LEAP PROJECT DESCRIPTION

Name	Department	Inst.	Role	Relevant Expertise
Pierre Gentine	Earth & Env. Eng.	Columbia	**Lead PI**, Center Director, Director of Diversity	Hydrologic and carbon cycles, land-atmosphere interaction
Ryan Abernathey	Earth & Env. Sci.	Columbia	**Co-PI**, Co-Director of Corporate Engagement	Ocean circulation, climate dynamics, high-performance computing
Galen McKinley	Earth & Env. Sci.	Columbia	**Co-PI**, Co-Director of Research	Carbon cycle, biogeochemistry
Carl Vondrick	Comp. Sci.	Columbia	**Co-PI**, Co-Director of Research	Computer vision, unsupervised learning
Jeannette Wing	Comp. Sci.	Columbia	**Co-PI**, Director of Computation	Computer systems, security, privacy
David Gagne	Comp. Systems	NCAR	Liaison for Model Development	ML, model parameterization
Lucas Joppa	na	Microsoft	Co-Director of Corporate Engagement	Corporate sustainability, ML, conservation science, innovation
Gaspare LoDuca	Info. Technology	Columbia	Director of Cyberinfrastructure	Technology management and strategy
Gavin Schmidt	Goddard Institute	NASA	Liaison for Model Development	Atmospheric chemistry, outreach
Tian Zheng	Statistics	Columbia	Director of Education	ML, pattern identification, education

9 Physics Senior Personnel: Cane, Gettelman (NCAR), Goddard, Horton, Kingslake, Lawrence (NCAR), Lall, Morrison (NCAR), van Lier-Walqui. **13 Computation Senior Personnel:** Bareinboim, Fish, Gelman, Giometto, Halem (UMBC), Jebara, Kaiser, Kim, Kpotufe, Pearson (IBM), Rubenstein, Sukthankar (Google), Tippett. **4 Biology Senior Personnel:** DeFries, Uriarte, Weng (NASA), Williams.

1. CENTER RATIONALE. Climate change is the most global of the Grand Challenges, posing dramatic dangers to environmental and human sustainability [1,2]. It is critical to predict future climate in order to craft more adequate response, remediation, infrastructural, and adaptation strategies in the face of mounting environmental risk [3]. Despite substantial progress made in Earth system modeling over recent decades, fundamental prediction biases and inherent uncertainties limit the accuracy of regional predictions (Figure 1). Uncertainties manifest across the climate prediction pipeline (Figure 1):

1. Climate-forcing factors – population and economic growth, technology development, social behaviors – are highly-complex and insufficiently understood, causing **climate emission uncertainties**.
2. For given greenhouse gas concentration pathways, models predict climate change. One major metric is the mean global temperature response to CO_2 doubling, known as **climate sensitivity**. However, model estimates of equilibrium climate sensitivity (assuming a steady-state response) range dramatically between 2.1° and 5°C. This intermodel spread has not decreased since 1979.
3. Predictions for regional climate response (temperature, sea level, and hydrological cycle) and extreme events (droughts and floods) are even less precise (Figure 1), even though such predictions matter most for policy makers and for the people and communities that must adapt. These regional uncertainties come at tremendous economic and social costs [2].

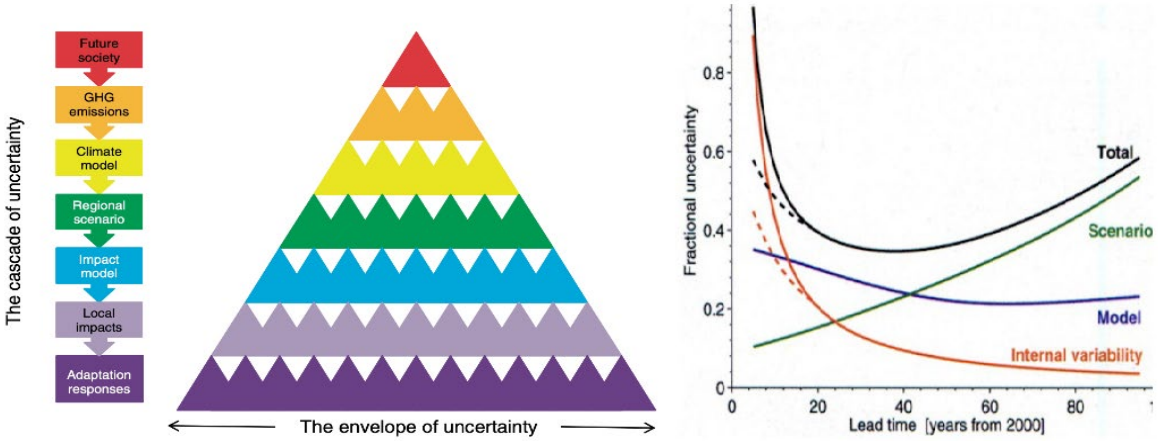
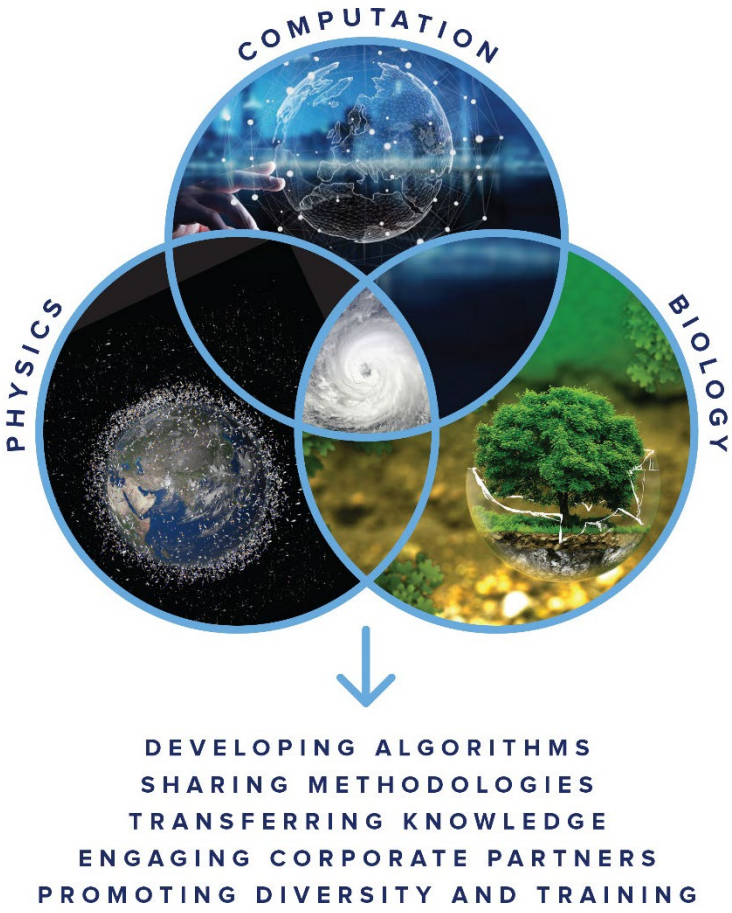


Figure 1: *(Left) Schematic showing the cascade of uncertainty of modeling, from robust adaptation uncertainty in climate change adaptation [4]; (Right) schematic of the impact of lead time on different components of uncertainties.*

Current Shortcomings. Prediction accuracy is limited by imperfect scientific representation and understanding of Earth processes [5–8]. **Most of the errors are caused by physical and biological processes represented on scales smaller than the model's grid resolution ("parameterization")**, which is of the order of 1° in the horizontal (or, ~100km). As guiding examples:

1. Atmospheric CO_2 concentration estimates for given emission scenarios are uncertain due to poorly understood oceanic, biospheric, and lithospheric carbon sink processes.
2. Given CO_2 concentrations, mean global temperature predictions have spread caused by imperfect representation of small-scale clouds and convection, as well as fundamental gaps in knowledge [9].



A New STC. Confronting this urgent scientific problem, our transdisciplinary team proposes a **novel and integrated approach to climate prediction by refining Earth system models' physical and biological subcomponents via novel machine learning algorithms**, aggressively leveraging the wealth of recent datasets to **simultaneously interpolate and extrapolate**. To deploy research, education, and diversity and corporate commitments towards this common goal, we propose a Science and Technology **Center for Learning the Earth with Artificial Intelligence and Physics (LEAP)** dedicated to machine learning (ML)-based development of novel subgrid-scale models ("parameterizations") within the open-source Community Earth System Model (CESM). Partnering with the nation's top modeling centers, LEAP pledges a **sustained and high-impact contribution to the climate modeling ecosystem**. To make this ambitious contribution, LEAP assembles 36 personnel – geoscientist and data scientist faculty, government scientists, and industry executives – towards a shared goal. To emphasize the quantity and quality of engagements, personnel names will henceforth be highlighted in **bold**.

Intellectual Merit. Though ML and geoscience are traditionally studied independently, their pairing reveals complementary advantages previously unachievable. Traditional simulations are grounded in physics to allow future extrapolation, yet fail to represent processes not fully captured by traditional parameterizations (for example, biology, turbulence, and clouds). While ML can learn representations from data without an underlying physical model, it has difficulty generalizing to unobserved conditions. By integrating both domains of expertise – physical knowledge (including biology) and ML – **LEAP will pioneer a new class of data-driven learning algorithms that respect physical laws and, unlike traditional ML, can extrapolate to future unseen conditions.** These algorithms will have broad applicability across other physics-rooted disciplines (especially biophysics and astronomy).

Transfusing ML with geoscience is particularly timely due to the recent proliferation of massive datasets, such as high-resolution simulations that resolve many small-scale and fast processes (for example, deep clouds [9–12]), remote sensing observations that can monitor those processes from meters to tens of thousands of kilometers in scales, and *in situ* observations (increasingly from autonomous sensors) in the atmosphere, on land, and in the ocean. However, extracting relevant information from multi-petabyte datasets hinders physical intuition and understanding; as a result, these data do not currently optimally

inform Earth system model (ESM) development, meaning predictive abilities lag behind their potential. LEAP will accelerate ML's integration into subgrid-scale ESM components via observational and high-resolution data, specifically implementing novel ML-based subgrid parameterizations of physical and biological processes within the CESM.

At the conclusion of the five- or ten-year NSF funding period, LEAP will have laid a foundation for broad research at the intersection of physics and artificial intelligence. This integration is expected to have long-term impact across fields and society by accelerating the discovery of new physics, enabling robust machine intelligence and robotics, and providing planetary insights for planning and development.

Exciting Challenges. Four challenges halt deriving optimal insight from existing data:

1. Extracting physical knowledge from very large observational and high-resolution simulation data;
2. Deriving ESM components (parameterizations) that learn from combined observational and high-resolution data, going beyond traditional data assimilation approaches;
3. Developing ML algorithms that better extrapolate and represent extremes, while preserving invariants and physical laws (for example, mass and energy conservation – required in climate models); and
4. Better quantifying and representing internal uncertainties (for example, stochasticity).

Recent ML developments, informed by high-resolution simulations, satellite, and *in situ* observations, will trigger advancements in modeling turbulence, convection, and biologically-mediated processes, thereby improving regional climate prediction to better anticipate future risks. Novel ML algorithm developments are critical for this endeavor, and must be guided by data and prior physical knowledge while simultaneously respecting physical laws, constraints, and invariances (for example, mass conservation, rotation, and translation). This approach aligns with developments in CapsNet deep learning [13], which includes a pose (translation plus rotation). Pursuing this goal will advance ML's frontiers via confrontation with out-of-sample generalizations, constraints, and invariances [14].

LEAP targets three categories of Earth system subcomponent uncertainties (Figure 3):

1. Compared to the true system, model systems possess inherent **structural errors** because current parameterizations only *approximate* complex biological and physical systems; as such, "traditional" parameterizations cannot fully resolve processes (for example, there are no clear cloud microphysical equations). Furthermore, physical and biological process knowledge gaps still exist: How does deep convection work, and how do microbes affect the carbon cycle? Without answers to these and other critical questions, parameterizations will deviate from the true system;
2. Physical model parameters (the "knobs" controlling model sensitivities) may not be optimally "tuned" to correctly represent the climate system, henceforth referred to as **model parameter errors**;
3. The climate system exhibits **internal stochasticity** (for instance, modes of internal variability) that limits capacity to provide exact, deterministic prediction, and which must be better quantified.

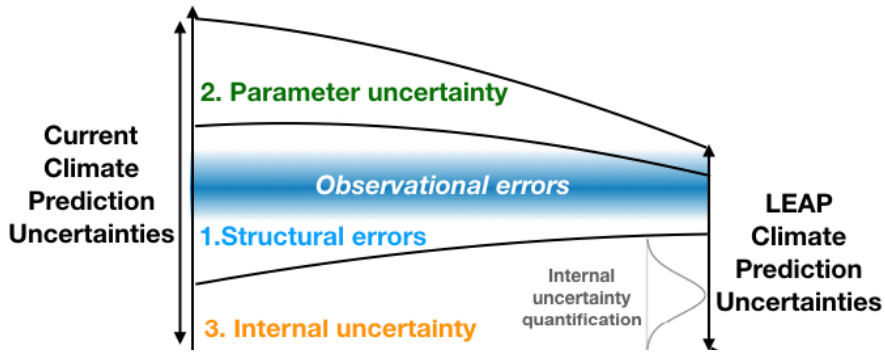


Figure 2: *Schematic showing the reduction in model structural and parameter errors and improved quantification of internal uncertainties by LEAP.*

For climate prediction, the most consequential uncertainties lie in structural errors of biospheric, cryospheric, oceanic, and atmospheric models. Indeed, for many processes, scientists do not fully understand relevant physics, meaning that new techniques are necessary to generate a leap in model physical description and prediction. LEAP will reduce these structural (#1) errors and parameter (#2) uncertainties, and will better quantify and represent internal uncertainties (#3) (Figure 2). Observations are fundamentally imperfect measurements of true climate systems given observational errors and data sparsity; this will be accounted for when deriving and constraining these new subgrid models.

This paradigm shift affords increasingly accurate ESM representation to better predict its future, and requires developing new infrastructures (particularly aligned with the cloud computing work of Co-PI **Abernathey**'s *Pangeo* project), as well as re-engineering subgrid physics codes to leverage the latest trends in computer hardware (such as towards GPUs, TPUs, and other specialized devices) for increased efficiency. This will trigger a dramatic acceleration of ESM simulations, as demonstrated by PI **Gentine** and colleagues with a tenfold increase in computational efficiency of the whole atmospheric model with an ML approach to deep clouds [11].

Competitors Analysis. LEAP's hybrid approach has not yet been pursued. The Center addresses key model structural errors and uncertainties by replacing *many* model components with learned parameterizations. Caltech's CLIMA effort is developing novel approaches to parameter optimization, but their models are based upon assumed deterministic physics; their approaches address neither structural errors with ML nor stochasticity. Secondly, LEAP will leverage existing community model infrastructures and open-source codes for the community; CLIMA is being built from scratch. Finally, LEAP will reduce uncertainties within all Earth system components, including the biosphere, cryosphere, and ocean; CLIMA focuses exclusively on the atmosphere and ocean.

The UK Research and Innovation recently funded three centers (Reading, Oxford, and Exeter) at the intersection of ML and environmental applications. With LEAP, the NSF will ensure that the US remains the leading epicenter of climate prediction and ML.

2. CENTER PLAN. LEAP will tackle structural errors in subgrid-scale models and parameter estimations, and will better assess and represent inherent uncertainties and stochasticity (Figure 3) while meeting physical constraints and invariances [14,15]. Three disciplines guide these efforts:

1. **Geoscience**: Building hybrid, physics-aware, ML-based parameterizations informed by high-resolution simulations and observations (atmosphere, ocean, cryosphere, and biosphere).
2. **Machine Learning**: Developing ML techniques informed by physics that respect physical constraints and invariances, thus better predicting, generalizing, and extrapolating outside of initial training sets.
3. **Computer Architecture**: Leveraging ML's symbiosis with recent hardware architectures (GPUs and TPUs) to improve the computational efficiency of ESMs that typically rely on parallelized CPUs.

Subgrid Models. LEAP will employ **five strategies** to define subgrid-scale models Y (Figure 3):

1. ML will learn subgrid-scale model processes from observations or high-resolution models. For example, observations can inform diurnal rainfall and cloud cover timing, yet are deficient in models. Subgrid-scale models can be deterministic *or* stochastic (for example, noise can be time- and state-dependent, and defined using Generative Adversarial Networks or variational autoencoders);
2. Via post-processing, ML will correct structural errors of physically-based models using the mismatch to observations [14] to improve the model's trajectory (more than a bias correction).
3. ML will learn coarse-grained (for example, at the coarse resolution of the ESM ~1°) output Y from high-resolution models (previously pioneered by PI **Gentine** [10,11,14]). Thereafter, another ML algorithm (emulator) will optimize the remaining parameters of the high-resolution physics based upon the coarse-grain variable mismatch with observations, yielding the most optimal model by uniquely leveraging high-resolution models and observations. For example, high-resolution cloud-resolving simulations can reproduce heating and moistening tendencies of deep clouds, yet are still depend upon smaller-scale parameters (turbulence and microphysics). An emulator will be used with, for example, Markov chain Monte Carlo methods, to infer the remaining parameters of turbulence and microphysics, best fitting the observations (for example, precipitation and top-of-the-atmosphere radiation).
4. Governing equations (such as nonlinear partial differential equations) can be theoretically upscaled from fine- to coarse-scales via multiscale averaging approaches; ML then targets the remaining phenomenological models dependent upon the coarse-scale state X (for example, a diffusion coefficient dependent on the flow characteristics) [16].
5. LEAP will develop additional novel algorithms to optimally merge high-resolution simulations and observations, with a focus on extrapolation.

These five strategies will be iterative. For example, a physical model can be optimized parameter-wise to best reproduce observations, yet may prove deficient in specific cases because of inherent structural errors.

ML will then be used to diagnose these errors (either via post-processing, following #2 above, or by comparing a direct ML-fitted algorithm to the model, point #1 above) and understand new physics. These errors will then be targeted for developing and refining the subgrid model.

Research Thrusts. LEAP encompasses four research thrusts, each greatly interdependent and co-led by one geoscientist and one data scientist. **Most investigators are assigned to two or three thrusts.**

Thrust 1: Atmosphere (**Gentine** (Co-Lead), **Vondrick** (Co-Lead), **Cane**, **Gagne**, **Gentine**, **Gettelman**, **Giometto**, **Goddard**, **Halem**, **Horton, Kim**, **McKinley**, **Morrison**, **Pearson**, **Rubenstein**, **Schmidt, Tippett, van Lier-Walqui**, **Wing**). New subgrid physics components will be informed by high-resolution data (direct numerical simulations that resolve turbulence, large-eddy simulations, and cloud resolving models), combined with data from satellites (for example, A-train constellation), lidar, radar observation, and other sources. LEAP will better represent several processes: microphysics, turbulence, shallow clouds, deep convection, and waves, and reproduce radiation with ML to speed up computing time. LEAP will employ a multiscale approach, starting from finer-scale processes that will be inserted into larger-scale models to better represent increasingly coarser scales. LEAP will leverage team members' expertise with using ML for deep convection parameterization and turbulence.

Thrust 2: Ocean (**McKinley** (Co-Lead), **Zheng** (Co-Lead), **Abernathey**, **Bareinboim**, **Cane**, **Giometto**, **Goddard**, **Jebara**, **Joppa**, **Kpotufe**, **Lall**, **Schmidt**, **Sukthankar**, **Tippett**, **Vondrick**). Ocean heat transport shapes climate, and sea temperature anomalies drive atmospheric responses, resulting in extreme weather fluctuations. The ocean also continuously absorbs heat from the atmosphere in response to greenhouse gas warming [17]. Properly simulating these transports requires parameterization of sub-grid mesoscale and sub-mesoscale eddy effects, and current models contain too many biases. LEAP will leverage new high-resolution simulations and satellite products to develop physics-constrained, ML-guided parameterizations of these processes. The ocean is also a key player in the global carbon cycle, absorbing approximately 30% of carbon emissions each year; improving ocean physical and biological representations will lead to significant improvements in sink predictions. Satellite data can resolve surface ocean chlorophyll and biomass, which LEAP will use to inform models. The biological removal of carbon to the deep ocean is much more sparsely observed [18–21], but new datasets are becoming available, particularly from NASA's EXPORTS project, which will be used to develop ML-constrained carbon export representations.

Thrust 3: Biosphere (**Gentine** (Co-Lead), **Gagne** (Co-Lead), **DeFries**, **Fish**, **Gagne**, **Gelman, Gentine**, **Joppa**, **Kaiser**, **Kim**, **Kpotufe**, **Lawrence**, **Pearson**, **Sukthankar**, **Vondrick**, **Weng**, **Williams**, **Wing**, **Zheng**). The biosphere thrust focuses on the terrestrial carbon cycle, water partitioning (evaporation, transpiration, and runoff), and land surface characteristics (vegetation type, biomass, and height), and on deriving parameters that interact with the atmosphere and ocean (such as albedo and phenology). It tackles physical and biological components for which high-resolution simulations rarely provide additional information (for example, biology is still not fully understood at a fine scale). The biosphere model processes scale from leaf-level physiological activities (photosynthesis, respiration, and transpiration), to individual demographic processes (reproduction and mortality), to site-level population dynamics and structural organization (vegetation 3D structure), and to regional and global vegetation dynamics and distribution, as well as disturbances (such as fires). LEAP will harvest data obtained from multiple scales, including long-term site censuses and experiments (FluxNet, LTER, FIA, and NEON) and remote sensing data (SIF, LiDAR, MODIS, vegetation optical depth, and fire monitoring), combined with expensive numerical models (for example, ED2 and PPA) to optimally define ecosystem functions. These data constrain diverse processes of terrestrial ecosystems at temporal and spatial scales.

Thrust 4: Cryosphere (**Kingslake** (Co-Lead), **Rubenstein** (Co-Lead), **Bareinboim**, **Fish**, **Gelman**, **Horton**, **Jebara**, **Kaiser**, **Lall**, **Lawrence**, **Vondrick**, **Zheng**). Sea-level rise projection uncertainties stem from misunderstanding of ice sheet physics. Large volumes of new data are available on ice sheet flow, topography, melting, and retreat; ML can transform representations of these components. Currently, ice sheet models simulate changes in ice volume using partial-differential equations over spatial scales of ~10^6m. Parameterizations describe sub-grid processes controlling, for example, heat exchange, basal sliding, and fracturing, and are tuned using present-day observations and geological constraints on past ice-sheet extent. Finally, these tuned models are used with projected climate forcings as boundary conditions to predict sea-level rise. All parameterizations suffer from resolution limitations. For example, ocean-ice heat transfer is the primary reason for Antarctica so rapidly losing ice [22]. The turbulent boundary-layer processes controlling this transfer operate at the sub-meter scale [23], but ice-dynamic response involves stress transmission over hundreds of kilometers [24]. Models cannot explicitly capture all relevant scales,

yet current parameterizations are far too crude [24,25]. ML can drastically improve these processes' representations [26]. In addition, ML will improve tuning hindcasts and quantifying projection uncertainties when large ensembles are prohibitively expensive [25,27] (Figure 4). With physical process understandings insufficiently developed, as with basal sliding and ice fracturing, ML will develop in tandem with physics-based models to better comprehend physics using recent datasets.

Cross-Cutting Themes. Conjoining these four aforementioned thrusts are **four cross-cutting themes**: commitments and intellectual infrastructures steering research thrust priorities. Once again, each theme is co-led by a geoscientist and data scientist; most investigators are assigned to one or two thrusts.

Theme 1: Coupling (**Fish** (Co-Lead), **Abernathey** (Co-Lead), **Cane**, **Gettelman**, **Giometto, Kaiser**, **Kpotufe**, **Pearson, Rubenstein**). Coupling ESM subcomponents can trigger specific modes of variability and internal behavior, as shown by the close coupling between air-sea fluxes and El Niño. For spatial-scale impact in the coupling, LEAP will systematically derive spatial coarse-graining of various processes at the coupled interface to define subgrid interactions. Another issue pertains to the fact that ESMs must be retuned once coupled. To address this, LEAP will use ML algorithms' computational efficiency to define grouped cost functions that can be targeted simultaneously and with different sensitivities. This should lead to an optimally-tweaked coupled model or subcomponent coupling.

Theme 2: ML Generalization (**Vondrick** (Co-Lead), **Lall** (Co-Lead), **Bareinboim**, **Gagne**, **Gentine**, **Jebara**, **Joppa, Kaiser**, **Kim**, **Halem, Lawrence**, **Morrison**, **Pearson, Sukthankar**, **Wing**, **Zheng**). ML excels when 1) large amounts of data are available; and 2) training and testing set distributions match [28]. Both conditions do not hold if future climate statistics are unlike the recorded past, because existing training data will be rendered obsolete. This non-stationary distribution makes extrapolation exceedingly challenging and ill-defined. Therefore, LEAP will investigate a new class of ML algorithms that integrate physical knowledge for robust generalization and extrapolation. The synthesis of physics with learning will furnish strong inductive biases to guide learning in the absence of data. LEAP will operationalize this integration in four ways:

- **Physical Constraints:** Since most current ML models lack physical constraints, LEAP will directly inject these constraints into predictive models to provide key insights for ESMs and climate predictions. PI **Gentine** and colleagues recently implemented physical constraints within deep neural networks, and there are other recent examples using physical invariances (such as rotation and translation [13]) or constraints within layers (such as OptNet [29]). One further advantage of imposing the physical constraint is reducing the data requirement, since the system's dimensionality is reduced. ML models will also be used to directly infer phenomenological models, for example a diffusion coefficient (Figure 3) instead of the entire physical relationship or equation.
- **Distributional Dynamics:** Although climate may be non-stationary, the underlying physical processes likely do not change, and we will leverage this fact to improve learning. This requires identifying loss functions that model stationary dynamics distributions instead of distributions of values (which are non-stationary). To achieve this, LEAP will directly model relative transformations over time to more accurately learn dynamics processes. Additionally, dimensional analysis can help collapse multivariate distributions into similar distributions of fewer variables [30];
- **Compositional Dynamics:** LEAP will investigate compositional predictive models that use logic and reasoning to seamlessly combine physical rules, and learned models to predict previously unobserved outcomes. For example, simple physical laws of dynamics can be combined to efficiently describe complex behaviors. LEAP will create chained-together learnable modules of dynamics, which can efficiently represent complex dynamics and generalize to novel combinations outside of the training set. These modules will be combined with physical simulations to enable hybrid physics-learning models.
- **Dataset Generation:** Though impossible to obtain natural observations of such events, LEAP can instead synthetically generate new datasets using physical simulation (such as cloud-resolving simulations at higher temperature [11]). These datasets will provide high-resolution data that span sufficient conditions to correctly inform ML representations. LEAP can directly train on these datasets, and use domain adaptation approaches to blend natural and synthetic domains.

Through this four-pronged approach, LEAP will answer two questions: 1) can an ML model generalize outside of its training dataset?; and 2) are physical processes indeed non-stationary? The first question is not fundamentally different for ML approaches compared to physical parameterizations, as the parameterization is typically tested and tuned for a select few cases (for example, trade-wind shallow cumulus

from a particular campaign). However, it is generally assumed, though not definitively proven, that physically-based parameterizations will generalize better than ML-based parameterizations. While there are times when physically-based models cannot correctly predict future states, the integration of physics and learning will create models that robustly extrapolate to unobserved climate events.

Theme 3: Uncertainty Quantification (**van Lier-Walqui** (Co-Lead), **Gagne** (Co-Lead), **Gelman**, **Gettelman**, **Giometto**, **Goddard**, **Jebara**, **Kpotufe**, **Lall**, **Morrison**). Observation-informed model systems cannot avoid uncertainties directly related to observational uncertainties (for example, instrument noise and sampling errors). In parameter estimation and model selection problems, uncertainty quantification is achieved by employing Bayesian inference methods that return estimates of probability to define both optimal solution and uncertainty. Historically, these methods have been too computationally-intensive to use in full ESMs; however, the use of ML as model emulators allows for dense exploration and constraint of model parameter distribution by observations, with the additional benefit of estimating model uncertainty. Additionally, remaining model structural uncertainties can be quantified by making use of probabilistic methods for deep neural networks [31].

Theme 4: Interpretable Learning (**Vondrick** (Co-Lead), **Sukthankar** (Co-Lead), **Bareinboim**, **Jebara**, **Kaiser**, **Kpotufe**, **Pearson**, **Rubenstein**, **Zheng**)**.** To examine representations acquired by machine learning, LEAP will interpret learned models to reveal both limitations and new physics, employing expertise in interpretable ML [32,33] to extract governing equations from neural network-type models. For example, given a black-box model to forecast ocean transport, LEAP will determine what quantities and relations the model automatically discovers. The full model will be distilled into a compact set of differential equations that approximate the dynamics process learned by the neural network. To maintain interpretability, equations will be regularized towards simplicity while balancing their approximation's fidelity. To validate this approach and empirically guide the research, LEAP will evaluate this approach on known physical equations that gradually escalate in complexity; for example, by using a model hierarchy approach going from: 1) toy models, such as chaotic models of convection and their interaction with the mean flow [34]; to 2) turbulent models (such as 3D direct numerical simulations, large eddy simulations, and quasi-geostrophic 2D models); to 3) ESM's full complexity modules.

Learning from Combined Models and Data. Weather predictions have dramatically progressed due to advances in data assimilation techniques, such as filtering [35,36]. Although weather models have evolved in resolution and process representation quality, the frequent assimilation of observational data has been the primary reason for alleviating many model structural errors and reduces departures of the forecast from observations [34]. As such, weather forecasting is a fundamental initial-value problem: unlike longer-term climate prediction. Therefore, techniques such as data assimilation cannot be directly used for climate prediction. Instead, techniques must be defined to 1) reduce model structural errors; 2) estimate underlying physical model parameters; and 3) quantify uncertainties (Figure 4). Estimating parameters and quantifying uncertainties have largely been an afterthought [37] because of the limited human resources in climate centers, and because of limited systematic and automated approaches for tuning climate model parameters [37].

Major challenges remain in physics model representation (for example, in model structures like clouds and convection) that require new types of parameterizations (Figure 3). Furthermore, prediction uncertainty quantification is currently limited: Earth system models are exceedingly computationally-demanding, with only 3-10 simulations per scenario recommended for CMIP5 or 6 [38] – a dramatically low number of replicates for characterizing the distribution of such a dynamically-rich global climate system. There are dedicated efforts to include more ensemble members, but they are computationally expensive. Given the current cost of model complexity, there must be greater effort to adopt models that are more efficient than the current generation of ESMs in order to yield better uncertainty quantification [1,5,39–42].

Here, LEAP will leverage the increased boost in efficiency from integration with ML-based modules with modern GPU and TPU architectures, which will substantially reduce the cost of model runs, thereby affording generating many more ensemble members at a cost comparable to a previous single simulation. The roadmap for hybrid CPU-GPU platforms by vendors (particularly IBM, involved in this proposal) employing these improvements will significantly reduce this bottleneck, thereby enabling the combined use of ML-enabled ESMs. At first, LEAP will keep the horizontal/vertical resolution of the current CESM model, but, if computational cost is dramatically reduced, we will assess the advantage of using finer resolution LEAP will also develop deterministic *and* stochastic parameterizations, in which the noise structure and its state dependence are also learnt based on the data (from both high-resolution models and observations).

Engineering Contributions. PI **Gentine** and colleagues implemented ML algorithms of deep clouds, and created routines to convert TensorFlow outputs into Fortran: ESM's primary language (for example, CESM and NASA GISS's ModelE). LEAP will follow this same strategy for rapid implementation in Fortran. New algorithms will be developed in collaboration with the NSF-supported open-source CESM, plus NASA GISS. **LEAP will release the new subgrid models on the community ESM (CESM)**, and will collaborate with GISS on process models like large-eddy simulation and remote sensing observation. The Center will hire software engineers at Columbia and in collaboration with NCAR to optimally implement the new algorithms within CESM; these implementations will use recent hardware (GPUs and TPUs) to its full potential. High-resolution simulations will run on NSF supercomputers (particularly Cheyenne), with remote sensing analyses completed at NASA GISS.

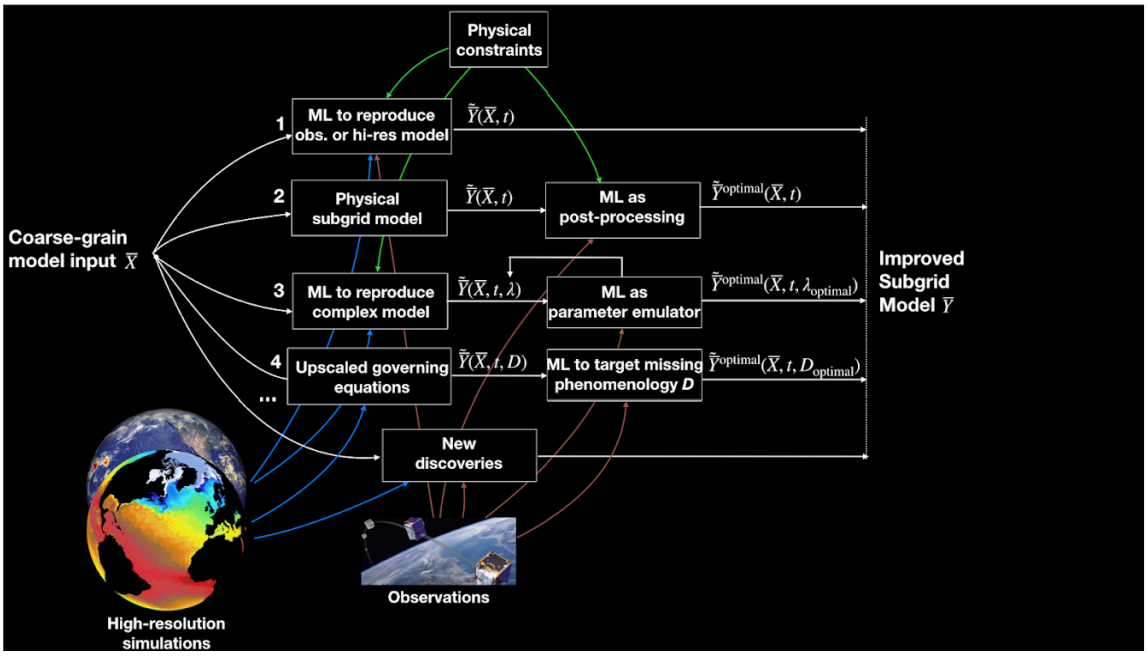


Figure 3: *Schematic describing different strategies to leverage both observations and high-resolution simulations using ML to inform subgrid parameterizations.*

Anticipated Bottlenecks. LEAP's primary challenges lie in 1) evaluating new hybrid models for enhanced generalization and extrapolation; and 2) achieving numerical stability. Contrary to typical ML techniques that focus exclusively on structures and interpolation, LEAP focuses upon extrapolation and generalization. The Center will develop high-resolution simulations in future scenarios (for example, higher sea surface temperatures or higher greenhouse gas concentrations) aimed at improving and testing generalization to unseen climates. LEAP will also improve extrapolation integrating a combination of physical knowledge and physical constraints directly into ML algorithms as described in more detail in section Theme 2. LEAP will additionally formulate improved cost functions minimized by ML algorithms to include physical constraints, and simultaneously focused on improved extrapolation and better representation of the full distributions and their tails, not just the mean.

A further difficulty relates to the ML subgrid-scale model's numerical instability [43]. LEAP will develop diagnostic tools for numerical instability, either by rejecting ML models that lead to instability, or, in certain cases, limiting ML algorithmic behavior. For example, LEAP will compute the Lyapunov exponents (error growth characterization) by computing the eigenvalues of the ML algorithm's Jacobian. These eigenvalues will inform error growth exploration; LEAP will limit them in the algorithms before implementation. A more gradual challenge relates to using different languages across different communities. To bridge this divide, LEAP will facilitate a range of interdisciplinary training, seminars and working groups to foster transdisciplinary exposure.

Broader Impacts: Education and Diversity. LEAP blends research with education, using the convergence between data science and geoscience to educate and develop the next generation of scholars. This will be accomplished through embedding four categories of trainees within each of the Center's four research thrusts, four cross-cutting themes, and multiple knowledge transfer initiatives:

1. **Postdocs**: PhD-level research scholars housed within the Data Science Institute. Postdocs will hold three-year appointments, and will be primarily physically located in either NASA GISS (a five-minute walk from Columbia) or NCAR (in Boulder, CO). LEAP will budget substantial travel costs to support the frequent movement of postdocs and faculty between New York City and Boulder, and will employ a range of telecommunications technologies to support large-distance group meetings.
2. **Graduate Students**: PhD students across any of Columbia's nine participating departments, jointly-supervised by one geoscientist and one data scientist. Students will frequently travel to NCAR to collaboratively implement subgrid components, and will obtain the doctorate in five years' time.
3. **Diversity Post-Baccalaureates**: LEAP will partner with Columbia's Bridge-to-PhD program to enhance the participation of students from underrepresented groups in STEM PhD programs. LEAP will support Bridge students who recently obtained their Bachelor's degree and who can use LEAP to advance their graduate school prospects. Bridge students will be closely supervised and mentored by the faculty leads within the research thrust in which they are placed, and receive further mentorship, professional development, and general support from Bridge administrative staff.
4. **High School Students**: LEAP will partner with the Eagle Academy for Young Men of Harlem, a 374-student, African-American charter high school, whose educational philosophy seeks to "foster positive livelihood outcomes for young men in communities where they are at high risk for incarceration." Located 20 blocks from Columbia's campus, Eagle students will engage for-credit, semester-long internships in LEAP's four research thrusts and four cross-cutting themes. Each year, up to four Eagle students will travel to NCAR for weeklong, immersive research trainings.

Each year, LEAP will support **six postdocs, ten grad students, six Bridge students, and 12 high school interns**: Adding trainees to the 36 proposal personnel, LEAP will comprise **70 researchers** of greatly diverse backgrounds and perspectives, plus staff. Each trainee will spend one year embedded within a research thrust or theme (with the exception of high school students, whose internships will last one semester); in the following year, each trainee will "rotate" into another thrust or theme.

Broader Impacts: Climate & Society MA Program. LEAP will partner with Columbia's Program in Climate & Society, a 12-month interdisciplinary Master of Arts program training professionals and academics to understand and cope with climate impacts on society. Through courses and research, students gain knowledge in both geoscience and social science. This partnership will take two forms:

1. Launching a new elective course, *Climate Prediction Challenges*, to provide future policymakers with training in data science; and
2. Co-hosting the Annual Symposium by contributing policy content and audiences.

Broader Impacts: Center for Science and Society. Residing within Columbia's Department of History, the Mellon-funded Center for Science and Society facilitates research and outreach between natural scientists, social scientists, engineers, and policy scholars. Through public programming series, seed grants, and visiting scholar programs, the Center is one of Columbia's leading units for engaging the general public in interdisciplinary scholarship. LEAP and the Center will partner to develop a "*Public Conversations in Climate Prediction*" speaker series. Open to the public and free-of-charge, these *Public Conversations* will host roundtable discussions from leaders in climate prediction and data science. Each STC-supported graduate student will host one *Public Conversation* per year, and will receive budgetary, faculty, and administrative mentorship support to propose, design, and execute this public event. With six graduate students at any time, there will be six *Public Conversations* per year, two of which at NCAR.

Broader Impacts: Annual Symposium. Beginning in Year 2, LEAP will organize an annual **Climate Prediction & Data Science Symposium** for the broader community, to showcase its research processes and products, its overarching motivating interests and concerns, and the current and future roles of its corporate partners and trainees. Featuring keynote speakers, corporate roundtable discussions, presentations from NSF, DOE, and NASA program officers, lightning talks, student posters, and travel grants, this event will be Live Streamed with the video feed posted to the Center's website.

Broader Impacts: Evaluation. If invited to submit an STC full proposal, LEAP will recruit a data scientist with expertise in education programs to serve as Senior Personnel chiefly responsible for trainee assessment and reporting through a combination of summative and formative instruments and protocols. This evaluator will lead a for-credit course for STC-supported graduate students and Bridge students, training them to conduct their own quantitative and qualitative evaluation interventions.

3. TEAM DESCRIPTION. Leading LEAP will be **Columbia University in the City of New York**, which leverages its three most expansive and high-impact basic science initiatives: the Data Science Institute, the Earth Institute, and the Lamont-Doherty Earth Observatory. Complementing Columbia is the **University of Maryland, Baltimore County**, contributing expertise in ML for climate prediction and quantum computing. Algorithms will be deployed in the **National Center for Climate Research**'s CESM model, along with NASA's **Goddard Institute for Space Studies**.

LEAP Leadership. PI **Gentine** will serve as **Center Director**, having pioneered using ML for climate convection modeling, continental biosphere observations, and atmospheric turbulence. PI **Gentine** and colleagues are developing physics-based hybrid models for the atmosphere, energy conservation for convection, and the biosphere. As Center Director, **Gentine** will be responsible for all Center operations and liaising with the NSF. He will Chair the Center's Executive Committee of Director-level Co-PIs and Senior Personnel, supervise staff, and liaise with LEAP's two advisory committees.

Success ML-geophysics convergence can only happen in tandem with supporting academia to look more like the society in which it is situated; based at Columbia University *in the City of New York*, LEAP is uniquely and optimally poised to accomplish this. Therefore, PI **Gentine** will also serve as the **Director of Diversity** so that diversity initiatives come from the Center's top executive; in this capacity, Gentine will manage LEAP's partnership with the Eagle Academy and the Bridge-to-PhD program. In close collaboration with Director of Education **Zheng**, he will design the diversity and teacher engagement portions of the Annual Conference.

Co-PIs **McKinley** and **Vondrick** will serve as **Co-Directors of Research. McKinley** studies ocean carbon cycling and its sensitivity to climate variability and change. She engages with the Data Science Institute to improve quantification of ocean carbon uptake through the joint use of surface ocean data, climate and ocean models, and data science techniques. **Vondrick** is an emerging leader in machine learning and vision, focusing on dynamics. They will Co-Chair the Research Subcommittee.

Co-PI **Abernathey** and Senior Personnel **Joppa** will serve as **Co-Directors for Corporate Engagement**. **Abernathey** extensively collaborates with Google, Amazon, and Microsoft on cloud computing through his active *NSF Pangeo* project building scalable environments for Big Data climate analytics; **Joppa** is Microsoft's first Chief Environmental Officer. Both will Co-Chair the Industry Subcommittee.

Co-PI **Wing** will serve as the **Director of Computation and Data**, and leads the Columbia-wide and presidential-level Data Science Institute, which builds upon the University's collective expertise in the foundations of data science (computation, ML, optimization, and statistics) and applications across all disciplines. Prior to joining Columbia, **Wing** was Corporate Vice President of Microsoft Research and Assistant Director of the NSF's CISE Directorate. This pre-proposal is highly motivated by her report emerging out of the NSF CISE's 2018 workshop [44].

Senior Personnel **Zheng** will serve as **Director of Education**, responsible for developing innovative teaching and learning programs for LEAP's trainees, and collaborating with PI **Gentine** in developing inclusive educational platforms that promote diversity in STEM. **Zheng** is the Chair of Columbia's Department of Statistics, and Associate Director for Education for the Data Science Institute.

Senior Personnel **LoDuca**, Columbia's Chief Information Officer and Vice President of Information Technology, will support LEAP as the **Director of Cyberinfrastructure** to coordinate the acquisition and development of shared facilities and cloud-based computing resources, and train all Center community members in the advanced usage of these resources.

Senior Personnel **Schmidt** will serve as the **Liaison to GISS Model Development**, having been involved in large-scale climate modeling for 20 years, and is the Director of NASA GISS and Lead PI on the ModelE contributions to CMIP6. He co-founded an annual workshop on climate informatics [45,46].

Senior Personnel **Gagne** will be the **Liaison to NCAR Model Development**, and focuses on developing ML for predicting high-impact weather and parameterizing subgrid processes in numerical models. He developed an ML hail guidance system transitioning to the National Weather Service.

Investigator Team. Supporting Director leadership are 26 senior personnel spanning three disciplines:

1. Physics. Cane devised the first numerical model able to simulate El Niño, used in 1985 to make the first physically-based El Niño forecasts. He continues to work on El Niño prediction, and has worked extensively on the impact of El Niño on human activity. **Gettelman** is an expert in physical parameterization of clouds and aerosols, having developed and integrated cloud parameterizations in global models via a variety of tools including different types of emulation. **Goddard** directs Columbia's International Research Institute for Climate and Society and leads its research efforts on understanding and predicting climate change on the 10-20 year horizon. She is a globally recognized expert on El Niño and La Niña, decadal prediction, and near-term climate change. **Horton** holds expertise in extreme events and techniques for integrating diverse types of information into low-probability, high-consequence, risk-based scenarios. He has developed risk-based scenarios to support adaptation across the globe. **Kingslake** studies the growth of ice sheets, primarily using satellite remote sensing, field observations, and physics-based modeling to examine the physics of ice and of water flow in ice sheets. **Lawrence** studies land surface processes and climate change, with a particular emphasis on Arctic terrestrial climate system feedbacks, including the impact of permafrost degradation on carbon, water, and energy cycles. **Lall** studies hydrology, nonlinear dynamical systems, nonparametric methods of function estimation, and their application to spatio-temporal dynamical systems. **Morrison** holds expertise in cloud and convective dynamics and microphysics, having led or co-led development of cloud microphysics parameterizations in several models including the Weather Research & Forecasting model and Community Earth System Model. **Tippett** detects and attributes climate change in models and observations on centennial timescales, decadal prediction of Atlantic sea surface temperatures, seasonal forecasts of the El Niño-Southern Oscillation, and seasonal outlooks for temperature and precipitation. **van Lier-Walqui** employs Bayesian inference to improve and quantify uncertainty in atmospheric models, focusing on developing and improving models of cloud and precipitation physics using advanced polarimetric and Doppler radar observations.

2. Computation. Bareinboim focuses on causal and counterfactual inference with applications to data-driven fields. **Fish**'s multiscale software has been employed by over 150 companies in the aerospace, automotive, energy, consumer goods, and biomedical industries. **Gelman** focuses on multilevel modeling and Bayesian statistics, including theory, methods, and computation. His published work on Earth science includes reconstruction of climate from tree ring data. **Giometto** employs high-fidelity computational techniques (such as direct and large-eddy simulation) and analytical tools to gain insight into atmospheric turbulence and reduced-order models. **Halem**'s studies high-performance and service-oriented computing, adiabatic quantum computing, information systems, and satellite data fusion. He has served as Assistant Director, Information Sciences and CIO for the NASA Goddard Space Flight Center, and Chief of Earth and Space Data Computing. **Jebara** focuses on probabilistic graphical models with applications in recommendation, spatiotemporal, and visual data. He has published articles that leverage ML in oceanic dynamics, ENSO, and subsurface sea temperature domains. **Kpotufe** works at the intersection of ML and nonparametrics, with an emphasis on high-dimensional data analysis problems with noisy and highly-nonlinear patterns. **Kaiser** specializes in finding logic errors in ML and big data applications; she was the first to apply metamorphic testing to ML, developing foundations for determining the metamorphic relationships and transformations used for ML. **Kim** studies computer architecture, parallel programming, compilers, and low-power computing. Her work explores low-cost chip manufacturing systems, reconfigurable communication networks, and fine-grained parallel application profiling techniques. **Pearson** develops HPC and ML technologies for IBM. **Rubenstein** designs and analyzes computer communication network systems, with interests in security and wireless technologies.

3. Biology. DeFries uses satellite images and field surveys to examine how food demands change tropical land use. Her research quantifies how land use changes affect climate, biodiversity, and other ecosystem services. **Uriarte** studies tropical forest ecology, specifically demographic analyses and forest responses to disturbances. She works with field and remotely-sensed data to develop a range of modeling approaches. **Weng** develops ecosystem biogeochemical models, vegetation demographic models, and Bayesian approaches to study terrestrial vegetation dynamics, fire-vegetation feedbacks, ecosystem carbon-water interactions, and biogeochemical cycles. **Williams** examines fires and the causes and consequences of hydrological extremes such as drought and extreme precipitation.

4. INTEGRATION STRATEGIES. LEAP will launch a **Certificate of Professional Achievement in Climate Prediction**: A three-course, 12-credit sequence spanning ML, geophysics, and biogeochemistry,

available to all doctoral students at New York City institutions; if invited to submit a full proposal, LEAP will initiate internal applications for certificate approval.

LEAP will also hire an Executive Director, reporting directly to PI **Gentine**, responsible for financial management, staff supervision, coordinating the Certificate program, marketing and internal communications, and staffing the Executive Committee and all subcommittees. The Executive Director should possess an advanced degree plus over ten years of relevant work experience.

Two Advisory Committees. LEAP will establish an External Advisory Committee (EAC) of non-affiliate executives in key focal areas. This EAC will meet annually, timed with the NSF's annual site visit, and will place a premium on assembling a diverse team in terms of ethnicity, gender, gender identity, and ability.

LEAP will create a Leadership Advisory Board (LAB) of senior executives spanning partner institutions. LAB will meet annually to receive PI **Gentine**'s Annual Report presentation. The LAB will be responsible for identifying novel institutional partnerships and resolving any arising intellectual property concerns in order to propel the Center towards convergent impact. It will be Chaired by Columbia's Executive Vice President for Research (PhD, Geophysics).

Executive Committee. PI **Gentine** will chair an Executive Committee (EC) comprised of the nine other LEAP Directors. The EC will meet monthly to granularly review Center progress, including selecting seed grant recipients and trainees receiving funding, resolving internal conflicts, and dissolving institutional barriers. The EC will be administratively staffed by the Executive Director and one administrator from Columbia's Data Science Institute and the Office of the Executive Vice President for Research.

Research Subcommittee. Co-Chaired by Co-PIs **McKinley** and **Vondrick**, and including Leads of the four research thrusts and four cross-cutting themes, the Research Subcommittee will administer a seed grant competition pairing one geoscience faculty member with one data science trainee (or vice versa) to develop year-long research projects. This Subcommittee will design application and review processes, and plan the Center for Science and Society's *Public Conversations* discussion series.

Industry Subcommittee. LEAP's key knowledge transfer contribution will be industry stakeholdership, given the profound implications of its climate prediction research for business, namely pertaining to investment risk. LEAP's corporate strategy takes four forms:

- Industry collaboratively using LEAP's physical facilities and computational platforms.
- Summer internships for graduate students embedded into the company's headquarters.
- Co-hosting the Annual Symposium, including one panel discussion on sustainable finance risk.
- Executives-in-Residence, whereby corporate leaders are embedded in Columbia's community to teach courses, advise student entrepreneurship, and accelerate paths of promising technologies.

Corporate engagement will be coordinated by an Industry Subcommittee, co-chaired by Co-PI **Abernathey** and Senior Personnel **Joppa** (Microsoft). Senior Personnel **Sukthankar** (Google) and **Pearson** (IBM) will be key members of this Subcommittee, among others to-be-named. The Subcommittee will include *ex officio* members from Columbia's Technology Ventures and Data Science Institute.

A Great LEAP Forward. Shortcomings in both geoscientific knowledge and machine learning stunt scientists from sufficiently predicting climate for the benefit of human populations; most importantly, historic divides between these two aforementioned perspectives hinder each discipline from pivotal advances. Our team sees tremendous promise in developing a transdisciplinary effort to forge prediction methods that dramatically improve climate prediction at both global and regional scales. This highly-ambitious endeavor is only possible with trailblazing engagement by our nation's top climate research laboratories and major computing organizations, and, most importantly, with the talents and commitments of new and diverse trainees. LEAP requires interconnected thrusts and educational programs, which can only be made possible by a long-term funding mechanism as robust as the NSF's Science and Technology Centers program. Ameliorating this critical shortcoming is a worthy and necessary venture for our nation, and will provide a legacy for the next generation of scientists, policymakers, and professionals both studying and affected by climate change.