horizon RESEARCH, INC.

2022 Trustworthy Artificial Intelligence for Environmental Science (TAI4ES) Summer School Feedback

Laura Craven
Jessica A. Dula

September 2022

Submitted to: Dr. Amy McGovern, PI
University of Oklahoma
660 Parrington Oval
Norman, OK 73019

Submitted by: Horizon Research, Inc.
326 Cloister Court
Chapel Hill, NC 27514

TABLE OF CONTENTS

Page

Introduction 1
Participants and Patterns of Participation 3
Participant Opinions of Summer School 8
Impacts and Next Steps 18
Evolution of the Summer School 23
Summary 24

INTRODUCTION

In June 2022, the NSF AI Institute for Research on Trustworthy AI in Weather, Climate, and Coastal Oceanography (AI2ES), in collaboration with the National Center for Atmospheric Research (NCAR), held the second Trustworthy Artificial Intelligence for Environmental Science (TAI4ES) Summer School. The Summer School was advertised to a worldwide audience and was held entirely online over four days. Registrants came from 64 countries (see Figure 1) and represent a range of education levels (see Figure 2).

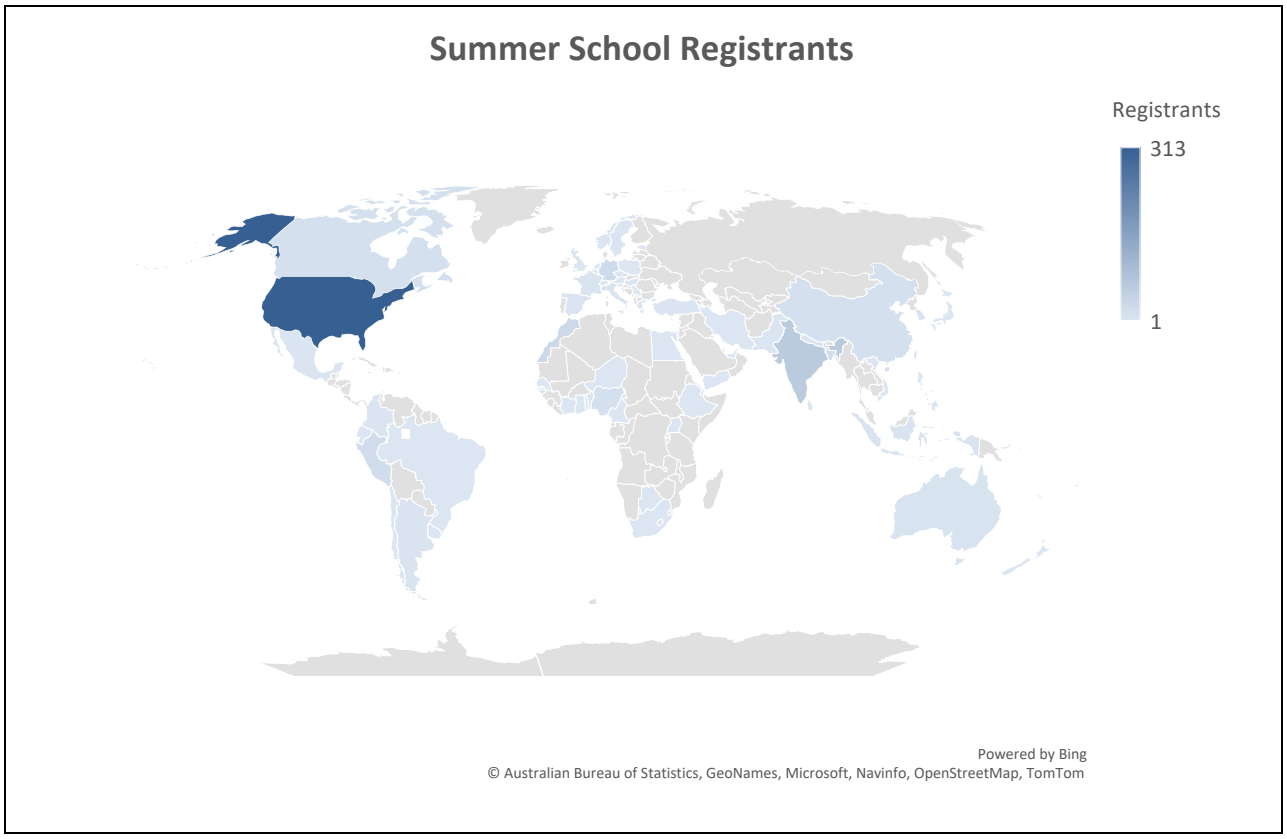


Figure 1

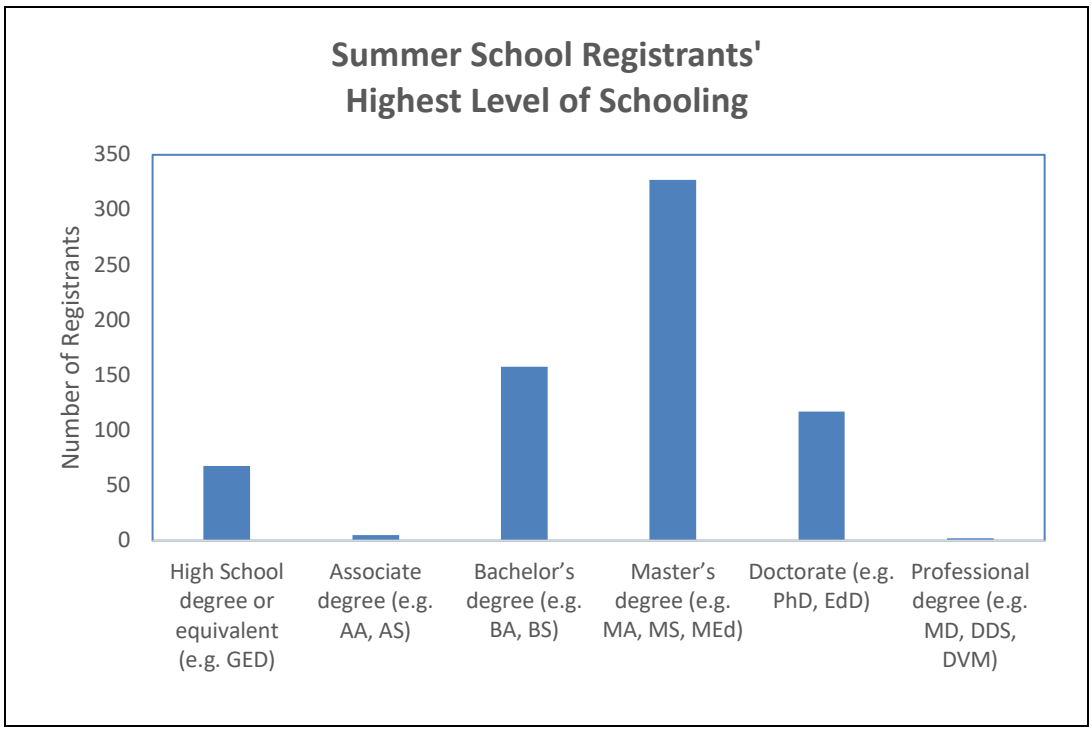


Figure 2

Sessions consisted of lectures from and discussions with leading researchers in trustworthy AI for environmental sciences. Lectures covered topics including the foundations of trustworthiness for AI, Explainable AI (XAI), Interpretable AI, and how machine learning systems have been developed for a range of environmental science applications. The Summer School also offered a Trust-a-Thon for a subset of attendees (available on a first-come, first-served basis) to gain hands-on experience with evaluating the trustworthiness of pre-trained machine learning algorithms to solve real-world environmental science challenges. This experience took place in the afternoons following the lecture portion of the Summer School and consisted of three data sets (severe weather, space weather, and tropical cyclones) with a set of personas (short descriptions of an end user that would be using the model participants develop) randomly assigned to small groups of participants.

The goals of the Summer School were for attendees to:

1. Learn about the nature of trust and trustworthiness in AI as well as evaluation metrics for evaluating trust.
2. Learn about explainability and interpretability, how users think about these concepts, and how to use attribution maps.
3. Understand the importance of trustworthy data and workflows as well as the benefits of using case studies.
4. Learn about the uncertainty lifecycle, metrics to evaluate uncertainty, and strategies for communicating uncertainty to different audiences.

As part of the AI2ES project evaluation, Horizon Research, Inc. (HRI) observed the Summer School (both the lectures and Trust-a-Thon), invited all participants to complete a survey, and interviewed a sample of participants. HRI also analyzed registration data collected by NCAR.

The survey was developed to assess participants' perceptions of the quality of the experience, impacts, and next steps. All registrants (N = 711) were invited by email to complete the survey after the Summer School ended. Of those invited, 282 responded, a response rate of 40 percent, and of those, 96 individuals attended the Trust-a-Thon. As discussed below, the respondents were representative of the total group of registrants, but the results should still be interpreted with caution due to the possibility of non-response bias.

The survey included a question asking respondents if they were willing to be interviewed. From those who agreed, a representative sample of 16 participants were invited to participate in the interview.[1] The interview addressed the same topics as the survey but in greater depth.

The purpose of this memo is to provide feedback to help AI2ES leaders reflect on the Summer School. The first section provides data about participants and patterns of participation. The second section focuses on participant opinions of the Summer School lectures, Trust-a-Thon, and the event overall. The third section discusses participant impacts and next steps, and the fourth describes the evolution of the Summer School from 2021 until now. The memo concludes with a summary and recommendations.

PARTICIPANTS AND PATTERNS OF PARTICIPATION

On the registration form, participants were asked how they heard about the Summer School. As seen in Table 1, 39 percent of respondents reported hearing about the event from a friend or colleague. Emails to the AI2ES list and social media were also common methods (20 and 17 percent, respectively). Other email lists (15 percent) and announcements on the AI2ES website (14 percent) and NCAR (12 percent) website were less common.

[1] Of the 16 participants initially contacted, 9 agreed to participate, 1 declined, and 6 did not respond. Participants were initially contacted by email, and two follow-up emails were sent to those who did not respond. Six backups were then contacted and none of them responded.

Table 1
How Participants Heard About the Summer School

	Percent of Registrants (N = 711)
Friend or colleague	39
Email to the AI2ES list	20
Social media	17
Other email	15
Announcement on the AI2ES website	14
Announcement on the NCAR website	12
Other website	5

† Percentages add to more than 100 because respondents could choose more than one option.

Table 2 shows the location, race/ethnicity, and gender of Summer School registrants and those who completed the survey. Overall, the characteristics of the survey respondents are similar to those of the registrants. Over a third of survey respondents originated from North America, followed by a quarter from Asia. Forty-four percent of attendees identified as Asian, 22 percent as White, 11 percent as Black/African American, and 9 percent as Latinx or Hispanic. In terms of gender, 67 percent identified as men, 31 percent as women, and 1 percent as non-binary/gender non-conforming (the remainder preferring not to answer).

Table 2
Registrant and Respondent Characteristics

	Percent of Respondents	
	Registration (N = 711)	**Survey (N=282)**
Location		
North America	51	37
Asia	19	26
Europe	13	16
Africa	10	13
South America	6	6
Australia	1	2
Race/Ethnicity[†]		
Asian/South Asian/Southeast Asian/East Asian	41	44
White or Caucasian	27	22
Black/African/African American/Afro-Caribbean	8	11
Latino/a/e/x and/or Hispanic	11	9
Middle Eastern or North African	6	6
Multi-Racial	3	1
American Indian/Native American/Alaska Native	1	1
Another race or ethnic origin	1	0
Native Hawaiian or Other Pacific Islander	0	0
Prefer not to answer	8	9
Gender[†]		
Man	62	67
Woman	34	31
Non-binary/gender non-confirming	1	1
Transgender	0	0
Prefer not to answer	3	2

[†] Percentages add to more than 100 because respondents could choose more than one category.

Table 3 shows the position, institution, and AI2ES affiliation of survey respondents. Sixty-one percent were graduate students, 14 percent were research staff or scientists, and 11 percent were undergraduate students. Postdocs, university faculty, non-research staff (e.g., software engineers), and others made up around 15 percent of participants. A large majority of attendees did not work at an institute of higher education (IHE) in the US. Among those who did work at an IHE in the US, the vast majority worked at a non-minority serving institution. Almost all participants were either not affiliated with AI2ES or were unsure about their affiliation.

Table 3
Attendee Roles

	Percent of Respondents	
	Registration (N = 711)	**Survey (N=282)**
Position		
Graduate Student	60	61
Research Staff or Scientist	12	14
Undergraduate Student	13	11
Postdoc	7	6
University Faculty	4	4
Non-Research Staff	2	2
Other	2	2
Institution†‡		
Not applicable, I do not work at an IHE in the USA		78
Non-minority-serving institution (i.e., serving predominantly White students)		10
Asian American and Native American Pacific Islander-Serving Institution		4
Hispanic-Serving Institution		4
Tribal College or University		3
Historically Black College or University (HBCU)		1
Alaska Native or Native Hawaiian Serving Institution		0
Native American-Serving Nontribal Institution		0
AI2ES Affiliation‡		
No		86
Yes		6
Not sure		8

† Percentages add to more than 100 because respondents could choose more than one category.
‡ Item was not included on the registration form.

Looking at the AI experience of the Summer School registrants, 17 percent had no experience with AI, and just over half reported only having some experience (see Table 4). Two-thirds had either a master's degree or a doctorate.

Table 4
Expertise of Registrants

	Percent of Registrants (N = 711)
AI Experience	
No experience	17
Some experience	54
Quite a bit of experience	22
Professional	7
Highest Degree	
High School degree or equivalent (e.g., GED)	9
Associate degree (e.g., AA, AS)	1
Bachelor's degree (e.g., BA, BS)	22
Master's degree (e.g., MA, MS, MEd)	46
Doctorate (e.g., PhD, EdD)	22
Professional degree (e.g., MD, DDS, DVM)	0

Participants were asked which lecture sessions they attended each day. As seen in Table 5, attendance appeared to wane as the week went on. Almost all lecture sessions on Monday or Tuesday were attended by at least two-thirds of respondents. Most lecture sessions on Wednesday and Thursday were attended by less than 60 percent of respondents.

Table 5
Participation in Summer School Lecture Sessions

	Percent of Respondents (N = 249)
Monday	
Evaluation metrics for classification and regression	81
Insights on (meaningful) interdisciplinary work in the AI/ML development process	80
What does it mean to trust? An interdisciplinary perspective	78
XAI for traditional machine learning	77
Tuesday	
Explainability vs Interpretability	73
XAI Techniques for deep learning (Part 1)	69
XAI Techniques for deep learning (Part 2)	63
Wednesday	
Trustworthiness of data and implications for workflows from collection through to designing analyses	64
The importance of case studies and tips for using them effectively	58
Thursday	
Uncertainty quantification methods (Part 1)	59
Uncertainty quantification methods (Part 2)	55
Communicating uncertainty	53

HRI used attendance data, along with the Summer School agenda, to calculate the amount of time each participant spent attending lectures. Attendees were categorized in five levels according to their extent of participation in the synchronous sessions[2] (see Table 6). Just over one-quarter of respondents reported attending every lecture session on every day, and another 18 percent attended more than 8 hours of lecture sessions. Thirteen percent of respondents indicated they did not attend any sessions. Common reasons for not attending the live sessions included not being in the same time zone and unplanned work or personal conflicts.

Table 6
Overall Levels of Participation in Lectures

	Percent of Respondents (N=282)
Did not attend any sessions	13
Light (attended less than 4 hours)	19
Moderate (attended between 4 hours and 8 hours)	23
Heavy (attended more than 8 hours)	18
Attended all sessions, all days	27

PARTICIPANT OPINIONS OF SUMMER SCHOOL

Lectures

The Summer School survey included several items asking participants about the quality of their experience in the lecture portion of the event. As can be seen in Table 7, over 85 percent of respondents moderately or strongly agreed that the speakers were well prepared, the event was interesting, the Summer School reflected careful planning and organization, time was used effectively, the event was a good use of respondents' time, participants were encouraged to ask questions, and the goals were made clear. Responses related to the few remaining aspects of the Summer School were not as positive, though still generally complimentary. It is worth noting that 25 percent of respondents disagreed to some degree that the Summer School included opportunities for them to make connections with other attendees.

[2] This time includes only the sessions listed in Table 5. It does not include any introductions or breaks.

Table 7
Participants Opinions of Summer School Quality

	Percent of Respondents (N = 249)					
	Strongly Disagree	**Moderately Disagree**	**Slightly Disagree**	**Slightly Agree**	**Moderately Agree**	**Strongly Agree**
The Summer School lectures were well prepared.	0	0	1	7	39	52
The Summer School lectures were interesting.	1	1	1	12	36	49
The Summer School lectures reflected careful planning and organization.	0	1	1	7	43	47
Time was used efficiently during the Summer School lectures.	0	2	3	13	42	41
The Summer School lectures were a good use of my time.	1	1	3	12	43	40
The Summer School lectures encouraged me to ask questions and participate in discussions during the live (or synchronous) Q&A sessions.	0	2	3	13	42	39
The goals of the Summer School lectures were made clear.	0	2	2	5	53	37
The Summer School lectures were relevant to my own work.	1	2	3	24	40	31
The social science sections of the Summer School lectures were very helpful.	0	3	6	19	43	28
The Summer School lectures included sufficient opportunities to reflect on what I was learning.	0	2	5	20	48	24
The Summer School lectures included opportunities for me to make new connections with other attendees.	4	6	15	29	27	20

Like the survey respondents, interviewees also had positive opinions of the quality of the lecture portion of the Summer School. Many described aspects of the lectures that they found particularly useful, both related to the content of the lectures and the format by which they were presented. Many indicated that they liked learning about the various applications of machine learning systems. One explained how these applications were directly related to their own analysis:

> *I really enjoyed the lectures, and especially some of the things that I'm using for my analysis, . . . some of the charts that they introduced were very cool. Although, I would say maybe they introduced it in a more advanced level, but it challenged and it pushed me like, "Oh, yeah, that's cool stuff I can use."*

Interviewees also described aspects of the lecture format they found useful, including question-and-answer opportunities, the lecture organization, and the selection of speakers. In the words of two:

> *The polls I really liked because the polls kind of helped us to interact. Interaction is good because it requires concentration. It basically creates focus. . . . It's not possible to bring everybody on board, hundreds of attendees on board, and have them talk to each other, or have them pose a question orally.*

> *[There was a slide deck of] 100 and something slides, and it was divided [among] different people, so it was easy for me to navigate. Maybe you don't get bored with someone just talking about the same thing, but it was divided in two different people and each one has his own role. Those are the things I like most about the lecture structure.*

The survey also asked about the pace of the Summer School. As can be seen in Table 8, more than two-thirds of respondents indicated that the pace was appropriate, and just over a quarter indicated it was somewhat too fast. Only 2 percent rated the pace as much too fast. Some interviewees described the lectures as "intense" with a high "density of the material," though they acknowledged the difficulty of covering the intended material with such a large audience any other way.

> *It's really hard with professionals and lots of people anyway, but it would be nice to have it in smaller chunks of time. It was really intense.*

Table 8
Participants' Opinions of the Summer School Pace

	Percent of Respondents (N = 249)
Much too slow	0
Somewhat too slow	1
Appropriate	68
Somewhat too fast	28
Much too fast	2

In interviews, summer school participants were asked to describe what aspects of the lecture they would change. Responses varied and described changes to the question-and-answer structure (i.e., Slido), faster delivery of session recordings, more inclusion of numerical data in addition in imagery data, and information about how end users and researchers collaborate with one another. The following quotes are representative:

They have talked at length about what the end users want, but then having some case studies where the end users wanted something, and the researchers, they were looking for something. How did two things get married? How did two things get paired up? That thing actually never showed up in the format. Having for each day, where if there is a detailed case study of what the researchers did great, amazing work. After that, okay, what were the end users actually thinking and how the researcher actually changed, would be really instrumental in this thing.

The one thing I really didn't like is the built-in question-and-answer system because you would have to like, even if you liked a question, you would have to pay attention to see if someone answered it. Because sometimes they would just disappear and you'd be like, "Oh, I'm really interested in that." And by the time, because you're still listening to all the lectures, you go back to see, "Well, what's the answer, or what did people say about it?" They would be gone and then you would never know.

Trust-a-Thon

About a quarter of the Summer School participants had the opportunity to participate in a Trust-a-Thon to gain hands-on experience with evaluating the trustworthiness of AI methods to solve real-world environmental science challenges.

Trust-a-Thon participants who were interviewed indicated that they spent the majority of their experience working with code in a Jupyter notebook. They then responded to a blog post about how to explain their model to the end user described in their persona, as one interview described:

We had [personas] given to us at the beginning, and so we were supposed to use this persona you have to explain to them, portray to them, the details of the notebooks. So, the notebook has training. So whatever training that is done within the notebook, you would have, let's say, a graph visualization at the end, and then you're supposed to interpret this visualization to the user persona you had.

The Summer School survey included several items asking participants about the quality of the Trust-a-Thon experience. As can be seen in Table 9, over 75 percent of respondents moderately or strongly agreed that the Trust-a-Thon was interesting, it was a good use of their time, and there were sufficient opportunities to reflect on what they were learning. Responses related to other aspects of the Trust-a-Thon were also quite positive. It is worth noting that between 12 and 16 percent of respondents disagreed to some extent that the Trust-a-Thon included opportunities for them to make connections with other attendees, that the team approach worked well for them, and that the goals were made clear.

Table 9
Participants Opinions of Trust-a-Thon Quality

	Percent of Respondents (N = 95†)					
	Strongly Disagree	**Moderately Disagree**	**Slightly Disagree**	**Slightly Agree**	**Moderately Agree**	**Strongly Agree**
The Trust-a-Thon was interesting.	0	2	3	16	43	36
The Trust-a-Thon was a good use of my time.	0	3	1	9	53	34
The user personas helped me think about what was important for the end users.	1	3	5	17	38	35
The Trust-a-Thon included sufficient opportunities to reflect on what I was learning.	0	1	5	17	48	29
The Trust-a-Thon activities reflected careful planning and organization.	0	2	5	18	46	28
The Trust-a-Thon was accessible to me at my current knowledge level.	0	2	3	23	43	28
The Trust-a-Thon included opportunities for me to make new connections with other attendees.	0	1	11	22	40	26
The Trust-a-Thon was relevant to my own work.	0	3	4	24	45	23
The Trust-a-Thon team approach worked well for me.	1	5	10	23	39	21
The goals of the Trust-a-Thon were made clear.	0	1	12	15	52	20

† Only those who reported participating in the Trust-a-Thon saw these questions.

Interviewees also had positive opinions of the quality of the Trust-a-Thon. Most commonly, interviewees liked the technology platforms that were used for this portion of the event. They appreciated the ability to access code on GitHub, communicate with their team and the Trust-a-Thon organizers on Slack, and access all necessary resources on the event website. In the words of one:

> *I think [the technology platforms] were used very well, I must say, because most of the code we found directly on GitHub, which was really key. And then Slack was good for the communication because most of the channel-based details, we could all reach it at once and instantly, which is really nice. . . . The website was also really good because it was intuitive to find the content you're looking for. And also, there was like support for the times when we were having issues with accessing resources.*

Interviewees also particularly liked working with code in the Jupyter Notebook and the real-life context of the problems they were working to solve. As two interviewees explain:

I really like the functional code. . . . And I liked it because I think it's a useful thing to see a functional code function. So, you can think with it and see what breaks it and see how it's changed it. That's, uh, it's a good way to function; examples are good way to learn.

I also like the idea of the personas because it gave you an example, a user that you're trying to provide this model for. To build this model for, let's say, someone working in a metrology center or something like that, and this is a person going to be using the model that you've built. So having this understanding of how to relay information to them was an amazing thing, that I think a key takeaway that we all would've gotten from the sessions altogether.

The survey also asked about the pace of the Trust-a-Thon. As can be seen in Table 10, about two-thirds of respondents indicated that the pace was appropriate, and just over a quarter indicated it was somewhat too fast.

Table 10
Participants' Opinions of the Trust-a-Thon Pace

	Percent of Respondents (N = 95)
Much too slow	0
Somewhat too slow	7
Appropriate	64
Somewhat too fast	27
Much too fast	1

† Only those who reported participating in the Trust-a-Thon saw these questions.

In interviews, Trust-a-Thon participants were asked to describe the aspects of the experience that they would change. The most common response was to have more control over which data set and persona they were assigned to work with. In one participants' words:

When they grouped us, they didn't ask us about our preferences. . . . They didn't ask, "Where would you want? What project do you want to partake in? What are you interested to investigate for this Trust-a-Thon?" So, at some point, for myself, I was thinking maybe I would've done better in the tropical [group] because it deals more on what I'm currently looking at.

Several interviewees also mentioned that they would change some aspect of the small group structure or how their team communicated with one another. More specifically, participants indicated that many of their team members did not interact with them on Slack and it became challenging for them to get feedback on their work and ideas. Participants suggested that larger

group sizes or more purposeful group assignments (e.g., grouping based on interests or background) could help address this challenge.

> *The only thing lacking I felt was team coordination because you could not always get people to speak about a topic or interact. You couldn't participate fully because you didn't know what other team members were thinking. I couldn't have written a blog on my own. It was supposed to be a team effort.*

> *Especially on the last day, I was feeling like, "I'm just speaking to myself. Even if I have a challenge, nobody's responding to me." Yeah, I think those are some of the things maybe we can look into if we are planning for another one. Give people a chance to select [their group]. And I also think that can even improve the participation.*

> *I think having a slightly larger group size would be helpful. Or a larger, effective group size, because two of the group members, we didn't get a chance to interact with them. We basically ended up being three of us.*

Other aspects of the Trust-a-Thon that interviewees would change include the nature of the asynchronous work, the naming convention for the Jupyter notebooks to make it clearer which one they should use each day, and the technology platforms that were used.

Summer School Overall

In interviews, participants were asked to describe the connections they saw between the lecture portion of the Summer School and their experience with the Trust-a-Thon. Most interviewees (7 out of 9) noted the content alignment between the two sections. Three of 9 interviewees also appreciated the ability to get hands-on experience with the lecture content. In the words of one:

> *Each project was basically a subset of whatever was being taught in the lectures. So, the Trust-a-Thon helped to actually reinforce the learning, what we had from the lectures. The lecture covered the theory part. The Trust-a-Thon showed us how to employ that theory into practice.*

To gather additional information about participants' perceptions of the quality of the Summer School, the survey included an open-ended question asking which aspects of the experience were most useful. HRI coded their responses in several categories. As can be seen in Table 11, respondents provided a variety of responses, which included both specific lecture topics and structural aspects of the Summer School. Of the 144 individuals who provided a response, 54 percent mentioned at least one aspect of the lectures as most useful. Twenty-two percent specifically mentioned the sessions related to XAI methods, and 11 percent mentioned the sessions related to quantifying and communicating uncertainty. In the words of one:

I really enjoyed the lectures, especially those about XAI and uncertainty quantification because they gave me a global view of both fields.

Additionally, 29 percent of respondents pointed to an aspect of the Summer School structure as being most helpful. These included the well-organized nature of the Summer School, the interactive elements (i.e., polls, Q&A, Slido), the wide range of topics, the speaker selection, and the balance between practical applications and theoretical discussions. As two explained:

I thought for most of the lectures, the pace and style was really well done. There was a really good mix of technical methods, practical application, and 'sit-back-and-synthesize' commentary.

I like the fact that the hybrid meeting platform was much easier to use this year, and although a bit fast-paced, the talks were nicely distributed across various topics.

Table 11
Most Useful Aspects of the Summer School†

	Percent of Respondents (N = 144)
Lectures	**54**
XAI methods	22
Quantifying and communicating uncertainty	11
Explainability vs. Interpretability	5
Applications of AI (case studies)	4
Introduction to Trustworthy AI	3
Interpreting AI for End Users	3
Summer School Structure	**29**
Easy to attend/ accessible/ well organized	13
Interactive elements (polls, Q&A, Slido)	12
Wide range of topics	6
Selected speakers	3
Good balance among practical and theoretical	2
Trust-a-Thon	**15±**
Resources shared by speakers	**8**
Communication between speakers and participants	**4**

† Percentages add to more than 100 because respondents may have mentioned more than one category.
± This underrepresents the percentage identifying the Trust-a-Thon as the most useful aspect because only a subset of the 144 respondents participated in it, whereas all respondents had an opportunity to participate in lectures.

Like the survey respondents, interviewees noted various aspects of the Summer School they felt had a particular impact on their learning, including both the lectures and the Trust-a-Thon. Specifically, interviewees thought that the lectures provided them with new perspectives and information and could see direct applications to their own work. As two explained:

The entire summer school had an impact on my learning because the lectures were clear, and I think they were meticulously organized, . . . and most of what I learned from the summer school, I learned from the lectures.

It is not possible to give an in-depth learning in a short period of time, but their training is tremendous, and I'm totally fascinated by the kind of breadth they provide. . . . So that is what I feel is the major contribution of summer school, to make you aware that these kinds of ideas are there.

The survey also included an open-ended question asking participants how the Summer School could have been improved. Of the 99 respondents who commented, 38 percent suggested improvements to the content (see Table 12). The most common request of this type, given by 13 percent of respondents, was for more practical examples or applications. Seven percent wished for less advanced/technical lectures, and 5 percent wanted more or different sessions to be offered. In the words of one:

For people with a little or no preliminary knowledge, some basics may be added. Practical or hands-on sessions are more expected.

In addition, 32 percent suggested improvements to the structure of the Summer School. Eighteen percent suggested improvements to the timing or pace, 10 percent requested more interaction among the participants or with the speakers, and 5 percent suggested improvements to the Q&A sessions or Slido. As one explained:

Time the information better for the time slots allotted (so you don't have to go so fast in sections or skip things) [and] have a place where questions and answers can live somewhere less temporary (it is hard to keep checking to see if questions get answers while mid-lessons and then sometimes the questions were cleared before I had a chance to see answers at all).

Twenty-six percent of respondents provided suggestions related to the Trust-a-Thon. Eleven percent mentioned improving group collaboration, and 7 percent wanted the number of people allowed to participate to be increased. Other suggestions, each mentioned by 3 percent or fewer of respondents were various suggestions for technical issue fixes, having more instructions available, and having a team facilitator. One respondent commented:

Encouraging direct engagement among the team members. . . . I feel in my team we were not able to ever experience the cross-domain expertise collaboration, which I was looking forward to in the summer school.

Finally, 14 percent suggested improvements to the Summer School logistics—more specifically, holding the Summer School at a different time or in person, making the session materials available sooner,[3] and creating Slack channels for all participants, each mentioned by 5 percent or fewer respondents.

Table 12
Suggestions for Improvement†

	Percent of Respondents (N = 99)
Improvements to Summer School Content	**38**
More practical examples or applications	13
Less advanced/technical lectures	7
More/different sessions	5
More hands-on activities	3
More specialized details included in presentations	3
Inclusion of pre-course material	2
Improvements to Summer School Structure	**32**
Improvements to timing/pace/breaks	18
More interactive (among participants/with speakers)	10
Improvements to Q&A/Slido	5
Longer duration	2
Improvements to the Trust-a-Thon	**26**
Improve group collaboration	11‡
Increase number of participants	7
Suggestions for tech issue fixes	3‡
Have more instructions available	2‡
Have a team facilitator	2‡
Improvements to Logistics	**14**
Hold at a different time	5
Hold in person	5
Make session materials available sooner	2
Create Slack channels for all participants	2

† Percentages add to more than 100 because respondents may have mentioned more than one category.
‡ This underrepresents the percentage because only a subset of the respondents participated in the Trust-a-Thon.

A couple of interviewees mentioned aspects of the structure of the Summer School they would change. These broader suggestions include ideas like having an information session about

[3] Summer School leaders sent all participants links to session videos and presentation materials. Some participants completed this survey prior to receiving the materials.

available opportunities in the AI for ES field and making sure that the research shared has been published or is available to participants. The following quotes are illustrative:

> *They can have an information session at the end. Maybe . . . they can say, okay, what kind of early career scientist, what kind of opportunities they have. Or they can give information about what kind of opportunities [are] available in AI for environmental sciences domain, presently, in other organizations, because they would know. They are the pioneers in this field.*

> *[There was] at least one time where they had tables and data from articles that they used to support their points. And the articles are unpublished and not accessible to anybody in the classroom. . . . I get it, but it's mildly annoying to have that be presented as justification and then refute it as well. So maybe refine the sources that are used to prove the points. . . . Because if a student is going to take these concepts and spread them and push them forward, we need to be able to show our resources or our references.*

IMPACTS AND NEXT STEPS

HRI used responses to survey items to describe impacts the Summer School may have had on participants. Interview responses also helped illustrate the nature of these impacts. In the survey, participants rated their understanding of certain lecture topics before and after the Summer School, using a retrospective-pre/post approach[4] (see Table 13). The statements align with the Summer School organizers' goals for participants. Prior the lecture portion of the Summer School, only 13–36 percent of participants felt that they had "good" or "strong" understanding of any of the topics. These percentages rose considerably to between 68 and 86 percent after the Summer School.

[4] The retrospective pre approach is appropriate when respondents may not have realized how much they did not know about a topic before an event.

Table 13
Participants Indicating Good or Strong Understanding† of the Lecture Topics

	Percent of Respondents (N = 249)	
	Retrospective Pre	**Post**
What trustworthy AI means	22	86
The importance of thinking about end users	36	83
The importance of trustworthy data and workflows	24	79
Available techniques for XAI	15	72
The difference between explainable AI (XAI) and interpretable AI	13	69
How to communicate uncertainty	22	68
Uncertainty quantification methods	20	68
How to pick case studies	17	68

† Includes those indicating "Good" or "Strong" understanding on a four-point scale from 1 "None" to 4 "Strong."

Participants were also asked to rate their understanding of Trust-a-Thon topics before and after the Summer School, again using a retrospective-pre/post approach (see Table 14). Prior the Trust-a-Thon, no more than 18 percent of participants felt that they had good or strong understanding of any of the topics. After the Trust-a-Thon, at least 59 percent or more participants reported these levels of understanding of the topics.

Table 14
Participants Indicating Good or Strong Understanding† of the Trust-a-Thon Topics

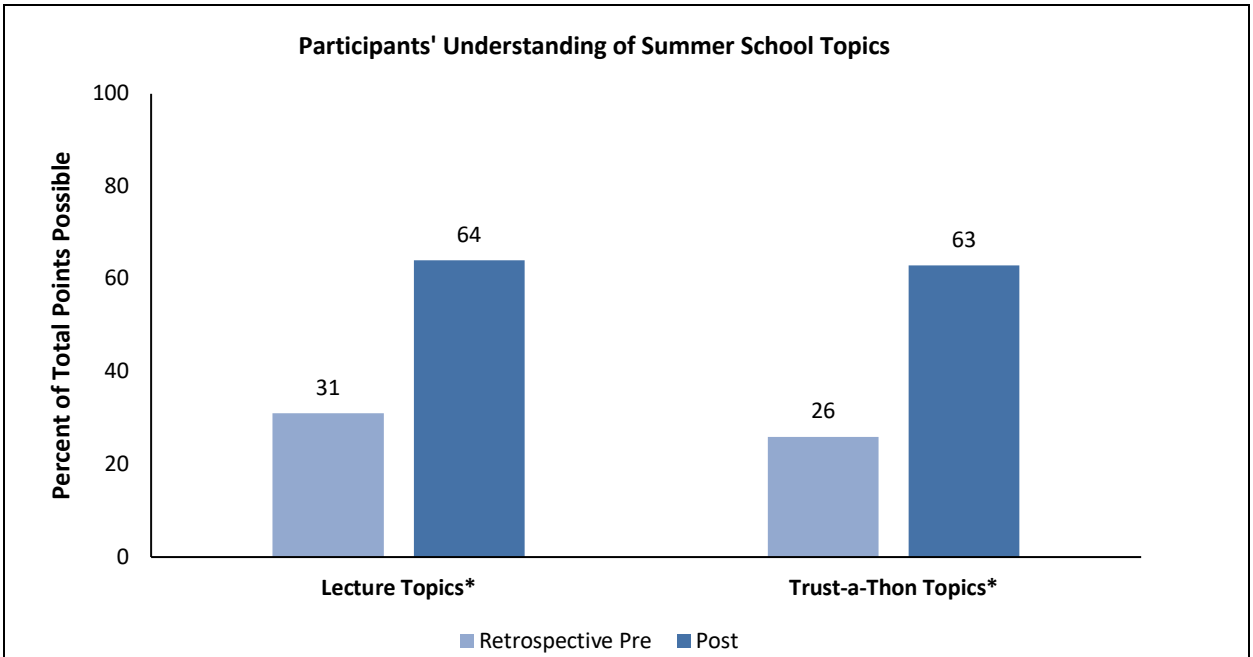
	Percent of Respondents (N = 95)	
	Retrospective Pre	**Post**
The importance of thinking about the end user when developing AI	16	84
How to assess AI trustworthiness	13	73
How to implement machine learning systems that have been developed for a range of environmental science applications	16	72
How to implement explanatory AI (XAI) methods	18	69
How to implement ethical AI methods	16	59

† Includes those indicating "Good" or "Strong" understanding on a four-point scale from 1 "None" to 4 "Strong."

To test for changes in understanding before and after the Summer School, items in the tables above were combined to create two corresponding composite variables.[5] As displayed in Figure 3, participants on average felt that they had some understanding of the lecture and Trust-a-Thon topics before the Summer School (mean scores of 31 and 26, respectively) and good

[5] Composite variables have the advantage of being more reliable than individual questionnaire items. The composites were calculated by summing the responses to the relevant items and then dividing by the total points possible. Composite scores can range from 0 to 100 points; someone who marks the lowest point on every item in a composite receives a score of 0, and someone who marks the highest point on every item receives a score of 100.

understanding by the end of the week (mean scores of 64 and 63, respectively). There is a statistically significant difference between participants retrospective-pre and post understanding composite scores. On average, scores after the Summer School were 1.84 standard deviations higher for lecture topics and 1.75 standard deviations higher for Trust-a-Thon topics.[6]



* There is a statistically significant difference between participants' understanding scores before and after the Summer School (two-tailed paired samples t-test, $p < 0.05$).

Figure 3

Interviewees also reported increased understanding in relevant content areas. Specifically, they described learning about (1) AI applications in the context of environmental science, (2) explainable or interpretable AI, and (3) recent advancements in the field. They reported these outcomes as expectations they had for the Summer School that were met, and some specifically described how attending sessions was personally beneficial for them. In the words of one:

[6] Because different studies may use different instruments or report different kinds of scores, it is difficult to compare results across studies. Effect sizes are used to report results that can be more easily compared across studies. In addition, effect sizes take into account the amount of variation in scores, aiding interpretation of results. The effect size for each of these comparisons was calculated as the difference in means of composite scores at each time point divided by the pooled standard deviation. Effect sizes of about 0.20 are typically considered small, 0.50 medium, and 0.80 large. Cohen, J. (1988). Statistical power analysis for the behavioral sciences. Lawrence Erlbaum Associates.

First of all, in the aspect of acquiring more knowledge about environmental science, that was highly achieved because I can say that through the Summer School, I didn't leave as I joined. It was an intensive week. . . . I covered quite a lot through that summer school. . . . And then in the direction of trustworthy AI, I can say that it helped me to see how best I can use to trustworthy AI and not just in the environmental science space, but in my application of AI generally, because I've been using AI for a while, and I didn't see interpretability as a thing to always show people, as a primary thing to show.

Survey participants responded to a list of ways they might use what they learned in the Summer School. As shown in Table 15, 70 percent or more of participants expect to use what they learned for coding AI methods themselves and interpreting output. At least half of those who responded indicated that they expected to use what they learned to apply AI methods others have coded, to develop AI methods to address novel problems, and for more effectively collaborating with others who are coding AI. Around one-fifth reported that they expected to teach others or supervise others who are collaborating or coding AI.

Table 15
Participants Expecting to Use What They Learned in Various Application Areas

	Percent of Respondents (N = 249)
For coding AI methods myself	75
For interpreting output of AI methods	70
To apply AI methods that others have coded	57
To develop AI methods to address novel problems	51
For more effectively collaborating with others who are coding AI	50
To help others interpret output of AI methods	40
For assessing ethical impacts of AI applications	34
To teach my students or others who will be collaborating on AI projects	20
To supervise others who are coding AI	19

Similarly, a majority of interviewees mentioned ways that they would continue to use what the learned in the Summer School. They discussed incorporating AI models and other ideas into their own work and sharing the information they learned with colleagues. As one explained:

I came to know a lot of terms which I had no idea about. Actually, one of the things I've been taught in the summer school, I'm applying right now—how to do permutation feature importance techniques for a ML model, to understand how the ML model is functioning. That is exactly the kind of stuff I'm doing with random forest, right? So definitely there are many things I've come to know which I was not aware.

Survey respondents were also asked to rate the extent to which they would like more information about topics addressed in the Summer School. They recorded their responses on a 5-point scale from 1 (not at all) to 5 (to a great extent). As shown in Table 16, the results align somewhat with ways they plan to use what they learned (see Table 15). Around three-fourths of survey respondents expressed strong interest (i.e., gave a rating of 4 or 5) in learning more about uncertainty quantification methods and how to communicate uncertainty. Interpretable AI, XAI, and trustworthy data and workflows were also topics that about 70 percent of participants wanted more information about. About 60 percent or fewer were strongly interested in learning more about how to pick case studies and trustworthy AI.

Table 16
Participants Wanting More Information About Various Topics

	Percent of Respondents (N = 249)				
	Not at all [1 of 5]	**[2 of 5]**	**Somewhat [3 of 5]**	**[4 of 5]**	**To a great extent [5 of 5]**
Uncertainty quantification methods	2	2	22	25	49
How to communicate uncertainty	2	4	24	25	45
Interpretable AI	2	3	25	26	44
Explainable AI	1	3	29	25	42
Trustworthy data and workflows	4	3	24	30	39
How to pick case studies	5	5	29	26	35
Trustworthy AI	3	6	35	24	32

Survey respondents had the opportunity to describe in their own words what additional resources would help them best use what they learned in the Summer School. The most common types of responses are shown in Table 17. Of those suggested additional resources, 21 percent each said having datasets/coding activities for practicing and examples of code would be helpful.[7] Similarly, a number of participants noted that the session materials[8] (including lecture slides/recordings and Jupyter notebooks) and lists of papers referenced in presentations would be helpful. Others noted that links to other resources (like videos or blogs) and more examples or practical information would be helpful.

[7] In addition, 18 respondents had no suggestions for additional resources, and 12 made suggestions that were not interpretable.

[8] As noted earlier, session materials for the Summer School were made available online after the event, though some survey respondents may have indicated these as resources they would like access to before finding them.

Table 17
Additional Resources Participants Would Like Access To[†]

	Percent of Respondents (N = 71[‡])
Datasets/coding activities for practice	21
Examples of code	21
Session materials (current and previous courses)	18
List of papers referenced in presentations	11
Other resources (videos, blogs)	11
More examples/ practical information	10
Internet/Software	8
Other trainings, tutorials, online courses	6
A way to connect/discuss with others	3
Mentorship	3

[†] Percentages add to more than 100 because respondents may have mentioned more than one category.
[‡] Although more respondents made comments, only 71 were complete enough to be interpretable.

These responses were also echoed by interviewees. They mentioned wanting (1) other courses/tutorials; (2) access to datasets, code, AI models, and cloud computing resources; and (3) ways to form connections and stay in contact with researchers and other Summer School participants. In the words of one:

> *There are so many resources of AI in the open domain, but when it comes to AI for geosciences, you wouldn't find a lot of it. I think AI2ES is very well positioned to post tutorials, videos teaching stuff on AI for geosciences. And they can post this, whether it is on videos, on YouTube, or study materials on their website.*

Finally, interviewees were asked what they thought a subsequent Summer School should focus on. Their responses varied considerably, though a few suggestions fell into broader categories of additional applications and recent advancements in AI for ES. Others wanted more specific topics, like additional information about developing models or guidance on translating data to the general public.

EVOLUTION OF THE SUMMER SCHOOL

AI2ES, in collaboration with NCAR, held a TAI4ES Summer School in the summer of 2021, and HRI collected similar evaluation data about participants' perceptions of the quality of the experience, impacts, and next steps. The two Summer Schools were similar in many ways. They both attended to concepts related to trustworthy AI, XAI, and how machine learning systems have been applied to a wide range of environmental science applications. The participants of both Summer Schools joined these discussions virtually from around the world and had various

levels of background in AI and ES. That said, the two Summer Schools differed in many ways. The content of the presentations was different, as were many of the speakers. Participants of the 2022 Summer School had the added benefit of a Trust-a-Thon, which provided hands-on experience with the ideas discussed in the lectures.

Overall, the location, race/ethnicity, and gender of Summer School registrants were very similar in 2021 and 2022. There were only small fluctuations, such as the number of registrants from Africa increasing from 5 percent to 10 percent and the number of registrants identifying as White or Caucasian decreasing from 31 to 27 percent.[9]

Summer School participants in both years were asked to rate their agreement with several statements related to the quality of the Summer School. Both groups of survey respondents agreed that the quality of the Summer School that they attended was very high overall. For instance, over 90 percent of respondents moderately or strongly agreed that the Summer School they attended reflected careful planning and organization and that the speakers were well prepared. Some minor changes in respondent perceptions of quality include those who strongly agreed that there were opportunities for them to make connections with other attendees appeared to rise from 10 percent to 20 percent and those who strongly agreed that they were encouraged to ask questions and participate in discussions rose from 26 percent to 39 percent.

Similar to those who attended the 2022 Summer School, participants of the 2021 Summer School reported a largely positive and valuable experience in interviews and the post-event survey. They also provided recommendations and constructive criticism, many of which were attended to in the 2022 Summer School. The most prevalent recommendation from 2021 was also the most substantial change in 2022—the addition of the hands-on Trust-a-Thon portion of the Summer School. Past participants requested more hands-on tutorials, more specialized details, examples of code, and examples of ML models they can manipulate. All of these features figured prominently in the Trust-a-Thon. In addition, past participants suggested adding more interactions among other participants, which was addressed with the addition of individual Slack channels for small groups of Trust-a-Thon participants.

SUMMARY

Over 700 people registered for the 2022 TAI4ES Summer School, and 40 percent of them completed a survey in which they described their perspectives on the experience and the impacts

[9] None of the apparent differences in this section were tested for statistical significance and should be interpreted cautiously.

they perceived. They also provided information about themselves, their reasons for attending, and what they would like from similar events in the future. Nine participants completed interviews in which they provided similar information with greater detail.

Survey respondents overwhelmingly agreed with many statements about the Summer School, indicating that it went well and that it was valuable to them. More than 90 percent agreed to some extent with almost all of the statements about the lectures. Participants echoed these positive sentiments when asked to rate their agreement with statements related to the Trust-a-Thon. However, 16 percent disagreed to some extent that the team approach worked well for them. They generally thought the pace of the lectures and the Trust-a-Thon was appropriate, though nearly a third thought they were too fast.

Most respondents also indicated that their understanding of various subject areas, corresponding to the subject areas that event planners aimed to focus on, increased as a result of participating. Most also reported that they intend to use what they learned to code AI methods and interpret output in their own work.

Though most survey respondents reported a largely positive and valuable experience, they also provided helpful recommendations and constructive criticism that may be valuable in planning similar events. Many commented that they would have benefitted from information about how end users and researchers collaborate with one another in the lectures as well as more practical examples or applications. In addition, respondents would have liked more control over the content their Trust-a-Thon group worked with and changes to the small group structure, as well as access to code, AI methods referenced in presentations, and datasets. Slower pace, more interaction among participants, and changes to the Q&A structure were also mentioned.