LEAP RFP form (unlike standard proposals, the information required is very minimal and to some degree quite informal): <https://docs.google.com/forms/d/e/1FAIpQLSdH5SlG6BUsqJ7aqIhZ5L8fPFglWVYWJu6X6Dwfb5jHkP3PaQ/viewform>

Deadline: End of day 15 November.

**Main questions:**

**What is the goal of your project and how will it advance LEAP’s vision? Describe your bold idea with minimal jargon (i.e., understandable to both climate and ML communities). (1 paragraph)**

This project aims to implement a ML based parameterization in CESM that will capture sub-grid spatial heterogeneity’s impact on heat and momentum exchange at the air-sea interface. This work will be complementary to, but not overlap with, ongoing LEAP efforts incorporating wave effects and leveraging in-situ and satellite observations for improving pointwise bulk formulae of air-sea flux. The air-sea interface, where the Earth's atmosphere meets the ocean, is very dynamic and critical for the coupled climate system. This interface plays a central role in regulating long term climate, shaping weather, and maintaining marine ecosystems through the exchange of heat, momentum and gases. These exchanges are a result of an intricate and coupled interplay between oceanic and atmospheric flows, and significant variability across scales in these flows poses a challenge for accurate modeling of these exchanges. Research over the last two decades has highlighted the significance of heterogeneity in the air-sea exchange, and the impact this has on the boundary layers, surface flows, winds, clouds and rainfall. These small-scale impacts also cascade to larger scales, influencing crucial aspects of large-scale atmospheric and oceanic circulation, including coupled phenomena like ENSO. Recent advances in understanding small-scale coupled processes have revealed that the current generation of climate models do not adequately capture the large scale impacts of air-sea interface heterogeneity, suggesting that improvements in parameterizations are needed for better physical fidelity. This project will pioneer a parameterization for air-sea exchange of heat and momentum that captures heterogeneity at scales below 100-200km using ML.

**How will the research deliver innovations to climate science? Explain the methodology and the datasets. (2-3 paragraphs)**

The work in this project can be broadly divided into three phases, each providing tangible benefits to climate science and the broader community. The first phase of the work will include the curation of high resolution coupled simulation data, and the subsequent creation of the dataset containing filtered fields and sub-grid fluxes that will be used for training ML models. We will work with data from two coupled simulations, a low-fidelity (LF) and a high-fidelity (HF) dataset. The LF data will be from CESM1.1 run providing data for O(100years) with a resolution of ~10km in the ocean and ~25km in the atmosphere. This simulation resolves the major scales of heterogeneity at the air-sea interface but only marginally resolves the coupled interactions that ensue at the finest scales. The HF data will be from a MITgcm-GEOS simulation run for O(1year) at NASA JPL as part of the DYAMOND2 effort, which has a resolution of ~3km in the ocean and ~6km in the atmosphere. This simulation is able to resolve a larger range of scales of heterogeneity and also the associated coupled flow mechanisms. Much of the CESM1.1 data variables required for this project have already been moved to the LEAP-Pangeo cloud storage, and the PI plans to move the MITgcm-GEOS data to the LEAP-Pangeo cloud storage over the next few months using his NASA supercomputer allotment. The next step in this phase is filtering the state variables (using <https://gcm-filters.readthedocs.io/>) and calculating filtered (large-scale) and sub-grid air-sea fluxes (using <https://github.com/xgcm/aerobulk-python>) to generate training data for ML, this will be done over a range of spatial filtering scales. Some of the pipelines for generating this training data using CESM1.1 have already been setup and tested on the LEAP hub (as part of a paper that the PI is leading along with Dr. Julius Busecke), and this experience will be used for developing pipelines to work with the MITgcm-GEOS data.

The second phase of the project will involve doing a statistical analysis to understand the spatio-temporal patterns of the sub-grid air-sea fluxes and using this experience to train ML models that learn the appropriate relationships between the large-scale fields and the sub-grid air-sea fluxes. To the best of our knowledge, there are no studies that quantify even the mean and variance of the air-sea fluxes that correspond to sub-grid heterogeneity, as this is a relatively new research area and a lot of the effort over the last two decades has been on developing a mechanistic understanding of the associated coupled processes rather than parameterizations for the net effects regardless of mechanism. Thus, our statistical analysis will provide the first global (high-resolution model based) metrics quantifying the effect of sub-grid heterogeneity on air-sea fluxes, which could be used to evaluate aspects of future global climate models and compared against observational datasets. In this analysis, we will also quantify the impact that different components of the heterogeneity (purely due to atmospheric processes, purely due to oceanic processes, and resulting due to coupled processes) have on the air-sea flux, using a technique recently developed by the PI. This work will also lay the groundwork for discovering the appropriate associations between the large-scale fields and the sub-grid fluxes, which will help guide the choice of input variables and design of ML architecture. Also, the ML design choices will be guided based on the PI’s experience developing and implementing ML parameterizations for MOM6 (CESM’s ocean component) and discussions with the NCAR collaborators about the nuance of the CESM coupler, which connects atmospheric and oceanic components of climate models. These ML parameterizations will be first tested for offline skill, and since, currently no parameterizations for sub-grid heterogeneity’s impact on air-sea exchange exists, this work will result in important and much needed scientific discoveries, which may also help guide development of conventional parameterizations in the future.

The third, and most important, phase of this project will involve the implementation of the ML model in CESM and testing its performance in idealized atmosphere-ocean coupled simulations. During this phase we will identify and tailor appropriate idealized atmosphere-ocean coupled CESM setups, which will be used for testing and fine tuning of our parameterizations. Examples of such idealized setups include aqua and ridge planet simulations run in CESM by Wu et al 2021 (JAMES) and coupled channel simulations that have been used to study the impact of ocean heterogeneity on mid-latitude storm tracks (Foussard et al 2019, J.Clim.). The exact details for the implementation will be decided in discussions with Wayne Chuang as we approach the middle of phase two, as the technology for implementing ML parameterizations in climate models is rapidly evolving and because the details of the approach would have to be tailored according to the architecture that is developed in the second phase. During this third phase we anticipate going through a few development cycles that reach back into the later stages of the ML model development that was done in phase two, to ensure model stability and generalizability. Finally, we will evaluate the impact and success of our parameterizations by comparing against the large scale properties of high resolution idealized simulations. We will run experiments where the inputs going to the coupler in these high resolution idealized simulations are filtered, and compare if adding our parameterization to these filtered runs helps the model stay closer to the control high resolution simulation. This approach will ensure that we can isolate the impact that the small-scale heterogeneity has in the simulations, and verify if our parameterizations can represent them.

**What is your strategy to bring your project to level 3 readiness? Specify how this will improve CESM. Explain how you will interface with the CESM community; i.e., please identify who your CESM contact is that will help move the project along. (1 paragraph)**

This proposal is directly addressing the task of developing and implementing a new parametrization for the impact of subgrid heterogeneity on the air-sea fluxes in CESM, which is likely to improve the physical fidelity of the next generation of climate simulations. In the development, implementation and testing of the parameterization, we will be supported by Dr. Justin Small and Dr. Frank Bryan (agreement to be unfunded collaborators through personal communication) from NCAR, who bring a wealth of experience in developing and running coupled and ocean only models over many decades. Also, the filtered and sub-grid flux datasets generated here using high resolution coupled simulations can help in evaluation of future coarse resolution simulations and guide observational efforts. While running fully coupled global ESM simulations are beyond the scope of this project, we hope that successful demonstration of our parameterization in only atmosphere-ocean coupled simulations would lead to wider adoption by modeling centers for future CMIP simulations.

**How will it either use state-of-the-art ML or advance fundamental ML technology? (1-2 paragraphs)**

Since the main goal of this project is to implement a new parameterization into CESM, the foremost guiding principle for the choice of ML will be based on practicality of implementation. Recent experience has shown that physics-informed small neural networks or equations discovered using ML are two viable options, as they ensure manageable computational cost on CPUs and allow the model developer to choose appropriate numerical implementations. Also, the equation discovery techniques allow for greater interpretability and novel physical insights. We will explore both these approaches in this project.

In addition to these pragmatic architectural choices, we plan to ensure that a few key elements are represented in our parameterizations. First, we will ensure that the model is scale-aware by training on datasets that are generated across a variety of filter scales. Second, since we will be curating datasets from simulations that have been run at different grid resolutions (multi-fidelity data), we will use techniques being established in the multi-fidelity surrogate modeling literature to train models that can leverage information from these disparate data sources. Finally, we expect that stochasticity would be an important element for these parameterizations based on our physical knowledge of air-sea interactions, so we plan to train ML models capable of representing this stochasticity, by predicting large scale flow dependent distributions of sub-grid forcing rather than just the deterministic means.

**How will your project help forge a transdisciplinary convergence between climate science and data science? (1 paragraph)**

One of the primary aims of this project is to train a PhD student at the rare intersection of climate science and machine learning, with an eye towards climate model development. The PI is also listed as a potential mentor on Columbia’s Bridge-to-PhD program (INSPIRE Geoscience), and will leverage aspects of this project to provide a post-baccalaureate student hands on experience in doing research at the intersection of climate science and data science. Also, if funded, the PI and student would naturally be active participants of LEAP’s parameterization development focus group, hoping to have two-way-learning interactions with other LEAP members involved in improving CESM.

**How will you evaluate success? Explain your anticipated project outcomes, metrics and/or datasets, if different from Q2. (1 paragraph)**

We anticipate tangible elements to result from each phase of the project. The first phase will result in two types of datasets: (i) High resolution coupled model dataset will be made publicly available for the wider community through the LEAP-Pangeo cloud storage resources, (ii) The data generated for our ML training will be made available publicly in the form of a benchmark dataset (similar to ClimateBench or OceanBench) along with some baseline models, also using LEAP-Pangeo cloud resources and reproducible methods according to LEAP best practices. The second and third phase will result in three papers. (i) Paper documenting the impact that sub-grid heterogeneity (decomposed into contributions from atmosphere-only, ocean-only, and coupled heterogeneity) has on air-sea fluxes and how this changes as a function of scale. (ii) Paper describing how the impact of sub-grid heterogeneity (both deterministic and stochastic components) can be functionally related to the large scale flows using a ML model, (iii) Paper describing the successful implementation and testing of the ML parameterization in atmosphere-ocean coupled CESM simulations.

**How will you integrate research with DEI, education and KT activities? (1 paragraph)**

This project will support the training of a PhD student in the Applied Physics and Applied Mathematics (APAM) department at Columbia University - Prani Nalluri, who has been part of the PIs research group since summer 2023. The student is of South Asian origin and identifies as queer, which are not groups that have been traditionally well represented in climate sciences. Also, this project can support REU students, Bridge-to-PhD fellows, or LEAP momentum fellows each summer, further extending the educational and DEI potential of the project. Finally all the datasets and workflows will be shared publicly through LEAP-Pangeo cloud resources and github, which will ensure broad dissemination of the research and allow the wider community to easily extend on this work or adopt pipelines to other tasks, like downscaling.

**Please describe quantifiable milestones in your project that can be evaluated every 6 months.**

* + Milestone 1: Curation of high resolution coupled simulations on to the LEAP hub. (made publicly available through the LEAP cloud storage resources)
  + Milestone 2: Generation of training dataset at different resolutions and making this available publicly (as a data challenge).
  + Milestone 3: Descriptive spatio-temporal analysis of sub-grid fluxes, linking to physics of air-sea interactions. (comes with paper submission)
  + Milestone 4: Training of a ML model, particularly with an eye towards implementation, scale-awareness, multi-fidelity learning, and stochasticity. (comes with paper submission)
  + Milestone 5: Implementation into CESM, based on the type of ML model identified in previous milestones. (leading to code contribution to CESM)
  + Milestone 6: Proof of concept runs showing evaluation and improvements in idealized coupled CESM simulations. (comes with paper submission)

**LEAP has funding to support research experiences for undergraduates (REU). Could your project offer REU in summer 2024? If so, please briefly describe what will be the research experience, who will be the faculty research mentor(s), what will be the anticipated learning/research outcomes for the REU, and what would be desirable preparation/prerequisites for a prospective undergraduate researcher for your REU project. (3-5 sentences)**

Yes, this project can support REU students. The mentors will be Dhruv Balwada (PI) and Prani Nalluri (PhD student). There are many parts of the project that can be tailored for teaching, while allowing for an introduction to research. Since the summer of 2024 will be during the phase of data curation, the REU students will use filtering and visualization tools from the Pangeo stack to explore the datasets with a focus on quantifying the variability at different spatio-temporal scales. Motivated students could further expand on this by applying dimensionality reduction techniques. Some past experience with basic coding, physics, mathematics or statistics would be desirable, and aspects specific to climate data science will be learnt at part of the REU.

**What programming language(s) will you use? How will you make your code and data open source and ensure that your results are easy to reproduce? (3-5 sentences)**

The analysis of high and low resolution simulations, generation of training data from high resolution simulations, and training of ML models will be done using Python, relying particularly on tools developed through Pangeo and ML libraries like JAX and PyTorch (in which the PI has extensive experience). These data analysis, data generation, and ML training pipe line will be developed on the LEAP-Pangeo platform and shared using github repos (as the PI has done for many of his recent papers). We anticipate continuing the PI’s long history of collaborating on both science and software with Dr. Julius Busecke in this phase, who is developing the best open science practices for the LEAP community. Simulation of idealized experiments, implementation and testing of the parameterization will be done in CESM, which is based on FORTRAN. This will be disseminated via contributions to the CESM code repo on github.

**Secondary information:**

* **If you are not part of LEAP yet, who invited you to submit?**
  + Pierre Gentine and Laure Zanna
* **What is the readiness level of your project?**
  + Level 1: initial proofs of concept and/or high-risk high-reward ideas
* **What are you submitting to?**
  + CESM track

**Budget justification:**

**Sr. Personnel**

2.5 months of salary per year is requested for PI Balwada (rate of $12046.56 per month). Balwada’s major contributions will be on the technical side of the project, which would include porting to and generation of datasets on the LEAP cloud, development of data analysis ML training pipelines that work with massive high resolution climate datasets, and the implementation of the ML model in the CESM coupler. Additionally, Balwada will advise and mentor a graduate student throughout their PhD, along the way helping them to learn the tools of climate data science required for scientific analysis of large simulations, how to run and modify aspects of coupled simulations in CESM, methods of scientific machine learning and parameterization development, and how to disseminate their science through publications and other open science practices. Also, Balwada will lead the model implementation paper.

As a Lamont Assistant Research Professor at Columbia University, Balwada’s time and salary is supported almost entirely from extramural funds.

**Graduate student**

Three years of support are requested for a graduate student, Prani Nalluri, in the Applied Physics and Applied Mathematics department at Columbia University. Apart from learning technical aspects by working closely with the PI, the student will work on the spatio-temporal analysis and ML model development that are part of the second phase of the project, and will help with the design, running and scientific evaluation of idealized coupled simulations with and without the ML parameterizations. Also, the student would lead the first two papers describing the spatio-temporal aspects of surface heterogeneity and functional relationships between sub-grid heterogeneity and large scale aspects of the flow.

**Fringe benefits**

Fringe benefits are calculated at a rate of 29.3%, which is provided by Columbia University.