Received: 5 July 2023 | Accepted: 6 October 2023
DOI: 10.1002/aaai.12160

SPECIAL TOPIC ARTICLE

AI2ES: The NSF AI Institute for Research on Trustworthy AI for Weather, Climate, and Coastal Oceanography

Amy McGovern[1,2] | **Imme Ebert-Uphoff**[3] | **Elizabeth A. Barnes**[4] | **Ann Bostrom**[5] | **Mariana G. Cains**[6] | **Phillip Davis**[7] | **Julie L. Demuth**[6] | **Dimitrios I. Diochnos**[2] | **Andrew H. Fagg**[2] | **Philippe Tissot**[8] | **John K. Williams**[9] | **Christopher D. Wirz**[6]

[1]School of Meteorology, University of Oklahoma, Norman, Oklahoma, USA
[2]School of Computer Science, University of Oklahoma, Norman, Oklahoma, USA
[3]Electrical and Computer Engineering & Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, Colorado, USA
[4]Department of Atmospheric Science, Colorado State University, Fort Collins, Colorado, USA
[5]Evans School of Public Policy and Governance, University of Washington, Seattle, Washington, USA
[6]National Center for Atmospheric Research, Boulder, Colorado, USA
[7]Del Mar College, Corpus Christi, Texas, USA
[8]Conrad Blucher Institute, Texas A&M University—Corpus Christi, Corpus Christi, Texas, USA
[9]The Weather Company, IBM Business, Armonk, New York, USA

Correspondence
Amy McGovern, School of Meteorology, University of Oklahoma, Norman, OK, USA.
Email: amcgovern@ou.edu

Funding information
Directorate for Geosciences, Grant/Award Number: ICER-2019758

Abstract

The NSF AI Institute for Research on Trustworthy AI in Weather, Climate, and Coastal Oceanography (AI2ES) focuses on creating trustworthy AI for a variety of environmental and Earth science phenomena. AI2ES includes leading experts from AI, atmospheric and ocean science, risk communication, and education, who work synergistically to develop and test trustworthy AI methods that transform our understanding and prediction of the environment. Trust is a social phenomenon, and our integration of risk communication research across AI2ES activities provides an empirical foundation for developing user-informed, trustworthy AI. AI2ES also features activities to broaden participation and for workforce development that are fully integrated with AI2ES research on trustworthy AI, environmental science, and risk communication.

INTRODUCTION TO AI2ES

The NSF AI Institute for Research on Trustworthy AI in Weather, Climate, and Coastal Oceanography (AI2ES) is a convergent center focused on AI for the Earth and environmental sciences (ES) (McGovern et al. 2022). We are developing novel AI methods for real-world high-impact environmental use cases that ensure that we address the entire chain of relevant issues.

AI2ES has **two primary goals**. First, we are advancing the state-of-the-art of foundational research in AI, ES, and RC. Second, we are advancing understanding and prediction of weather phenomena to improve societal resilience to climate change and save lives and property. To achieve

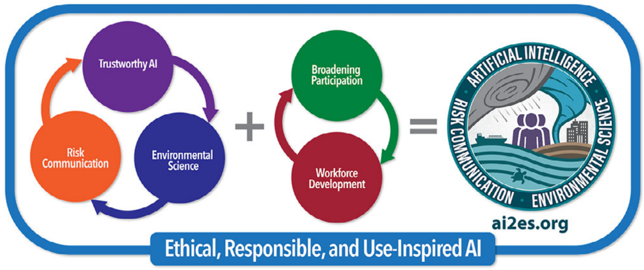

FIGURE 1 Underlying research structure of AI2ES, highlighting the synergistic cycle of AI, ES, and RC research. Our commitment to ethical, responsible, and use-inspired research underlies all that we do.

these goals, we address **three key topics**: (1) Ensuring that the developed AI methods are trustworthy[1] and meet the needs of diverse groups of end-users; (2) developing novel AI that revolutionizes our understanding and prediction of high-impact environmental phenomena; and (3) creating new educational pathways to develop a larger and more diverse AI/ES workforce.

Creating trustworthy AI for ES requires AI2ES to address **several key challenges**. First, many AI for ES applications directly impact lives and property. Furthermore, the AI methods are to be used in time-critical situations. The phenomena of interest are often rare, leading to a sparsity of labeled training data; the available data are multiscale and heterogeneous. ES applications provide an additional challenge that does not occur in many other AI applications: the need for AI models to be physically grounded while capturing complex nonlinear relationships. Finally, we must identify what is needed for the AI to be deemed trustworthy by a diverse set of end-user groups and create the necessary methods to achieve this.

The institute's work is structured as outlined in Figure 1. Shown on the left is our key synergistic cycle that connects foundational research in three areas: AI, ES, and RC. Likewise, our AI workforce development and broadening participation efforts (shown in the middle) are also synergistic and build on our foundational research. Underlying all that we do is a focus on ensuring that our AI is ethical, responsible, and use-inspired.

AI2ES RESEARCH

Given the overview nature of this paper, we highlight key research in each of the three areas of the synergistic cycle from Figure 1.

Generative AI: Generative AI, in particular attention-based models, such as transformers, are enabling the development of powerful AI-driven global weather forecasting models. These AI models execute up to 1000× faster than the numerical weather prediction (NWP) models currently used operationally for weather forecasting, and are starting to reach the resolution and accuracy of NWP models. However, pixel-based accuracy is not a sufficient measure for the usefulness of these models for real-world weather forecasting. Do these models properly predict the key features of storm fronts and tropical cyclones? Do they capture the extremes of heat waves and precipitation? Are the predicted fields consistent? Answering these questions requires expertise in both generative AI, but and meteorology, a perfect task for AI2ES. We are working with private industry and NOAA to evaluate such models and provide feedback to developers (Ebert-Uphoff and Hilburn 2023).

Robust AI: We are developing a set of robust methods for learning models from imperfect data. One aspiration is to address the *class imbalance* and *rare event* problems, including characterizing the sample sizes needed for achieving certain guarantees in performance. In Diochnos and Trafalis (2021), we show that any learning algorithm that generates a *probably approximately correct (PAC)* model can be extended to learn a model that also has high recall and high precision, while maintaining the efficiency of the original algorithm.

We are also working to ensure that our AI methods are *robust to noise and missing data*. In Flansburg and Diochnos (2022), we show that L_1-regularization can be an effective defense mechanism for regression models that are subject to certain training-time attacks, complementing a property of L_1-regularization known for classification models under test-time attacks. We are currently exploring *semi-supervised* learning to build-in robustness to labels that are noisy or that are missing altogether.

Explainable and interpretable AI (XAI/IAI): Explaining predictions of AI models is one aspect of trustworthiness for end-users. The need to peer inside the black box of AI and what the end-users need to see of what the AI model is predicting is context dependent (Wirz et al.). This includes visualizing input–output relationships on data and other aspects of explanation such as case studies and failure modes (Cains et al. 2023).

Most XAI methods were not developed for ES domains and understanding their applicability to the highly spatially and temporally auto-correlated data prevalent in ES domains needs to be investigated. Furthermore, most XAI methods related to image-based tasks yield so-called *attention maps* that indicate where in an input image, a neural network paid the most attention. However, different XAI methods yield greatly differing attention maps. Lastly, the resulting attention maps are usually interpreted visually by AI developers for clues on the strategies the neural network is using, introducing a large potential for subjectivity in interpretation. To address all of these aspects, we developed two synthetic benchmarks (Mamalakis, Barnes, and Ebert-Uphoff 2022; Mamalakis, Ebert-Uphoff, and Barnes 2022) representative of Earth science processes that provide ground truth not only for the neural network prediction, but also for the corresponding attention maps. Using these benchmarks, we identified key characteristics of XAI methods for attention maps. This allowed us to identify which XAI algorithms are most suitable to address certain science questions and how their differing results should be evaluated (Mamalakis, Barnes, and Ebert-Uphoff 2022; Mamalakis, Ebert-Uphoff, and Barnes 2022).

Furthermore, while *post hoc* XAI methods focus on the regions of the input that were most relevant to the network's prediction, they do not tell *how* the prediction was made (Rudin 2019). Our team is building a suite of interpretable neural network architectures that mimic the way scientists interpret weather and climate patterns, allowing the full decision making process to be tracked from start to finish. One example of our IAI networks uses prototypical samples (i.e., "this looks like that") from the training input to classify new samples at inference. However, for weather and climate prediction, the global locations of atmospheric and oceanic phenomena are vital in understanding and predicting their downstream impacts on weather and climate extremes. Thus, our modified IAI network encodes the spatial location of the prototypes as well, that is, "this looks like that there" (Barnes et al. 2022). In this context, we are also currently developing an explanation method for image classification, where we do not allow overlapping regions of an image to be plausible explanations for different classes.

Uncertainty quantification (UQ): Given that many environmental applications involve life-and-death decision making, integrating uncertainty quantification into AI models that support the decision making process is crucial. Our focus on UQ includes understanding the limitations of existing methods for UQ for ES, developing novel methods to represent uncertainty, and approaches for communicating and visualizing uncertainty for and with end-users. For example, in Haynes et al. (2023), we review six different UQ approaches for neural networks—from simple approaches to Bayesian neural networks—and apply them to two environmental case studies. We also highlight four different ways to evaluate such uncertainty estimates, and use the case studies to illustrate how to use them to identify which UQ methods yield reliable estimates.

Responsible and ethical AI for ES: The use of AI is growing exponentially across society, as well as within the sciences. With this use, comes an increased understanding of the need for ethical and responsible development and deployment of AI for many use-cases. However, many AI developers for ES do not see how the issues in the news with AI affected their work on AI for ES applications. We published an in-depth series of examples demonstrating how AI can go wrong when naively applied to ES problems (McGovern et al. 2022).

In our current work, we are focusing on the issue of bias in AI models, specifically examining it for ES applications (McGovern et al. 2023). While bias is not the only issue that needs to be addressed for ethical and responsible use of AI, it is a key issue. We have developed a classification of bias modeled after the more general one discussed in NIST's AI risk management work. In our current work, we are diving deeply into our four main bias categories and showing how AI developers can address and manage these risks for ES applications.

Use inspired research in ES: Grounding our research: All of our work is use-inspired by phenomena in atmospheric sciences and coastal oceanography. Our use cases include: (1) convective hazards, meaning those that are associated with strong thunderstorms, including wind, hail, lightning, and tornadoes; (2) winter weather, including visibility, snowfall, and freezing rain; (3) coastal phenomena, including fog, forecasting cold stunning events to save sea turtles, and understanding harmful algal blooms; (4) tropical cyclones; (5) subseasonal to seasonal prediction of diverse high-impact phenomena including excessive rainfall events.

Risk communication: A crucial but often neglected topic is how fundamental advances in AI for predictions of ES hazards can be developed to provide information that is needed, trusted, and used by professionals, scientists, and expert decision makers, such as weather forecasters, transportation officials, emergency managers, and natural resource managers. Treating AI models and XAI as forms of risk information, the RC team leads convergent, interdisciplinary, multimethod research across AI2ES on trustworthy AI. Initial studies have included ES use cases such as severe hail, storm mode, and coastal fog, with a focus on forecasters' assessments of potential use and trustworthiness of AI-based guidance, and how factors such as model verification and the ability to interact with the model output influence those assessments. Drawing on research from fields as diverse as risk communication and management, organizational and social psychology, human–AI teaming, and trust in automation, the team is contributing methodological insights, as well as fundamental insights into the contextual and subjective nature of trusting and assessments of trustworthiness.

EDUCATION, WORKFORCE DEVELOPMENT, AND OUTREACH

Our efforts to broaden participation and workforce development are strongly connected. We overview key parts here.

Developing an AI certificate program at Del Mar community college: Our core components in the area of education include close collaboration with a Hispanic Serving Institution (HSI) community college. Del Mar College (DMC) is a 2-year community and technical college located in Corpus Christi Texas and leads the AI2ES effort to bring AI and machine learning technologies to the local workforce. The research team of four educators at DMC has led the design and creation of a new Occupation Skills Award degree program, consisting of five courses. Three of the five courses are new for the Award and deal with AI in general and machine learning algorithms for GIS technology specifically. The courses have all been taught to three cohorts of learners, leading to two cohorts of graduates, most of whom transferred to our partner Texas A&M University—Corpus Christi (TAMU-CC) to continue their research and education in the ES.

Core diversity efforts: Both DMC and TAMU-CC are HSIs, with major minority student populations. DMC has made significant efforts to recruit both women and minorities into their AI program. The college has participated in numerous public recruitment efforts revolving around public events including annual GIS Day, Earth/Bay Day, and Hurricane Conference events. To create a sustainable pipeline of secondary students from high school to the college, annual summer STEM camps have focused on local middle and high school students, with an emphasis on minority girls. Research has noted the need for outreach and recruitment in the middle school years for girls, since they form their life goals and plans much earlier than boys of the same age. The college has completed two successful summer bootcamps, with 2023 adding returning campers to expand upon their technology exposure and reinforce their bonds with DMC.

The DMC activities, some in collaboration with TAMU-CC students and faculty, are the foundation of an AI2ES student pipeline. Students enrolled in the new DMC GeoAI classes and other computer science classes are introduced, or reintroduced, to AI2ES opportunities. The close collaboration between DMC and TAMU-CC faculty facilitates this bridge. While the two organizations are in the same city, the type of students enrolling at each institution is quite different due to financial constraints and cultural differences, including not being familiar with the opportunities, career paths, and financial support possible through higher education. Pairing a community college and a university with overlapping programs has been a very productive practice to recruit a broader range of students and launch them on an AI career for AI2ES. Community College students are more likely to be first generation, underrepresented minority, and not be aware of STEM opportunities or their own potential. At present, seven DMC students have been hired as undergraduate research assistants at TAMU-CC while still being enrolled at DMC. The context of a large AI institute has been invaluable for these students. The biweekly site-wide meetings, and particularly participating in the AI2ES first live meeting and presenting or co-presenting at the American Meteorological Society conferences (15 student presentations in 2023) have opened the eyes of many of these students to broader possibilities and will help diversify our field. They now dream bigger.

Internships and collaboration with industry: AI2ES industry partners comprise an Industry Advisory Board that offers their perspective on how the institute's research can help address important problems in the private sector. Industry collaborators are involved in many aspects of the Institute's research areas, participate as mentors for student research projects, and offer summer internships. These activities help prepare students for the workforce while also catalyzing the transition of AI2ES research into operations, thereby broadening the Institute's impact and enhancing its service to society.

EXAMPLE SUCCESS STORIES: TEXT BOXES

Uncertainty quantification for tropical cyclone intensity and track forecasts: Uncertainty quantification and communication can be incredibly challenging. We have explored a simple method for adding uncertainty to almost any neural network regression task via estimation of a general probability distribution (Barnes, Barnes, and DeMaria 2023). We showed that this intuitive approach can improve current tropical cyclone forecasts of intensity, as well as their track, by adding uncertainty estimates to an otherwise deterministic prediction. This approach and product is currently being tested at the National Hurricane Center.

Advancing the conceptualization of trustworthiness: Drawing on trust-related literature across multiple disciplines and fields, we have synthesized knowledge on interpersonal trust, trust and risk perceptions, and trust in automation (Bostrom et al. 2023; Wirz et al.). This synthesis of trust theory, along with our ongoing empirical research, informs our (re)conceptualization of trustworthiness as being in practice a user's subjective assessment for a specific situation, which may be affected by time pressure and decision stakes, even when an AI/ML model has been developed in accordance with trustworthiness standards. The resulting potential misalignment between developers and users is analogous to mismatches found historically between lay and expert risk assessments of other technologies, such as nuclear power. The AI2ES risk communication team is co-producing these syntheses and studies that build on them with AI/ML developers and environmental scientists to advance the evaluation and treatment of AI/ML trustworthiness in the ES.

CONCLUSIONS

AI2ES is leading the development of AI models for weather and climate applications. The foundational methods developed by AI2ES will revolutionize our ability to predict, understand, and communicate a variety of high-impact weather hazards.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. ICER-2019758.

CONFLICT OF INTEREST STATEMENT

The authors declare that there is no conflict.

ORCID

Amy McGovern https://orcid.org/0000-0001-6675-7119

ENDNOTE

[1] AI2ES Definition of Trustworthiness: Trustworthiness is a trustee's evaluation, or perception, of whether, when, why, or to what degree someone or something should or should not be trusted.

REFERENCES

Barnes, E. A., R. J. Barnes, and M. DeMaria. 2023. "Sinh-Arcsinh-Normal Distributions to Add Uncertainty to Neural Network Regression Tasks: Applications to Tropical Cyclone Intensity Forecasts." *EDS* 2: e15.

Barnes, E. A., R. J. Barnes, Z. K. Martin, and J. K. Rader. 2022. "This Looks Like That There: Interpretable Neural Networks for Image Tasks When Location Matters." *AIES* 1(3): e220001.

Bostrom, A., J. L. Demuth, C. D. Wirz, M. G. Cains, A. Schumacher, D. Madlambayan, A. S. Bansal, et al. 2023. "Trust and Trustworthy Artificial Intelligence: A Research Agenda for AI in the Environmental Sciences." *Risk Analysis*: 1–16. https://doi.org/10.1111/risa.14245

Cains, M. G., C. D. Wirz, J. L. Demuth, A. Bostrom, A. McGovern, I. Ebert-Uphoff, D. J. Gagne, A. Burke, and R. Sobash. 2023. "Exploring what AI/ML Guidance Features NWS Forecasters Deem Trustworthy." In *103nd AMS Annual Meeting*. AMS.

Diochnos, D. I., and T. B. Trafalis. 2021. "Learning Reliable Rules under Class Imbalance." In *SDM*, 28–36.

Ebert-Uphoff, I., and K. Hilburn. 2023. "The Outlook for AI Weather Prediction." *Nature* 619: 473–74.

Flansburg, C., and D. I. Diochnos. 2022. "Wind Prediction under Random Data Corruption (Student Abstract)." In *AAAI*, 12945–46.

Haynes, K., R. Lagerquist, M. McGraw, K. Musgrave, and I. Ebert-Uphoff. 2023. "Creating and Evaluating Uncertainty Estimates with Neural Networks for Environmental-Science Applications." *AIES* 2: 1–58.

Mamalakis, A., E. A. Barnes, and I. Ebert-Uphoff. 2022. "Investigating the Fidelity of Explainable Artificial Intelligence Methods for Applications of Convolutional Neural Networks in Geoscience." *AIES* 1(4): e220012.

Mamalakis, A., I. Ebert-Uphoff, and E. A. Barnes. 2022. "Neural Network Attribution Methods for Problems in Geoscience: A Novel Synthetic Benchmark Dataset." *EDS* 1: e8.

McGovern, A., A. Bostrom, M. McGraw, R. J. Chase, D. J. Gagne, I. Ebert-Uphoff, K. D. Musgrave, and A. Schumacher. 2024. "Identifying and Categorizing Bias in AI/ML for Earth Sciences." *BAMS*. https://doi.org/10.1175/BAMS-D-23-0196.1, in press.

McGovern, A., A. Bostrom, P. Davis, J. L. Demuth, I. Ebert-Uphoff, R. He, J. Hickey, et al. 2022. "NSF AI Institute for Research on Trustworthy AI in Weather, Climate, and Coastal Oceanography (AI2ES)." *BAMS* 103(7): E1658–68.

McGovern, A., I. Ebert-Uphoff, D. J. Gagne, and A. Bostrom. 2022. "Why We Need to Focus on Developing Ethical, Responsible, and Trustworthy Artificial Intelligence Approaches for Environmental Science." *EDS* 1: e6.

Rudin, C. 2019. "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead." *Nature Machine Intelligence* 1(5): 206–15.

Wirz, C. D., J. L. Demuth, A. Bostrom, M. G. Cains, I. Ebert-Uphoff, D. J. Gagne, A. Schumacher, A. McGovern, and D. Madlambayan. "(Re)Conceptualizing Trustworthy AI as Perceptual and Context-Dependent: A Foundation for Change."

How to cite this article: McGovern, A., I. Ebert-Uphoff, E. A. Barnes, A. Bostrom, M. G. Cains, P. Davis, J. L. Demuth, D. I. Diochnos, A. H. Fagg, P. Tissot, J. K. Williams, and C. D. Wirz. 2024. "AI2ES: The NSF AI Institute for Research on Trustworthy AI for Weather, Climate, and Coastal Oceanography." *AI Magazine* 45: 105–10. https://doi.org/10.1002/aaai.12160

AUTHOR BIOGRAPHIES

Amy McGovern is a Lloyd G. and Joyce Austin Presidential Professor in both the School of Meteorology and School of Computer Science at the University of Oklahoma.

Imme Ebert-Uphoff is a Research Professor in Electrical and Computer Engineering and the Machine Learning lead at the Cooperative Institute for Research in the Atmosphere, both at Colorado State University.

Elizabeth A. Barnes is a Professor in the Department of Atmospheric Science at Colorado State University.

Ann Bostrom is the Weyerhaeuser endowed Professor in environmental policy in the Evans School of Public Policy & Governance at the University of Washington.

Mariana G. Cains is a Research Scientist at the National Center for Atmospheric Research.

Phillip Davis is a Professor at Del Mar College.

Julie L. Demuth is a Research Scientist at the National Center for Atmospheric Research.

Dimitrios I. Diochnos is an Assistant Professor in the School of Computer Science at the University of Oklahoma.

Andrew H. Fagg is a Brian E. and Sandra O'Brien Presidential Professor and an Associate Professor in the School of Computer Science at the University of Oklahoma.

Philippe Tissot is the Conrad Blucher Institute Chair for Coastal Artificial Intelligence at Texas A&M University-Corpus Christi.

John K. Williams is a Senior Technical Staff Member and Head of Weather AI Sciences at The Weather Company, an IBM Business.

Christopher D. Wirz is a Research Scientist at the National Center for Atmospheric Research.