a great fit for this if there is a sense from the review team that this additional corporate partner would add value. If the NSF would like us to, we could approach Uber or Toyota as a partner for this purpose. Alternatively, we could keep our corporate partnership set as it is, but ask our current partner Google to connect us directly with Waymo (owned by Alphabet, the parent company of Google), for the purpose of gathering feedback from a representative of this corporate audience segment.

What does high-level success look like 3-years in from the perspective of each of the five leadership areas?

We interpreted the five leadership areas as our convergent strategies and defined responses below across those 5 Convergence strategies, with corresponding bullet points.

For **CS1:** Success in LEAPangeo in year 3 will be characterized by the following key performance indicators:

- *High volume of activity by LEAP participants within the platform*. All major research activity should be taking place within LEAPangeo, not because participants are "forced" to, but because they are genuinely more productive and collaborative this way. Conversely, failure would mean participants fall back to their own private workflows and analysis environments.
- *Active dialog between LEAP scientists and software developers*. The Agile development methodology we intend to use means that LEAPangeo will never be "finished." Rather, the platform should

continuously evolve in response to user needs. Users should be engaged in reporting bugs and suggesting features, and developers should be responding quickly to such feedback.

- *Significant data usage by partner institutions and the broader community.* A major thrust of our knowledge transfer strategy is to expose climate data to the world in a useful format using cloud computing and the LEAPangeo platform. By year 3, partner institutions and other non-partner institutions in the broader community should be actively "consuming" LEAPangeo data and using it for diverse purposes, including research and education.

For **CS2:** Success in the new algorithms in year 3 will be characterized by the following key performance indicators:

- *Several implementations of hard constraints and physical invariance built into deep learning algorithms* will have been developed. By year 3, we will specifically target conservations of energy and mass which are critical for climate change simulations. These algorithms will be usable across Earth components (land, atmosphere, ocean, cryosphere) but we hope that they could also be used in other disciplines (which are at the intersection between physics and data rich environments e.g. mechanical engineering). **Two to three papers** will have been published in either conferences or journals. Codes will be shared on GitHub and their adoption will be used as an assessment metric too.
- *Embedding causal inference into ML* to improve generalization (e.g. unseen climates). Algorithms will have been developed that can include causal inference into ML to reduce spurious correlations and consequent instabilities in the models. An example of such an issue was highlighted by Brenowitz et al. 2020: spurious correlations between the stratosphere and the lower part of the atmosphere (boundary layer) were present when using a deep neural network to replicate moist convection in the atmosphere. These spurious correlations led to instability of the deep neural network model when implemented as a parameterization of convection in a coarse-resolution model. **At least two papers** will have been published, one showcasing the methodology and the other one demonstrating the application to one of LEAP's subcomponents (e.g. moist convection in the atmosphere). Codes will be shared on GitHub and used as an assessment metric too.
- *Understanding of processes using equation discovery*. By year 3, we will have discovered new understanding of **two or three** processes (either physical or biogeochemical) proving interpretable explanations, including via distilled equations. Specifically we hope to understand processes such as atmospheric convection or photosynthesis, in addition to refining ocean turbulence work pioneered by Zanna (Zanna and Bolton 2020) to more dynamical regimes and climate conditions. We expect to publish **two manuscripts** on this strategy.

For **CS3:** Success in terms of CESM integration in year 3 will be defined based on the following criteria:

- *Corrections of model structural errors using ML:* One of the important goals of LEAP is to reduce model parameterization structural errors (incorrect representation of the physics or biology leading to large biases). In three years we will have defined success if LEAP can implement **several (3-5) ML-based parameterizations** into CESM component models (atmosphere, land, ocean, land ice). Potential candidate parameterizations which are already at or close to the Level 2 stage (ready for implementation into component models), but need to be tested further, e.g., for numerical stability, impact on other coupled processes, and suitability in different climate regimes (e.g., warm/cold) would be (1) atmospheric convection (Gentine et al. 2018), (2) ocean turbulence and eddies (Bolton and Zanna, 2019), (3) warm cloud bin microphysics (Gettelman et al., 2020), and (4) physics-guided ML for streamflow (Yang et al., 2019).
- *Metrics and data products*: In years 1-3 LEAP scientists will utilize existing metrics packages to more comprehensively evaluate new ML-based parameterizations. This exercise will have two benefits. (1) It will provide a more comprehensive assessment of impact of new ML-based parameterizations on component model performance. (2) It will reveal limitations of existing metrics packages (for example, for the most part existing packages do not assess the ability of a model to replicate statistics of higher order moments, like extremes). In parallel, we will have developed new data "products" with clear uncertainty quantification that will be used to refine model metrics performance. In particular, we want to go beyond typical mean square errors at the global for the global

evaluation of models. Instead, the uncertainty quantification in space and time of various products (e.g. pCO2 Gloege and McKinley 2020 or photosynthesis and evapotranspiration Alemohammad et al. 2017) will be used to refine standard metrics and attribute varying weights (spatially) depending on the uncertainty quantification - regions with high uncertainty will be given a lower weight during the model metric evaluation. We expect to have developed **several global products (2-3)** with quantified uncertainties and developed **2 metrics** using those products.

- *Parameter estimation*: For the parameter estimation, by year 3, we will have expanded on the work of Elsaesser and Dagon for systematic tuning of parameters based on existing metrics (e.g. ILAMB for Land) using a Bayesian parameter inversion and emulators of the model subcomponents (land, atmosphere, ocean, cryosphere). This relates to an earlier point of the reviewers that this component of LEAP has parallels with data assimilation in numerical weather prediction for parameter estimation (data assimilation can also be used for parameter estimation in addition to state estimate). This will be achieved with first dimensional reduction so we can achieve Bayesian parameter estimation and limit the dimensionality of the problem following on recent work from the group (Dagon et al., 2019; Elsaesser et al., 2020). Here, we will use an emulator of the CESM subcomponents to more easily compute the "adjoint"/sensitivity of the outputs and metrics' response to model parameters instead of the very computationally expensive model itself. This work will leverage recent NCAR's efforts in large parameter perturbation ensembles for parameter tuning. We expect that this strategy will have bee**n deployed across CESM subcomponents** by the end of year 3.

For CS4: Success in Education + Research Convergence in year 3 will be defined based on the following key performance indicators:

- *Fully developed curriculum* that attract strong student interest, quantified as:
  - Four existing courses updated with LEAP-research integrated modules, case-studies and projects, with virtual options, supported by LEAP's *Design Studio*, facilitated by the LEAPangeo platform.
  - One new project-based course (Climate prediction challenges).
  - Two new convergence seminars driven by active LEAP research.
  - Six Design Studio interns over three years for designing and supporting research-education integration.
  - 20 students (with > 30% from either data science or climate science) each year participating in LEAP's certification programs.
  - At least two students from each LEAP university partner participate in LEAP's certification programs.

- *Research Experience Programs* with robust participation, quantified as:
  - Six doctoral dissertations will be in the advanced stages of development (completed qualifying exams, published at least one manuscript or contributed significantly to code development).
  - > 20 applicants each year have participated in the summer Research Internship program.
  - 12 summer interns (three cohorts of four) will have produced publishable (in journals or conferences) high-quality research results, leveraging the LEAPangeo platform, and with >30% of them choosing to further their studies or careers at the interface between climate and data science (climate data science) or to use their newly acquired technical skills to find a position in private or public sectors (e.g., government, startups, reinsurance companies, weather and risk companies).
  - > 10 high school students each year have participated in LEAP's offering in the Science Honor Program.
  - ~ 10 public school sustainability coordinators (NYC K-12 science teachers) have participated in LEAP's "Lab to School" program. >30% of them will be willing to adopt their developed teaching units informed by LEAP's research in their classroom.

- *Broadened Participation* from students from underrepresented populations will have been increased towards a trajectory to be on par with US population within 5-10 years
  - Three LEAP summer visitations to NCAR with SOARS Integration.

    - Engage > 15 of returning SOARS protégés in LEAP's summer workshop over three years at NCAR.
    - > 2 SOARS protégés mentored by LEAP researchers over three years.
    - > 2 returning SOARS protégé participates in LEAP summer research internship program over three years.
    - Three Bridge-to-PhD scholars will have been supported by LEAP and will have settled PhD positions after their Bridge (typically lasts two years) with an acceptance rate above standard > 66%.
    - Up to 15 (three cohorts of five) students though the National Society of Black Engineers participating in LEAP certificate courses.

- *Broader faculty interest*: we will have witnessed strong interest from Faculty and Trainees in LEAP's education programs
    - > 10 faculty members participate in LEAP curriculum.
    - > 20 faculty members participate in LEAP "train the trainer" workshops
    - > 20 faculty members actively mentor research trainees through LEAP fellowships, post-docs and summer internships.

CS5: Success in Knowledge Transfer in year 3 will be defined based on the following key performance indicators:

- **DEI and Broadening Participation**:
    - DEI strategy overview, empirically grounded recommendations regarding recruitment and STEM DEI strategies will have been disseminated in the form of 3 annual reports, two major academic publications (e.g. Science and Harvard Business Review, Harvard Data Science Review) and presentation or workshop at a two major STEM conferences (e.g. NeurIPS, Joint Statistical Meetings, Academic Data Science Alliance, American Geophysical Union)
    - The Inclusive Science Module based on formal annual assessments of the LEAP DEI strategy will be made publicly available online and promoted as part of other dissemination activities (e.g. publication, conferences, workshops)
    - We will have co-hosted the first of two AI4EARTH conferences with Microsoft with over 300 attendees (this may expand if virtual) including broad participation from our educational program participants.
    - Two community based public organizations (e.g. NYC Office of the Mayor, Trust for Governors Island) will be trained, will provide user feedback, and will be utilizing LEAPangeo data and computing to support their distinct missions.

- **Corporate Engagement**
    - 1,000+ corporations sent annual corporate engagement survey to ask for information on what is useful to them and to disseminate information about the knowledge outputs currently available from LEAP
    - Feedback gathered from 100+ corporations regarding data and informational needs/preferences (via annual surveys, feedback from website portal, in addition to partner organization engagement)
    - Active use of the LEAP portal by 30+ corporations/ entrepreneurs/ organizations

- **Public and Media Engagement**
    - Building on our early co-production work with Climate Central, Earth Journalism Network and other partners, by year 3 LEAP will have an active interface (web-based) through which news media (and civil society organizations) trying to identify and cut exposure and vulnerability to climate-related hazards are actively drawing on LEAP's emerging 10-40-year climate projections. We anticipate at least a dozen references in year 3, with LEAP output referenced in maps, reports, news stories and web features pursuing climate resilience and adaptation capacity.

- By year 3 we will host an online conference (with at least 1,000 "attending" - in line with Sustain What segments which are getting several thousands viewers per episode) targeting climate-focused media, students and community organizations and featuring the insights and output on climate-data science of at least 20 LEAP Storytellers in Residence, Journalism Fellows and Translate-a-thon participants from years 1 and 2.
- Through sustained outreach to, and engagement with, data-focused environmental journalists, at least one major news feature story or series (e.g. Science Magazine, NYT) on LEAP's climate-data science innovations will provide a compelling overview of the project's genesis and impacts.
- A video series will be posted showing the genesis and learnings of LEAP, built around highlights from LEAP-focused Earth Institute webcasts and the various LEAP events.

Thank You!