Intelligent Simulation-based Learning About Natural Disasters (ISLAND)

Importance

Simulations are powerful tools for scientific inquiry into complex Earth systems, enabling investigations otherwise too challenging to conduct in traditional science labs or in classrooms [1], [2], [3], [4]. In natural hazards education, simulations may be the only viable tool for students to build knowledge of hazard systems by creating various scenarios, observing phenomena, exploring causal relationships, and making predictions [5]. This approach addresses the urgent global need to prepare future generations to understand and respond to natural hazards given the rising frequency and intensity of climate-driven disasters and their economic, social, and environmental impacts worldwide [6], [7].

Most students need scaffolding to engage productively in simulation-based inquiry. When learning science through simulations, students must actively define problems, run experiments, collect and analyze simulation data, and interpret results [8]. However, too often students explore only a limited set of parameters, improperly set up initial conditions to recognize important effects, or fail to collect comparable evidence across trials [9]. Our prior research revealed that only 28% of students created the necessary phenomena to answer a question related to inland flooding [1]. Another study revealed that students approached simulation-based inquiry linearly—moving directly from data collection to analysis, interpretation, and answering the question without revisiting simulations to collect new evidence when needed [10]. This linear approach overlooks the iterative nature of inquiry, where students ideally revisit and refine simulations to improve the quality of their evidence. Providing individualized support is needed but difficult for teachers in real world classrooms, especially as students use simulations independently and progress at their own pace [11].

Artificial intelligence (AI) technologies have been used to support simulation-based inquiry through automated feedback to students who did not engage in important simulation use behaviors, such as systematic control of variables [12], [13]. Fostering deeper epistemic engagement in open-ended, simulation-based inquiries goes beyond supporting essential simulation behaviors [14], [15]. Future AI feedback systems should accommodate multiple dynamic paths toward successful inquiry and recognize the diverse ways students express and develop their ideas [15]. Despite the success of machine learning (ML) algorithms, they have raised concerns about potential biases [16] as they may unintentionally overlook or discourage students' unique expressions of disciplinary ideas or actions [17].

This project will design a next-generation automated feedback system to scaffold student inquiry. By deploying large language models (LLMs)—advanced AI systems trained on extensive datasets to understand and generate human language— this project will create a personalized, adaptive automated feedback system for guiding students' simulation-based inquiry. This AI system will interpret nuanced student input, deliver personalized, context-appropriate feedback, and dynamically adjust guidance based on student progress. This design represents a substantial advancement from current LLM-based applications, which are primarily focused on scoring student-generated text [18], [19] and responding to questions through chat interactions [20]. A fully integrated LLM-based feedback system capable of assessing and supporting the entire inquiry process from start to finish would be groundbreaking. Moreover, designing such a system for seamless integration into teachers' daily practices in real-world classrooms will be transformative [21]. This project will generate critical insights for designing LLM-based feedback systems that can (1) be trained to uphold disciplinary standards, (2) systematically scaffold simulation-based inquiry across diverse student demographics, and (3) integrate meaningfully with teachers, who bring valuable contextual insights to classroom implementation.

Goal and Objectives

The Concord Consortium (CC), in collaboration with the American Meteorological Society (AMS) and Physics Front (PF), proposes a five-year, Level III Design and Development project in the Learning strand. WestEd will provide external evaluation on the project. The goal of the project is **to create intelligent simulation-based learning about natural disasters (ISLAND) for middle school students** by leveraging advanced AI technologies—including machine learning, data analytics, and LLMs. To accomplish this goal, the ISLAND project will build upon the knowledge, experience, and products of two NSF-funded projects that developed three nationally distributed natural hazard modules addressing

wildfires, floods, and hurricanes, as well as a text-based automated feedback system developed to support scientific argumentation with real-world data and simulation data (see Results from Prior NSF Support). To enhance students' simulation-based inquiry, the project will develop "Hazbot," so named because it is an AI system (or bot) trained on natural hazards. Hazbot will function as an advanced, personalized automated feedback system featuring a two-tier pedagogical agent design. The student tier will support students' simulation-based scientific inquiry while the teacher tier will examine automated feedback generated by the student-tier and communicate synthesized data about students and insights to teachers.

Objective 1. Develop automated scoring mechanisms to represent the four components of students' simulation-based inquiry (collecting, analyzing, and interpreting simulation data and constructing arguments from evidence) using computer logs, digital artifacts, and text data from extensive datasets gathered during prior implementations of the natural hazard modules.

Objective 2. Conduct design-based research on Hazbot integration to test the functionality of its automated scoring and feedback mechanisms, the usability of Hazbot's student-tier agent in enhancing student inquiry, and the feasibility of the teacher-tier agent in supporting classroom instruction.

Objective 3. Identify effective teaching strategies for utilizing real-time automated feedback provided by Hazbot. Refine the approaches and strategies with nine focus group teachers and integrate the strategies into just-in-time educative curriculum materials (ECMs) for the wildfire, flood, and hurricane modules.

Objective 4. Conduct a pilot study involving a randomized control trial to evaluate the impact of Hazbot-integrated modules on student learning outcomes related to understanding of natural hazards and construction of evidence-based scientific arguments.

Objective 5. Disseminate the Hazbot-integrated modules along with ECMs and share research findings through publications and conferences targeted at both researchers and teachers.

Research Questions (RQs). The project hypothesizes that (1) students, guided by Hazbot and their teachers, will improve simulation-based scientific inquiry, both in the process (collecting, analyzing, and interpreting simulation evidence) and the result (developing scientific arguments based on simulation evidence); (2) students will improve their inquiry practice across multiple tasks; and (3) students' deeper engagements with simulation-based scientific inquiry in the Hazbot-integrated hazard modules will result in significant improvements in their understanding of natural hazards and ability to construct evidence-based scientific arguments related to varying natural hazard scenarios. The research questions below will guide the design iterations of Hazbot (RQ1-RQ3) and a randomized control trial (RQ4):

RQ1. How does Hazbot's automated scoring represent students' simulation-based inquiry?

RQ2. How does Hazbot feedback support students' simulation-based inquiry—such as collecting, analyzing, and interpreting data, and constructing arguments based on that data—within each task and their inquiry development across multiple tasks?

RQ3: What combinations of teacher facilitation and automated feedback are needed to support students' simulation-based scientific inquiry?

RQ4. What is the impact of Hazbot-integrated modules on student learning?

In addressing these RQs, we will analyze data for students as a whole and by specific demographic groups (gender, race, and English as a primary or secondary language).

Theoretical Foundations

Automated feedback functions as scaffolding [22], [23] within digitally constructed learning systems by guiding learners as they tackle an initially challenging task beyond their current knowledge and abilities and adapting support in real time to meet the learners' ever-changing needs towards independent inquiry [13]. Drawing on existing literature, we frame simulation-based inquiry as a disciplinary practice and as a challenging yet accessible form of experiential learning. Following this, we review automated feedback systems in science education, focusing on their instructional benefits for guiding inquiry and their challenges in adapting to diverse student needs and classroom contexts.

Simulation-based Inquiry. Computer simulations have become essential tools in scientists' toolkits [24] and are increasingly valuable resources for classroom use [25]. In educational settings, simulations serve multiple purposes, from aiding conceptual acquisition [9], [26] to motivating student engagement [27], fostering procedural learning [28], and supplementing or even replacing physical labs [29], [30].

Simulations enable users to explore the factors and processes related to large-scale dynamic phenomena that are otherwise inaccessible [31], [32]. They can be used for explaining and predicting natural phenomena [33] and comparing simulated outcomes with real-world phenomena [34]. Simulation-based inquiry with phenomena such as natural hazards involves carrying out experiments with different parameters, boundaries, and starting conditions. When scaffolded, students can conduct controlled variable experiments by changing one parameter at a time with simulations [35] and compare relationships between simulation input and output data to real-world phenomena [36].

Computer simulations can engage students in scientific reasoning [35] and improve their scientific inquiry [37]. From the experiential learning perspective [38], using a simulation to conduct inquiry on a natural hazard serves as an experience that would be otherwise impossible or too dangerous to replicate in a classroom [39]. The simulation, then, serves as the source of knowledge about a phenomenon [40]. Through their experience with the simulation, students develop a sense of how complex phenomena emerge and change over time when variables are adjusted [41]. Students make observations of the simulation and are asked to engage in reflection and analyze outputs. Finally, students are able to develop complex explanations [42] and, in some cases, make predictions [35]. However, simulations do not automatically impart knowledge to students [8]; instead simulations must be used with purpose and scientific intent, and learning with them must be scaffolded to guide students toward meaningful insights. When studying complex Earth systems, providing individualized student guidance is challenging for teachers in real-world classrooms [43], [44].

Automated Feedback. Feedback plays an important role in engaging students in learning [45], [46], helping them correct conceptual and procedural errors [47] and reflect on their performance for deeper engagement with disciplinary knowledge and practices [48]. The impact of feedback is amplified when it is immediate and task-specific [49], [50] and further enhanced when tailored in real time to the learner's performance level [51]. While providing feedback during instruction has traditionally been the teacher's responsibility, advancements in AI technologies now allow digitally supported learning environments to share feedback responsibility [16]. With the advent of machine learning (ML) and natural language processing (NLP) algorithms combined with deep neural networks and empowered by big datasets, automated scoring is now possible for unstructured, student-generated text [52], images and drawings [53], concept mapping [54], and student interactions with digital objects [12], [55].

Advancements in artificial intelligence in education have made it possible to design automated feedback systems for science learning, which has traditionally been more challenging to implement than similar systems for computer programming or mathematics [16]. There is growing recognition of the potential for these systems to deliver personalized, adaptive, real-time guidance [56], with several successful applications already supporting science learning in classrooms [57]. Automated feedback systems enhance both the scientific inquiry process and outcomes [13], [58], while offering particular benefits for English language learners [59], [60]. Research has also shown that AI-enabled systems can improve teaching practices by tracking student progress in real time [61], [62], assessing performance [63], and delivering timely scoring and feedback [64]. Such systems provide valuable insights for teacher decision-making [65] and offer recommendations for instructional strategies [61]. Although gaining traction within the science education community [66], the application of ML-based AI technologies has faced several criticisms [67]. Chief among these is the potential for algorithmic bias, particularly with NLP and ML technologies used to uncover hidden data patterns, as bias can arise from factors such as data selection, annotation processes, input representation, model choice, and misinterpretation based on syntactic and semantic structures [68] and research design [69]. Advocates of automated feedback systems acknowledge these concerns and emphasize the need for empirical testing rather than theoretical dismissal [57]. Additionally, teachers and students can be taught how to assess automated feedback with a critical eye, an essential skill in this new era of AI immersion [70].

Recent developments in large language models might be harnessed to address these issues [71]. Compared to traditional ML-based NLP approaches, an LLM-based approach is expected to be more effective at handling the nuances and variability in student writing. Recent studies [19], [72] suggest that by presenting an LLM with a small number of demonstrative examples, accompanied by well-crafted background task information and a scoring rubric, the LLM can perform effectively as an automated

scoring tool for students' scientific writings [73]. However, while these models excel at individual tasks, there is still significant work needed to enable LLMs to provide cohesive, context-aware feedback across complex, multi-step learning processes. This is why the "in-context learning" (ICL) method [72], [73] of training the LLM is essential in the design of Hazbot since the ICL allows the LLM to be trained not only on task-specific examples enforcing disciplinary requirements but also on an individual student's history of responses, simulation interactions, and learning trajectory. Through ICL, an LLM can build a model of each student's unique learning needs, strengths, and areas for growth. This individualized approach, which moves beyond isolated task scoring, opens up new possibilities for personalized, adaptive feedback that evolves alongside each student's progress. Research is necessary to ensure that the automated feedback system functions as expected across various learner groups.

Results from Prior NSF Support

GeoHazard: Modeling Natural Hazards and Assessing Risks (PI: **Pallant**; Co-PIs: **Lee**, McAuliffe, McDonald, Larson; DRL-1812362; $2.8M; 2019-2023). ***Summary of results:*** The project developed three curriculum modules for wildfires, floods, and hurricanes, as well as simulations, teacher dashboards, educative curriculum materials (ECMs), and assessment instruments that measure students' understanding of hazards and ability to construct evidence-based scientific arguments. Cronbach's alpha values were 0.86, 0. 87, and 0.91 for the wildfire, flood, and hurricane assessment instruments, respectively. Significant pre-post gains were observed (ES = 0.69 SD for hurricanes; ES = 0.54 SD for floods; ES = 0.77 SD wildfires). Research found that (1) there is a significant correlation between students' simulation behaviors and the quality of their arguments [74]; (2) students who revisited and reran simulations produced higher-quality scientific arguments [10]; and (3) 28% of students observed the necessary scientific phenomena required to support their claims with reasoning [1]. ***Intellectual merit:*** The project provided new knowledge on how to integrate simulations into natural hazard curricula and how to support teachers in discussing socio-scientific topics with students; developed new techniques for analyzing student interactions with simulations; and created a pedagogical framework for natural hazards. ***Broader impacts:*** Since public release, the wildfire module has been implemented by 232 teachers and 11,411 students, the hurricane module by 208 teachers and 9,484 students, and the flood module by 127 teachers and 4,593 students. These teachers span all 50 states, showing that hazard instruction is of national interest, not confined to frequently affected areas. ***Publications:*** 2 published papers, 1 submitted paper, and 5 newsletter articles, which are listed under the project name in the References.

Investigating How to Enhance Scientific Argumentation through Automated Feedback in the Context of Two High School Earth Science Curriculum Units (ESAAF) (DRL-1418019; $2.5M; 2014-2019; PI: Liu at Educational Testing Service; Co-PIs: **Lee** and **Pallant**). ***Summary of project results:*** The project used machine learning-based natural language processing algorithms to automate the scoring of students' scientific arguments embedded in two online modules addressing climate change and freshwater availability [75]. The project developed HASbot, an automated feedback system and integrated it into the modules [13]. Research showed that (1) students significantly improved their arguments upon revisions prompted by HASbot [76]; (2) task-specific feedback proved more effective than general feedback across race, gender, and language [77]; (3) the more tasks students revised based on the automated feedback, the greater their gains in their argumentation abilities as measured from pre-test to post-test [58]; (4) discourse analysis indicated that critical thinking about evidence quality emerged most predominantly after receiving automated feedback [14]; (5) while feedback helped students improve their scientific arguments, on average, only 5% of students returned to simulations to redo data collection [13]; (6) with simulation feedback, an average of 35% of students revisited simulations to refine their evidence [13]; and (7) students were frustrated when they received identical feedback despite revising their arguments, when HASbot found their revisions insufficient to achieve a higher score [58]. ***Intellectual merit:*** The project pioneered automated scoring of simulation interactions, text mining, and image processing and developed assessments and rubrics for uncertainty in science argumentation. ***Broader impacts:*** 952 teachers and 46,684 students used the HASbot climate module; 965 teachers and 41,174 students used the HASbot water module without formal training from project staff after the public release. ***Publications:*** 12 published papers, 2 book chapters, and 1 report. See References.

ISLAND PI Lee and Co-PI Gweon have explored students' simulation-mediated inquiry practice as part of the **InquirySpace 2 (IS2): Broadening Access to Integrated Science Practices** (DRL-1621301; $4.5M; 2016-2022; PI: Dorsey; Co-PIs: Damelin, **Lee**, **Gweon**, Tinker) project. ***Summary of project results***: The project developed three high school modules for data-intensive, inquiry-based, independent scientific investigations in physics, chemistry, and biology. ***Intellectual merit:*** The project developed a pedagogical model of experimental inquiry, a new theory on epistemic engagement with scientific experimentation, and an epistemic knowledge instrument, and identified teaching strategies. ***Broader impacts:*** 50 teachers and 3,000 students used IS2 modules. ***Publications***: 7 published papers and one submitted paper. See References. Co-PI Price has experience studying simulated learning environments as part of **Investigating How Museum Experiences Inform Youths' STEM Career Awareness and Interest** (DRL-1906954; $1.2M; 2019-2025; PI**: Price** (2019-2022) due to employer change, PI: Applebaum (2022-2025). ***Summary of project results***: An experimental study measured the impact of experience with a human patient simulator on awareness of health careers and community health issues on students. ***Intellectual merit:*** Human patient simulators are used in modern medical education programs. This study examined if such simulations could provide a more authentic healthcare experience to secondary school students. ***Broader impacts:*** 1,299 students from 34 schools participated in a mixed-methods, sequential delayed research design. ***Publications:*** Project is ongoing. WestEd Co-PI Huang has experience with psychometric and statistical analyses for external evaluation in large-scale projects such as **Identifying Linguistic Factors Associated with Differential Student Performance on Middle School Science Assessments** (DRL-1348622; $1.1M; 2014-2019; PI: DeBoer; Co-PIs: Nelson-Barber, **Huang**). ***Summary of project results****:* The project distinguished cognitive and linguistic factors with a focus on English language learners (ELLs), through psychometric analysis, using more than 800 assessment items developed by AAAS and tested on more than 100,000 students. The project also developed and tested new versions of these items. ***Intellectual merit:*** Research specified the need for removing linguistic factors (e.g., complex sentence structure) for English language learners.

Hazard Learning Context

The three hazard modules address the performance expectation for the middle school Next Generation Science Standards [78]: MS-ESS3-2. Analyze and interpret data on natural hazards to forecast future catastrophic events and inform the development of technologies to mitigate their effects. The modules are based on three-dimensional learning elements [6], including *disciplinary core ideas* (ESS3.B: natural hazards), *science practices* such as using models, planning and carrying out investigations, analyzing and interpreting data, and engaging in argument from evidence, and the *crosscutting concept* of systems and system models. A classroom implementation of each module requires five to seven 45-minute class periods. In these modules, students use simulations during simulation-based scientific inquiry to develop understanding of the following hazard concepts [1]:

- *Scientific Factors:* Variables of the system that influence the progression of a hazard.
- *Impacts*: The types and severity of consequences for people living in and near a community.
- *Human Influences and Mitigation*: Human activities that increase or decrease potential risks and impacts, along with the values that influence an individual's risk perception.
- *Likelihood*: The probability of a hazardous event or impact occurring.

Figure 1. Wildfire Explorer

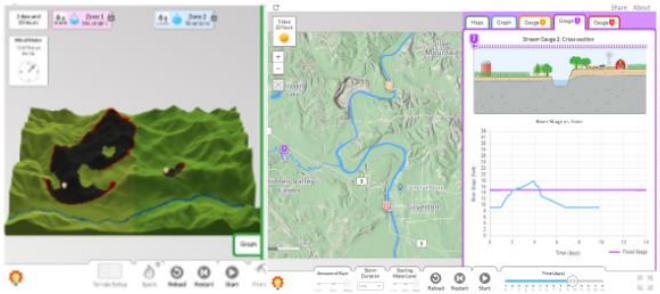
Figure 2. Floor Explorer

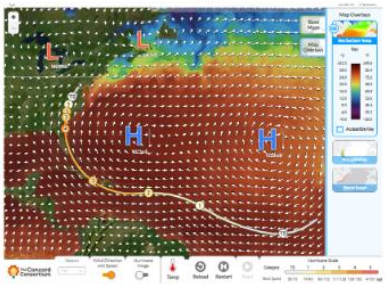

Figure 3. Hurricane Explorer

Simulations. In the wildfire module, students use the **Wildfire Explorer** (Figure 1) to investigate how

factors such as topography, fuel, moisture, and wind affect wildfire spread and intensity. They examine past trends in real-world data and predict future changes accelerated by rising global temperatures. Students also learn about and test mitigation strategies in the simulation, considering unintended effects of long-term wildfire suppression and the benefits of small, frequent wildfires for ecosystem preservation. In the flood module, students use the **Flood Explorer** simulation (Figure 2) to investigate factors that contribute to inland flooding, such as topography, surface permeability, water table levels, and precipitation. They explore the ecological and agricultural benefits of low-level flooding, as well as the hazards posed by extreme floods. Students analyze past trends and predict future flooding risks, considering the effects of increased human development and climate change. In the hurricane module, students use the **Hurricane Explorer** (Figure 3) to examine how sea surface temperatures, atmospheric pressure systems, and proximity to land affect the strength of a hurricane and its track. They investigate real-world hurricane cases to assess how scientific factors (wind, flooding) and social factors (infrastructure, population) are related to hurricane risks and impacts. Students then explore how rising global temperatures may alter these hurricane risks and impacts over the next century.

Hazbot Automated Feedback System Design

In the ESAAF project, HASbot operated as a single central agent scoring an individual student's record based on ML-based algorithms, matching the score to a pre-set feedback statement made by a domain expert, providing the text-based feedback to students, relaying the score to the teacher dashboard, and moving on to the student's next record without retaining any memory of the student. This limited feedback personalization when a student made multiple attempts to improve their arguments or simulation use within a task. In the ISLAND project, Hazbot will leverage large language models to tailor feedback more dynamically by processing a range of input data—such as simulation logs, data tables, snapshots, and text—for each student and retaining the student's record in each task. Feedback will include multiple media types, including graphics, videos, and pre-configured simulations and be adaptative to students' specific challenges. Hazbot will function as a two-tier system. The student-tier agent will provide individualized feedback based on all of the retained information while the teacher-tier agent will consolidate insights for individuals and as a class, suggesting targeted interventions for teachers. Below, we describe Hazbot design elements in more detail.

Simulation-based Inquiry Tasks. The wildfire, flood, and hurricane modules include a total of 28 simulation-based inquiry tasks (10, 10, and 8, respectively). These tasks address uncovering relationships between environmental factors and outcomes and making forecasts in hypothetical hazard scenarios. As shown in Figure 4, each task engages students in simulation-based inquiry, beginning with a guiding question such as, "How does wildfire spread across different vegetation types?," "How is the movement of a tropical storm affected by the Bermuda High?," or "Can people and their communities be protected even when extreme rainfall and urbanization continue into the future?"

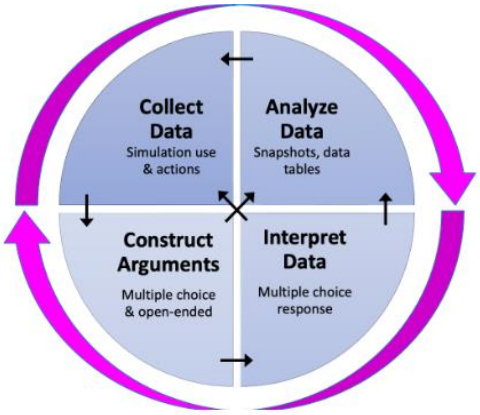

Figure 4. Four components of simulation-based inquiry

Within each module, the simulation tasks gradually increase in complexity, requiring students to manage additional variables and navigate growing uncertainties in simulation results. Within each simulation-based inquiry task students **collect data** in order to explore phenomena by setting up initial parameters, running the simulation, observing emergent phenomena, and collecting data. To **analyze data**, students take a snapshot of the simulation or create a table of data. To **interpret data**, students identify patterns from tables and snapshots by answering multiple-choice prompts. At the end, students **construct arguments** by making claims in the multiple-choice format and writing an open-ended justification to support their claims based on simulation evidence. The curriculum server logs each student's simulation interactions, data entries in tables or snapshots, multiple-choice answers, and open-ended responses. This information will be used by Hazbot to assess student performance and provide feedback in real time.

Automated Feedback. Automated feedback will be provided to enhance students' engagement with these four components of simulation-based inquiry (Figure 4). Note that the feedback examples described below are initial concepts, as the actual mechanisms and feedback will evolve based on the ISLAND project research, including the content, frequency, timing, and trigger (requested by students or enforced by the system). When triggered, automated feedback will be determined by Hazbot's automated scoring mechanisms developed with domain-specific rubrics. See RQ1 for rubric development. Students will be introduced to the rubrics in the first simulation task in each module so that they will be aware of what Hazbot is assessing. Feedback will be structured at three levels: a first-level general text-based prompt to orient student thinking and action, a second level specifying particular simulation setup or data collection using images or videos to highlight particular aspects for students' focus, and a third level offering simulations with preset configurations. Below we provide examples of automated feedback for each component of a simulation inquiry task found in the modules.

Figure 5. Collect feedback

Collect data. Students set up the simulation to generate phenomena and gather necessary data for analysis. For the first level, Hazbot will identify fundamental errors, e.g., incorrect initial parameter settings or insufficient model run time through a text-based message. If students click the "Check Simulation" button again, the second-level of feedback would highlight specific elements in the simulation needing adjustment (Figure 5). On a third request, Hazbot will reconfigure the simulation with correct parameter settings, for students to run and observe the phenomenon.

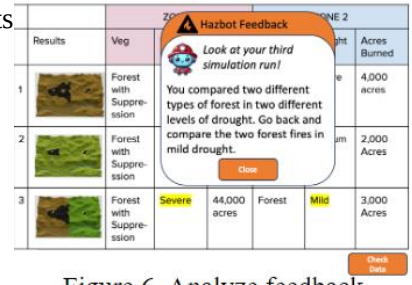

Figure 6. Analyze feedback

Analyze data. Students build a data table to identify patterns in the manipulated variable and the resulting outcome from multiple simulation runs. Hazbot will analyze the data table for the inclusion of the necessary evidence. If a student's data table lacks sufficient evidence, Hazbot will initially provide text-based feedback, prompting them to review their data. Should additional feedback be requested, Hazbot will deliver a second level of feedback visually, highlighting specific data points or variables within the table with text explaining what is missing (Figure 6). The third level would set up the simulation with pre-set parameters and have students rerun it and add new data to the table.

Interpret data. Students identify patterns in the data table or the snapshot by selecting an answer. When they click the "Check Answer" button, if wrong, Hazbot will disable immediate selection of any additional multiple-choice options, and feedback will prompt them to look at their data. Should they answer incorrectly again and need additional guidance, Hazbot will highlight relevant parts of their data table or snapshot to draw attention to critical patterns.

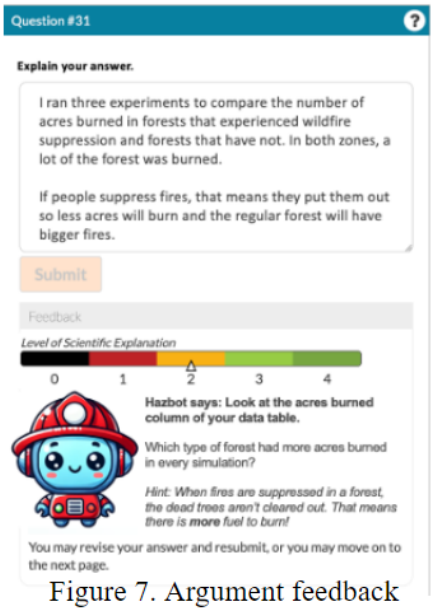

Figure 7. Argument feedback

Construct arguments. Feedback will be based on the level of justification based on evidence. First, if a student's justification lacks some critical data and reasoning, Hazbot will provide a generic prompt, encouraging students to add reasoning to support their claim. If further improvement is needed, a second level of feedback will specify areas within the justification where evidence or reasoning could be strengthened (Figure 7). A third level of feedback may prompt students to revisit earlier stages of inquiry—such as checking their simulation setup, data analysis, or interpretation stages—or direct them to review specific knowledge covered in the module to improve their justification.

Teacher Dashboard. Teachers will have real-time access to students' responses in the hazard modules via the enhanced

Hazbot Teacher Dashboard. The dashboard in the GeoHazard project displayed students' responses, with auto-scoring limited to multiple-choice items. In the Hazbot dashboard, teachers will be able to see the score and the type of feedback students received for each component of the simulation-based inquiry and how scores change in real time. The dashboard will also include a summary of class performance, flagging specific challenges individual students may be experiencing in each inquiry component and providing instructional strategies for addressing these issues. Teachers can also send direct feedback to each student or as a whole class through the dashboard.

Educative Curriculum Materials. The ECMs are designed as overlays to the entire module, not separate guides, to ensure that the teaching guidance provided within the ECMs is directly relevant to every student action and answer [79]. The ECMs developed in the GeoHazard project provide four types of support [80]: (1) essential subject matter and pedagogical content knowledge for teaching the module, (2) simulation tips covering design principles, features, functionalities, and suggested uses, (3) student support tips for conducting inquiry through simulations, and (4) student exemplar answers and related background information in each prompt in the module. The ECMs will be expanded to incorporate instructional strategies to teaching with a Hazbot system to enhance students' simulation-based inquiry.

Research Design

Research Timeline. Research will take place in two phases. In Phase I, we will employ design-based research [81], [82] to iteratively refine Hazbot through classroom implementations of hazard modules along with teacher dashboards and ECMs. We will collaborate with nine focus group teachers, three for each hazard module (see table at right), who previously taught hazard modules. The schools represent a range of geographical, demographic, and socioeconomic diversity and vary in proximity to the natural hazards targeted by the modules, with some close to regions impacted by the hazards and others farther away. In Phase II, WestEd will oversee a pilot study with a randomized control trial involving nationally recruited teachers to evaluate the impact of Hazbot-integrated hazard modules on student learning outcomes compared to modules without Hazbot integration.

Hazard	ST	School	Setting	Minority	Free lunch (%)
Wildfire	NJ	Thompson	Suburban	12.5%	3%
Wildfire	CA	Hong-Kingston	Urban	97.6%	71%
Wildfire	OR	Cascades	Suburban	10.7%	0%
Flood	KY	Rock County	Rural	5.3%	62%
Flood	IL	Gavin	Urban	43.3%	62%
Flood	NY	Highland	Suburban	28.3%	39%
Hurricane	FL	Park Avenue	Suburban	13.3%	20%
Hurricane	KY	Belfry	Rural	4.4%	71%
Hurricane	TX	Fort Worth	Suburban	32.3%	15%

Design cycle 1 (Year 1): Prototype testing. To begin, project partners will hold an in-person kickoff meeting to establish shared goals and project milestones and a virtual meeting with the focus teachers to discuss their experiences with the hazard modules through simulation-based inquiry and needs for scaffolding. Following these meetings, the project team will develop automated scoring models for three simulation tasks, one for each module along with prototype UI designs. In the next virtual meeting, focus teachers will review initial scoring rubrics developed by the project team and will test the prototype. Following the review from the teachers, we will revise the automated scoring models and the UI design and hold think-aloud sessions with six middle school students local to CC. Through these sessions, we will identify technical difficulties or usability challenges related to interface design, navigation, and system performance. Upon revising the models and the UI design, the project will continue to develop automated scoring models for additional simulation tasks and integrate Hazbot into at least half the simulation tasks for each module for a summer three-day, in-person workshop among project partners and focus teachers for fine-tuning automated scoring and feedback mechanisms.

Design cycle 2 (Year 2): Improving student experience with Hazbot. After the summer workshop, we will continue refining scoring models and integrating Hazbot into simulation tasks, aiming to finalize all tasks for the wildfire and hurricane modules by the beginning of the school year and the flood module by spring. During Design Cycle 2, focus teachers will implement their assigned Hazbot-integrated hazard module to their 450 students (50 per teacher), administer pre- and post-tests developed during the GeoHazard project, followed by post-implementation interviews and surveys. Project staff will conduct three classroom observations per teacher during implementation, totaling 27 visits, and will meet individually or in groups before and after implementation. Since module implementations are staggered

throughout the year, the project will utilize ongoing teacher feedback for iterative Hazbot revisions. At the second summer workshop, focus teachers will meet with the project staff and share their experiences, suggest revisions for the Hazbot automated feedback system, and discuss ideas and feature enhancements to add to the teacher dashboard.

Design cycle 3 (Year 3): Improving student and teacher experiences with Hazbot. Following the summer workshop revisions will focus on enhancing coordination between Hazbot's feedback for students' simulation tasks and the insights provided on the teacher dashboard. Focus teachers and their 450 students will implement the revised modules, with data collection aligned to Design Cycle 2. Project staff will meet with teachers in groups of three based on the module before implementations to go over the enhanced dashboard and ECMs and after implementations to gather their insights, instructional strategies, and recommendations. This analysis will inform further revisions to the Hazbot-integrated modules, dashboards, and ECMs.

Pilot study (Years 4 to 5). Prior to Year 4, all research tools—including assessment instruments, teacher interview protocols, and post-implementation surveys—along with the Hazbot-integrated modules, teacher dashboards, and ECMs will be finalized. Using a randomized controlled trial (RCT) design [83], we will conduct a pilot study with 72 nationally sampled teachers, 24 teachers per module. For each module, teachers will be randomly assigned to either the treatment group, using Hazbot-integrated modules, or the control group, using the hazard modules without Hazbot integration. This design minimizes bias from potential contamination (e.g., cross-class influence by the same teacher) and controls for teacher-level contextual factors, such as teaching experience and unrelated instructional practices. Teachers will participate in a 10-hour moderated online course to prepare to teach their assigned version of the module during the school year prior to their implementation. The RCT will be conducted at optimal times: the wildfire module in the first half of Year 4, the flood module in the latter half, and the hurricane module in the first half of Year 5. Pilot study teachers will complete online surveys, and selected teachers will participate in follow-up interviews. After the pilot study, control teachers will be supported to implement the Hazbot-integrated modules.

RQ1: How does Hazbot's automated scoring represent students' simulation-based inquiry?

a. What simulation interactions in the data collection stage are significantly correlated with students' data interpretation and argumentation?
b. How accurately do automated scores for student-generated argument texts align with human scores, and do these scores vary across demographic groups?
c. What relationships exist among automated scores for each inquiry stage—collect, analyze, and interpret data, and construct arguments—and do these relationships vary across demographic groups?

Rationale. This question will be explored when we develop automated scoring models. We will first identify simulation behaviors during data collection and analysis that are conducive to their interpretation and argumentation. We will also examine the alignment between LLM-generated and human scores for scientific argument texts and the relationships among the four components of inquiry using automated scores. Studying these relationships across demographic groups can detect and treat potential biases and transparency of Hazbot scoring, which is a basis for automated feedback [69], [84].

Extant Data Set. Since the modules became publicly available, all students' interactions with the elements in the module, including all four components of simulation-based inquiry (Figure 4) have been continuously collected on the server. We will use this growing dataset with each student assigned a unique computer-generated ID that links to log events and voluntarily provided demographic details, such as grade level, gender, race, and English as first or second language. We will use stratified sampling of 1,000 students for each simulation task to reflect diverse demographic groups and performance levels.

Data Analysis. For **RQ1a**, we will first score students' data interpretation and arguments. Multiple-choice claims will be scored as 0 (incorrect) or 1 (correct). Human scorers will use the reasoning-from-evidence rubric [85] that progresses as follows: from blank or off-task responses (0), to including no or incorrect knowledge and evidence (1), to mentioning relevant evidence or reasoning (2), to incorporating a single warrant between evidence and reasoning (3), and ultimately to including multiple warrants (4). To identify key simulation behaviors, we will extract events from computer logs of students' simulation

interactions, such as parameter setups, the range of parameters tested, number of runs, duration of each run, and total time spent on the simulation. Using machine learning algorithms, including decision trees and regression models [86], we will analyze which simulation behaviors predict students' data interpretation, claim, and justification scores and develop a rubric based on these behaviors for the data collection component of simulation-based inquiry.

To address RQ1b, we will train LLMs through in-context learning [18] for automated scoring of students' justifications. We will use randomly selected 200 responses out of the 1,000-response pool for training and the other 800 for testing. In-context learning of the training sample involves: assigning Generative Pre-trained Transformer (GPT) [87], a role of scoring student responses; describing the prompts that generated student responses including the simulation task; describing the scoring rubric for including score levels and criteria; providing student response examples for each score; and improving scoring performance through a chain of thought procedure. To examine accuracy of LLM-trained models, we will compare automated and human scores using various methods [88], including quadratic-weighted kappa (QWK) as moderate (0.41 - 0.60), good (0.61 - 0.80), and very good (0.81 - 1.00). We will aim to reach QWK values at 0.80 or higher [75]. We will then compare LLM and human scores across demographic groups for each simulation task, including gender, race, and whether English is the student's first or second language. By employing a confusion matrix, we will identify where the largest disagreements occur. Specifically, we will investigate whether certain groups are more frequently under- or over-scored by LLMs compared to human scorers and at which score levels and what simulation tasks these discrepancies are most prominent. For significant disagreements, a qualitative analysis will be conducted to examine specific linguistic or content features in student responses that might cause the automated model to score differently. They may include idiomatic expressions, informal language, or deviations from normative scientific language that may vary across student groups. With these findings, we will further refine rubrics and the automated scoring model training. If significant biases persist, these biases will be reported to teachers as inherent limitations of the Hazbot system.

For RQ1c, we will develop a rubric to assess analyze-stage artifacts (data tables and snapshots) for each simulation task, focusing on the sufficiency of data (enough trials to show a clear pattern between variables) and adequacy of evidence (controlling non-target variables across trials). Once the rubric criteria are set, we will use computer logs to autoscore the analyze component, eliminating the need for hand-scoring. This will provide automated scores data collection (RQ1a), construct arguments (RQ1b), along with auto-scored data analysis and interpretation components. We will then examine the correlations among these scores using Kendall's tau-b correlation coefficient, categorizing associations as very weak (tau < 0.1), weak (0.1–0.2), moderate (0.2–0.3), or strong (>0.3) to show whether increases in one component' score correspond with another. These correlations will also be analyzed across demographic groups (gender, race, language) to identify any variation.

RQ2. How does Hazbot feedback support students' simulation-based inquiry—such as collecting, analyzing, and interpreting data, and constructing arguments based on that data—within each task and their inquiry development across multiple tasks?

a. What actions do students take after receiving different types of Hazbot feedback?
b. Do students improve their scores after taking actions?
c. How do students reflect on their experiences using Hazbot feedback?
d. Do students' actions, score changes, and reflections vary across student demographic groups?
e. What patterns emerge in students' inquiry growth in connection with Hazbot uses over time?

Rationale. This analysis evaluates Hazbot's automated feedback functions to enhance student performance in each stage of simulation-based inquiry and gathers evidence to inform design changes based on students' actions and reflections across demographic groups. We will examine how students use Hazbot feedback across simulation tasks to determine if successful fading of Hazbot's guidance occurs, allowing students to maintain high levels of inquiry performance without continual feedback. We will use this set of questions each time design changes occur to Hazbot.

Data Sources. ***Student demographics.*** At the beginning of the module, students will answer questions about their gender, race, and whether English is their first or second language, and prior experience with

wildfires, floods, or hurricanes. ***Simulation task responses and computer log data.*** CC's server will log all student answers to the modules and interactions with the modules, such as simulation actions, feedback received, subsequent post-feedback actions, and any resulting score changes. All log events are time-stamped and linked to specific student and teacher IDs. ***Student feedback reflection.*** At the end of each simulation task, students will rate the usefulness of the feedback they received as very useful, somewhat useful, or not at all useful, with an optional field to explain their rating.

Data Analysis. For **RQ2a**, we will first extract relevant log events and categorize the actions into the four components of inquiry. We will identify patterns, such as whether students revisit simulations, modify data tables or snapshots, or revise their claims and justifications. Using time-stamped data, we will then examine the sequence of student actions post-feedback, tracking the frequency and time spent on these actions. We will also identify common behaviors or struggles. Frequent, meaningful, and longer interactions with revisions may indicate thoughtful engaging with the feedback [76]. To address **RQ2b**, we will examine whether students demonstrate improvement after receiving automated feedback. By comparing their automated scores before and after interacting with feedback, we will identify characteristics of students, tasks, rubrics, and feedback that contribute to the effective utilization of Hazbot. We will also examine cases of minimal or no improvement despite multiple revision efforts to identify potential issues. To examine whether students found the automated feedback useful for **RQ2c**, we will categorize the student ratings for each simulation task and compare rating distributions across tasks. We will then apply qualitative thematic coding to students' explanations of their usefulness ratings, identifying common themes such as clarity, relevance, or actionability. If certain simulation tasks or feedback statements consistently receive low ratings, these data may indicate areas for further refinement. For **RQ2d**, we will parse analysis results from RQ2a to RQ2c by student demographic groups (gender, race, language) using appropriate nonparametric tests. For example, we will examine whether certain student groups show greater gains in specific components of simulation-based inquiry or on specific simulation tasks. **RQ2e** explores students' cross-task development of simulation-based inquiry. Across multiple inquiry tasks (10 or 8) using the same simulation, we will track initial and final automated scores for each inquiry component for each student for each task. Scores for each component will first be normalized to a 0 to 1 range. Then, we will apply the Monte-Carlo Bayesian Knowledge Tracing (MC-BKT) algorithm to identify distinct patterns in inquiry growth, such as random fluctuation, improvement, decline, or consistent performance at either high or low levels [89], [90]. By integrating the levels of feedback utilized by students with MC-BKT results, we hypothesize several scaffolding patterns: (1) students who consistently require high-level scaffolding to maintain strong performance, (2) those who sustain high performance as support diminishes, (3) those who attempt to reduce scaffolding independently but experience declines in performance, (4) high-performing students who maintain success with minimal scaffold engagement, (5) students who rarely engage with scaffolding despite evident need, and (6) those who selectively use scaffolds based on particular task demands or conditions.

RQ3: What combinations of teacher facilitation and automated feedback are needed to support students' simulation-based scientific inquiry?

a. How do teachers use the Hazbot-generated information to support simulation-based inquiry?
a. What strategies do teachers employ using Hazbot classroom data, drawing from their own pedagogical content knowledge and the ECMs provided?

Rationale. We will collaborate with nine teachers as described in the project timeline using the focus user group methodology [91]. Teacher input will be solicited [92] by (1) orienting teachers to recognize feedback opportunities and teaching strategies to support student inquiry with simulations; (2) exposing teachers to Hazbot, the dashboard, and ECMs in their classrooms, (3) asking teachers to evaluate the Hazbot system features based on their implementation experience and to propose teaching strategies that were successful in their classrooms, and (4) soliciting their modification ideas to refine Hazbot and the ECMs and advise on the essential content for the online courses for the pilot study. Focus teachers will be supported to think critically about where Hazbot may fall short, failing some students even while advancing others as well as successful strategies [70].

Data Sources. ***Focus group meeting transcripts***. The focus teachers will meet several times a year over the first three years. Each year, they will implement a version of Hazbot-integrated simulation tasks and

participate in pre- and post-implementation meetings. They will attend two three-day in-person summer sessions to support Hazbot's design based on their implementation experiences and insights from project data analysis. All of these meetings will be audio-recorded and transcribed, and any written materials will be collected for further analysis. ***Class observations.*** We will observe each focus group teacher's implementation of simulation tasks three times per year and student interactions with the module and feedback. Observers will take notes on how teachers manage the classroom with Hazbot, including their use of the dashboard and ECMs, and the instructional strategies they employ. Observers will document students' interactions with Hazbot, noting whether the feedback is clear or causes confusion and identifying any challenges related to the interface design. Additional data will be collected from teacher logs that record how they use the synthesized dashboard data and supporting ECMs. ***Teacher post-implementation surveys.*** We will modify the post-implementation teacher surveys originally used for the GeoHazard project. The 40-question online survey collected demographic data, teaching experience, and credentials. It also included five sets of questions addressing how teachers used simulations, the instructional strategies employed for the hazard modules, their use of the teacher dashboard, their interaction with ECMs, and their perceptions of the strengths, weaknesses, and challenges of the module implementation, along with suggestions for revisions. We will add a set of questions related to Hazbot feedback, the modified teacher dashboard, and updated ECMs. ***Teacher interviews.*** We will develop semi-structured interview questions focusing on: (1) the instructional strategies teachers found effective for integrating automated scores, (2) the clarity and usability of the teacher dashboard, including how it presents student information and how easily it can be used during and after instruction, (3) opportunities, surprises, and challenges encountered with the current design of Hazbot, (4) any information or training they feel would enhance their use of Hazbot, and (5) instructional decisions teachers made, such as when to hold whole group discussions or identify students needing extra support, and their reasoning behind these decisions. This feedback will guide revisions to both the ECMs and the teacher dashboard, ensuring more diverse and targeted support for classroom adoption.

Data Analysis. We will triangulate patterns emerging from the various data sources to inform Hazbot's automated feedback functions for students and its synthesized data for teachers using a grounded theory approach [93]. This analysis will focus on: (1) identifying effective supports that maximize the learning potential of simulations across diverse classrooms; (2) determining the range of teacher actions that effectively scaffold the four components of students' simulation-based inquiry; (3) specifying the levels and types of feedback that can be automated to enable student progress without continuous teacher intervention; (4) designing summary information that allows teachers to efficiently assess students who would most benefit from direct teacher intervention; (5) identifying areas where individual or broader challenges emerge; and (6) gathering additional insights from teachers on what would enhance Hazbot's utility and close alignment with classroom needs. Recommendations from focus teachers for design adjustments will be carefully considered and incorporated into the revised Hazbot dashboard. Instructional strategies identified by focus teachers as effective in facilitating students' use of automated feedback will also be integrated into the ECMs to further support teachers. The instructional strategies and feedback support system co-developed with this focus group will inform the content of the online course designed to train pilot study teachers in the final two years of the project. The focus teachers will review this course and provide direct feedback for final revisions before dissemination.

RQ4. What is the impact of Hazbot-integrated modules on student learning?

- Do Hazbot-integrated modules improve student achievement in understanding natural hazards and constructing scientific arguments, as measured before and after the modules?
- What factors moderate the effects of Hazbot on student outcomes?
    - Do the effects of Hazbot vary by student demographic characteristics (e.g., gender, race, English as a primary or secondary language, hazard experience, or prior achievement levels)?
    - Is Hazbot more effective in improving student achievement when implemented by teachers with high fidelity compared to those with low fidelity?

Rationale. RQ4 will be addressed as an independent summative evaluation conducted by WestEd who will oversee teacher recruitment and selection, random assignment, and assessment instruments, and will analyze pilot test data to determine the impact of Hazbot on targeted student learning outcomes.

Teacher Recruitment. We will recruit pilot study teachers from two pools: One consisting of those who will have implemented the standard hazard modules through 2028 and the other consisting of 900 middle school science teachers who are alumni of AMS's teacher education courses, with half representing Title I schools. From the recruited list, WestEd will select 72 teachers (24 per module) based on criteria such as geographic diversity across urban, rural, and suburban schools, prior experience with the modules, classroom demographics, and teaching experience. Selected teachers are expected to teach at least 50 students. To address potential attrition, we will recruit six additional teachers per module on a waitlist. If necessary, these teachers will participate in the same online course and module implementation.

Treatment vs. Control Conditions. The recruited teachers will be randomly assigned to either the treatment group (using a Hazbot-integrated module) or the control group (using a standard module). Prior to implementation, each teacher will complete a 10-hour online course tailored to their assigned module, with two versions (standard and Hazbot) available for each hazard. These courses will cover subject matter, climate change, simulation-based inquiry for students, instructional strategies with simulations, formative assessment practices, and pilot study logistics. Treatment teachers will work with the Hazbot-integrated module, dashboard, and ECMs, whereas control teachers will use the standard versions.

Power Analysis. For *each* module, 24 teachers (12 treatment and 12 control) will participate with 50 students per teacher (totaling 1,200 students). Assuming (1) alpha = 0.05, (2) a two-tailed test, (3) power = 0.8, (4) the intraclass correlation coefficient is 0.15, and (5) the proportion of variance explained by covariates at each level is 0.6, the estimated minimum detectable effect size based on the PowerUp! Tool [94] is 0.31 for the proposed two-level hierarchical linear modeling (HLM) analysis where students are nested within teachers. As a reference, meta-analyses have found moderate to large impacts of feedback on student learning, ranging from 0.40 to 0.80 effect size in the past [95], [96].

Pre/Post-test. Before and after each module, students will complete an assessment instrument used by the GeoHazard project. Each instrument consists of 25 items designed to assess students' understanding of hazard concepts and their ability to construct scientific arguments from evidence. The instruments include a combination of multiple-choice items scored as correct or incorrect and constructed-response items scored 0 to 4 using the Knowledge Integration rubrics [97] or the evidence-based reasoning rubrics scored 0 to 4 [85]. WestEd will assess items and test properties using responses from 1,000 students pooled from Year 2 and the GeoHazard project based on the classical test theory (e.g., item difficulty, item discrimination, differential item functioning, and factor analysis) and item response theory (e.g., Rasch analysis). Based on the results, the instruments will be revised for the pilot study.

Impact Analysis. The impact analysis will be conducted separately for each module. Since students are nested within teachers, a two-level HLM will be used. To improve the precision of the treatment impact, we will include the pre-test measure of the outcome variable, student-level (e.g., gender, race, language, and hazard experience), and teacher-level characteristics (e.g., years of teaching science, science teaching certificate) as covariates in the impact model. The impact model takes the following form:

$$Y_{ij} = a_0 + b_1 \mathrm{Pre}_{ij} + b_2 \mathrm{Tx}_j + \sum b_s \mathrm{S}_{ij} + \sum b_t \mathrm{T}_j + u_j + \tau_{ij}$$

where subscripts *i* and *j* denote student and teacher, respectively; *Y* represents student learning outcome; *Pre* represents the baseline measure of the outcome measure; *Tx* is a dichotomous variable indicating student enrollment in a teacher/class assigned to the treatment or control condition (1 versus 0); and *S* and *T* are vectors of student-level covariates and teacher-level covariates, respectively, measured prior to exposure to the intervention. Lastly, u_j and τ_{ij} represent the random effect of teacher and student, respectively. In this model, the intervention effect is represented by b_2, which captures the treatment and control differences on the outcome variable after controlling for the covariates in the model. For the moderator/subgroup analysis, an interaction term of treatment indicator by subgroup will be added to the impact model. We plan to conduct the subgroup analysis by gender, race, language (English as primary vs. secondary language), and prior achievement level determined by the pre-test scores. To the extent possible, and for the treatment group, we also plan to conduct the mediator analysis to examine how the use of the intervention (e.g., time spent on activities, numbers of revisions after receiving feedback) interact with student learning processes and in turn, affect learning outcomes. A regression model or a structural equation modeling (SEM) approach will be used for the mediator analysis [98].

Implementation Fidelity Analysis. We will analyze backend log data descriptively to understand the

intervention dosage, including student interactions with Hazbot (e.g., time spent on activities, number of completed simulation tasks, and number of revisions after feedback during simulation-based inquiry) and teacher dashboard usage (e.g., frequency and duration of access and use of ECMs). We will construct a fidelity matrix to calculate the fidelity scores at each level, aggregated to the teacher level [99], and then determine teacher level of implementation (e.g., meeting expectations or not). We plan to conduct a three-arm impact analysis (treatment teachers with high fidelity of implementation versus treatment teachers with low fidelity of implementation versus control teachers). Should the intervention work, those students in the treatment group with high fidelity are expected to outperform the other two student groups, and those in the treatment group with low fidelity would perform similarly to the control students.

Mechanisms to Assess Success of the Project

A five-member Advisory Board will meet with project partners at the end of each project year to provide feedback on the design and revision of the Hazbot-integrated modules and teacher support resources, as well as on research planning, data collection, analysis, interpretation of results, and potential publication venues. Each member will provide written reflections to share with the project team, which will be included in the annual NSF report. **Wanli Xing, Ph.D.**, an Associate Professor of Educational Technology at the University of Florida, focuses on designing learning environments that leverage cutting-edge technologies—such as AI, computer simulations and modeling, and augmented reality—to support learning in diverse classrooms. **Libby Gerard, Ed.D.**, an Associate Adjunct Research Professor at the University of California, Berkeley, uses innovative learning technologies to develop automated scoring and real-time feedback for supporting student inquiry and assisting teacher instruction. **Jody Clarke-Midura, Ed.D.**, an Associate Professor at Utah State University, studies how to foster inclusivity in STEM education through the use of technology, helping students connect with STEM content in meaningful ways, and developing methods to measure STEM learning. **Sarah Fick, Ph.D.**, a Learning Scientist at Amplify, specializes in NGSS-based teaching and learning in K-12 science classrooms, employs design-based research to develop science curriculum and assessment materials, and has expertise in strategic planning for nationwide distribution. **Stephanie Harmon, M.A.,** is a distinguished science teacher and recipient of the 2014 Kentucky Outstanding Teacher of the Year award with extensive experience implementing the hazard modules and HASbot-integrated modules. She now provides professional learning services in Kentucky. In addition, **three scientists from the AMS membership** will provide specialized scientific advice focused on the simulations and subject matter knowledge needed for teacher resource development during the design phase. AMS frequently engages scientific advisory boards to support the development of teacher education courses and materials.

WestEd will conduct an external evaluation to ensure progress toward the project's stated objectives and timelines over the project period. WestEd will review the research plans at the beginning of each year and assess data collection, analysis, findings, and products at the year's end. By the end of Year 2, WestEd will conduct psychometric analyses of the wildfire, flood, and hurricane module assessment instruments to check validity. At the end of Year 3, WestEd will review research instruments, e.g., student and teacher assessments, surveys, and interview protocols—before the pilot study. For the pilot study taking place in Years 4 and 5, WestEd will oversee the random assignment of teacher participants to treatment or control conditions and independently analyze the pilot data to investigate the impact of Hazbot on student learning outcomes and conduct fidelity and moderator analyses specified in RQ4. WestEd will also contribute to dissemination work (e.g., conference presentations, peer-reviewed publications). At the end of each year, WestEd's evaluation reports will be submitted to the NSF.

Dissemination

The target population for the Hazbot-integrated wildfire, flood, and hurricane modules is middle school Earth science teachers who are implementing NGSS-aligned instruction. CC will create a designated website for the ISLAND project, featuring the Hazbot-integrated hazard modules along with teacher resources at no cost. AMS will also establish an ISLAND project webpage and provide a certification seal for curriculum resources to its affiliated teacher network. To reach potential users for dissemination, the project will use Earth science teacher lists, social media, and in-house publication venues run by CC and AMS. The project will publish articles in the *@Concord* newsletter—a biannual publication mailed to over 6,200 people and emailed to more than 65,000 subscribers, including teachers, administrators, and

education researchers as well as the *Bulletin of the American Meteorological Society* (a magazine of 12,000 subscribers, many of whom are educators). CC will utilize its blog, which has had over 47,000 views since July 2023 (according to Google Analytics), to post updates on module availability, research findings, pilot study opportunities, conference presentations, and publications. AMS will promote the modules to its list of nearly 1,000 K-12 educator alumni. The nine focus teachers will share the modules at science teacher conferences of their choice in Years 4 and 5, such as the National Science Teaching Association, and through teacher journals, such as *Earth Scientist* and *Science Scope*. We will disseminate findings at research conferences (e.g., American Educational Research Association, International Conference of Learning Sciences, American Meteorological Society, and National Association for Research in Science Teaching). The project will publish findings in academic journals in science education (e.g., *Journal of Research in Science Teaching, Science Education, International Journal of Science Education, Journal of Science Education and Technology*), teacher education (e.g., *Journal of Science Teacher Education*), Earth science education (e.g., *Journal of Geoscience Education*), and educational technology *(Computers and Education, Educational Technology Research and Development*).

Expertise

The PI at CC will oversee all project activities. CC, PF, and AMS will meet regularly to review progress, address challenges, and make any necessary adjustments to keep the project on schedule. WestEd will independently review research materials, analyses, and outcomes to maintain objectivity.

Hee-Sun Lee, Ph.D., (PI), a Senior Research Scientist at CC, will lead research on automated scoring and feedback and manage financial transactions. She has expertise in automated scoring and feedback system design, learning analytics, scientific argumentation, and simulation-based inquiry.

Amy Pallant, M.A. (Co-PI), a Senior Research Scientist at CC, will be responsible for overseeing the development of the Hazbot-integrated modules. She has directed numerous NSF-funded high-impact development projects that produced simulation-based Earth science curriculum materials.

Leslie Bondaryk, M.S. (Technology Lead), Chief Technology Officer at CC and an EdSafe AI Fellow, will lead the development of the Hazbot system. She has extensive experience in developing the safe, unbiased, and effective integration of AI technology in educational settings.

Sarah Haavind, Ed.D., a Senior Researcher at CC with backgrounds in supporting K-12 teachers in online and hybrid courses and professional learning communities. She will work with the nine focus teachers to develop teacher resources and online courses and coordinate teacher action research.

Gey-Hong Gweon, Ph.D., (Co-PI), the Founder and CEO of PF, a small business specializing in data analytics for educational settings with several in-house advanced algorithms, will analyze computer logs and build theories on AI-augmented scaffolding theory for simulation-based scientific inquiry.

Aaron Price, Ph.D., (Co-PI), Director of Education at AMS, will advise on scientific and pedagogical content, research methodology, analysis, and dissemination. He leads a team conducting online and hybrid K-12 science teacher PD programs with experience in experimentally designed research.

Chun-Wei (Kevin) Huang, Ph.D., (Co-PI), Senior Research Associate at WestEd specializing in measurement and applied statistics and designing rigorous experimental trials, will lead the psychometric analysis of the assessments, oversee the pilot study, and conduct the impact analysis.

Linlin Li, Ph.D., Research Director at WestEd, will review the research plan at the beginning of every year and evaluate the outcomes at the end of each year. She has expertise in highly sophisticated research design and evaluation and is a WWC-certified reviewer on group and single-case designs.

Broader Impacts

This project will design a next-generation automated feedback system to scaffold student inquiry in the study of natural hazards through scientific simulations. Designed as a fully integrated AI-enhanced two-tier pedagogical agent with supports for both teachers and students, the groundbreaking system will be capable of assessing and supporting the entire inquiry process from start to finish. With seamless integration into teachers' daily practices in real-world classrooms, the potential for student learning outcomes would be transformative. A total of 81 teachers and their 6,300 students will be involved in the project. Hazbot-integrated modules, dashboards, and teacher resources will be distributed nationwide at no additional cost. Products and findings from the project will be shared widely through conferences, workshops, and publications. Teachers will also be supported to present at conferences.