Responses to the Comments of Reviewer 1

Summary: In my opinion this is a thoroughly documented experiment and a well-written paper. The main message that different teaching styles appeal to different students is perhaps not entirely unexpected, but it might still be helpful to those with teaching responsibilities to see that this holds true within a hydrological teaching context.

1. **I do think that a few things can be clarified (see comments in the attached .pdf) and that section 4.2 and 4.3 in the Discussion could use some work. One element that could be added here is a brief discussion of the authors' experience of the benefits and drawbacks of creating materials for these various teaching styles, and touch on the question "is it worth my limited time?" that readers of this paper may be asking themselves.**

We thank the reviewer for their comments and suggestions. We understand the reviewer's point and will include a revised discussion. Please see our response to comment 28 from this reviewer for a detailed discussion related to "is it worth my limited time". We have also modified the title of discussion section 4.3 to be more inclusive of these themes. Please also see our response to comment 24 from this reviewer addressing the changes to section 4.2.

We respond to each individual comment below.

2. **Line 51: Is there a reason to assume that hydrologic education is somehow different from the broader STEM fields, that leads to this need?**

We will rewrite this section as follows:

"Hydrology possibly differs from broader STEM fields in several ways. First, geophysical challenges vary regionally, often resulting in region-specific educational case studies (e.g., flooding, drought, urbanization, agriculture runoff). STEM fields outside of the geosciences (e.g., medical research, chemistry) may have less regional variation in case studies. Second, compared to other STEM fields, hydrology curricula are decentralized. There have been reasoned calls to both decrease (Ruddell & Wagener, 2015) and increase the variation in hydrology curricula (Sanchez et al., 2016) concurrent with recommendations for new educational tools (Merck et al., 2021; Pérez-Sánchez et al., 2022; Lane et al., 2021; Slater et al., 2019). Given the regional nature of the geosciences, as well as variations in recent recommendations for future educational directions, there is a need to validate broader STEM findings within hydrology."

3. **Line 59: This is ambiguous. Do the authors mean an improved focus "on more material that is relevant [like high-res datasets are]" or "on material that is more relevant [than high-res datasets are]"?**

Thank you for identifying this phrase as ambiguous. We will change the language to more clearly reflect the "career-relevance" of using teaching tools such as high-resolution datasets.

4. **Line 87-89: I agree with this statement, but the authors may need to consider that studies investigation student perception of (hydrology) courses may be limited in the kind of assessments they can employ. Any assessment involving student grades can by definition be done at the level of the individual, but assessment involving student perceptions may need to be done through voluntary student participation in custom questionnaires. Teachers may be restricted by institutional policy in what they are allowed to measure through such informal assessment methods: policy may dictate that learning assessments must be performed at an aggregate level to guarantee anonymity for the students. Equally, students may be unwilling to participate in voluntary assessments at the individual level, if the perception exists that their answers may easily be traced back to themselves.**

We acknowledge that this study required substantial effort and institutional approval through our Institutional Review Board (IRB) to collect three different forms of assessments and de-identify students prior to analysis. This effort is likely infeasible to deploy in every class. We did not intend to suggest that these methods come at no expense. Instead, we meant that the results of prior studies that focused only on group mean perceptions may be missing detail and that this necessitates new studies that investigate this possibility. We will clarify this in the revised text as follows:

"Studies that focus exclusively on mean outcomes, often performed out of necessity to protect student privacy and ensure anonymity, may therefore only reflect the majority learning preferences of the students present and lose critical detail on the number of individual students reached with different teaching methods. Though anonymizing individual-level data to ensure student privacy may require additional personnel, the added data provides critical detail for assessing teaching modification impacts."

5. **Line 100: This terminology may be unfamiliar to readers outside of the US. Rephrasing this in terms of number of preceding years of hydrology education and expected knowledge/skill level will be more helpful to an international audience.**

We will change the terminology as recommended to be an "upper-level undergraduate course". We will also include expected courses taken prior to enrollment in this course.

6. **Line 110: If the following sections only discuss the final 4 weeks that were modified (as they appear to be), than it would be good to say so.**

We will restate that this modification happened in the final four weeks of the course as follows:

"The instructor remained the same across unmodified sections and all three modalities in the final four weeks of the course…"

7. **Line 114: The remainder of the text seems to use BMP as a noun to describe management infrastructure (i.e. things, instead of practices). Is this common use of the phrase?**

The term "BMP" to describe a stormwater best management practice installation is common practice and replaceable with "Green Infrastructure (GI)" and "Low Impact Development (LID)". "BMP" is the most commonly used abbreviation across academic fields, which is the reason we chose it. Unless it is recommended otherwise, we will continue to use BMP throughout the manuscript for consistency. And will ensure that all references use the same terminology.

8. **Line 131: Line 124 seems to suggest that this took 50 minutes.**

Thank you for identifying this error. The walking tour, though introducing specific stormwater management practices further studied in both the modeling and design sections, is considered a lecture as the students are primarily passively receiving information. This was the way we initially split the data for analysis, and the methods will be modified to reflect it. The total class times in each of the lecture, modeling, and design modalities will be adjusted to reflect the data. The tour did last for 50 min and the SWMM introduction was 15 min. We will clarify which time frames mentioned were total modality time versus an individual class period to reduce confusion.

9. **Line 141-143: I understand that these practices are part of designing infrastructure, but they seem to overlap in scope with the class periods dedicated to modeling. How can student preference for either modeling or design exercises be cleanly assessed if modeling is part of both?**

The "model" in the design phase of the class was a simple spreadsheet water balance (ΔStorage = Runoff - Infiltration - Overflow). Students were allowed to calculate runoff and infiltration with any method that had been introduced previously at any point during the course. It was misleading to refer to these calculations as a hydrologic model. We will clarify this in the revised text that these were spreadsheet calculations that did not differ substantially to the hand calculations that students were familiar with to this point in the class. Spreadsheets were used only to save time, not to facilitate the simulation of a more complex series of interrelated processes.

The Model phase of the class was carried out using a coupled H&H model simulating many physical processes in parallel (i.e., snowmelt, evapotranspiration, groundwater flow, runoff and stormwater detention in distributed Low Impact Development). Tasks centered on building-, compiling data for-, running-, and analyzing the results from-the model were substantially different from what the students had encountered previously.

We will clarify the differences between the two computer-oriented tasks in the revised text.

10. **Line 149: Were students given this survey before or after their grades for the modality were made known to them?**

All surveys were given at the end of the final class period before grades were known. We will clarify this in the revised text.

11. **Line 149: *Insert* 'And'**

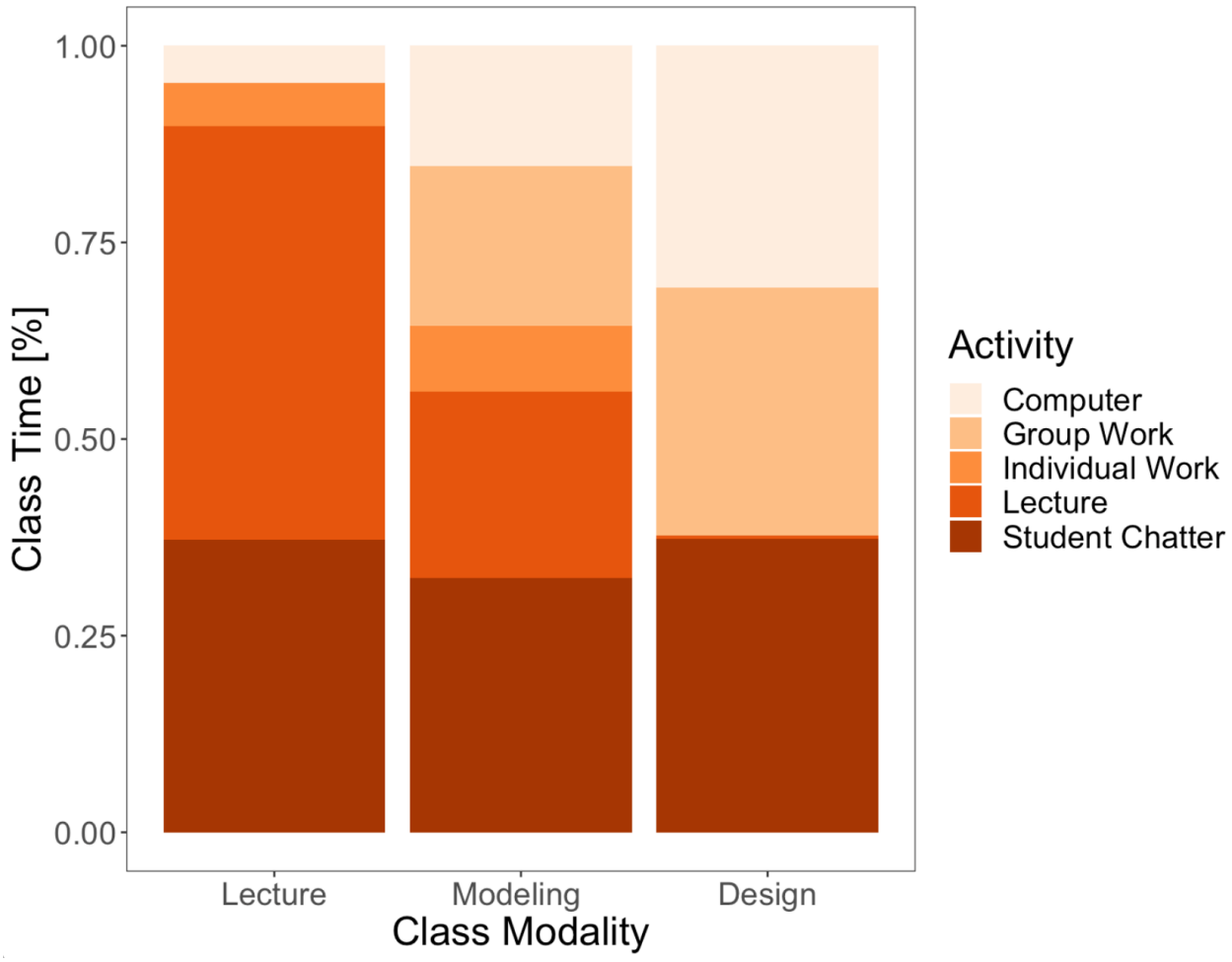
This was a typo that we will resolve in the revision.

12. **Line 151-152: How many out of the 25 students opted-in on linking their surveys to grades?**

The class had 25 total students. One student had to leave the course for personal reasons shortly after the midterm. A total of 20 students agreed to participate. The response rate was 20/24 = 83.3%. We will clarify this in the revised text.

13. **Why do the percentages of InstructorWalkTalk not match between Figures 1 and 2? They appear to me to describe the same activity.**

InstructorWalkTalk in the two figures described the same activity, but from two different perspectives: Figure 1 from the perspective of the instructor and Figure 2 from the perspective of the students. If we consider just the Design Module, while the instructor may spend most of the time walking around the classroom and engaging with students, the students in contrast are doing additional activities while the instructor cycles in this manner (e.g. working on a computer or in small groups). To clarify, we will remove InstructorWalkTalk from the student perspective figure to reduce co-occurring activities. Thank you for highlighting this confusing point and allowing us to clarify it. We will also change the color format of both figures to clarify that repeated colors do not reference the same activity. Instead we will use monochromatic color scales to increase accessibility of the figures as shown below in Figure 2:

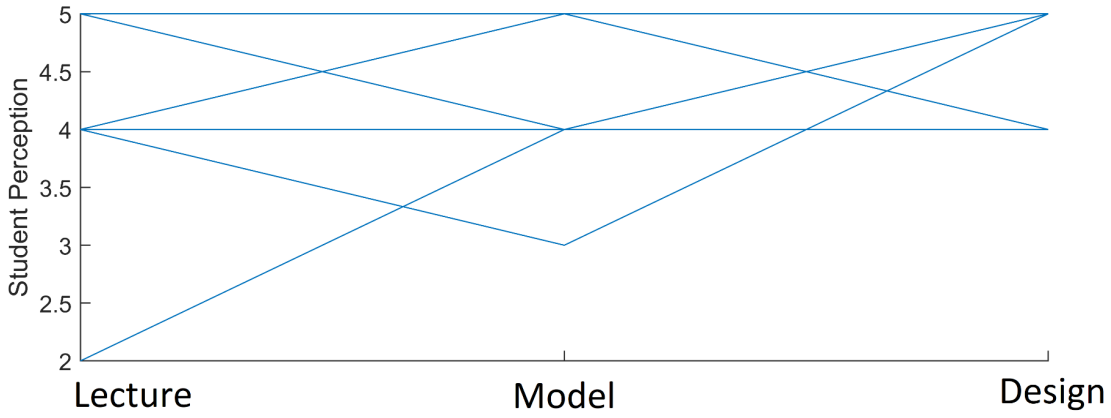


14. How many students filled out each of these surveys?

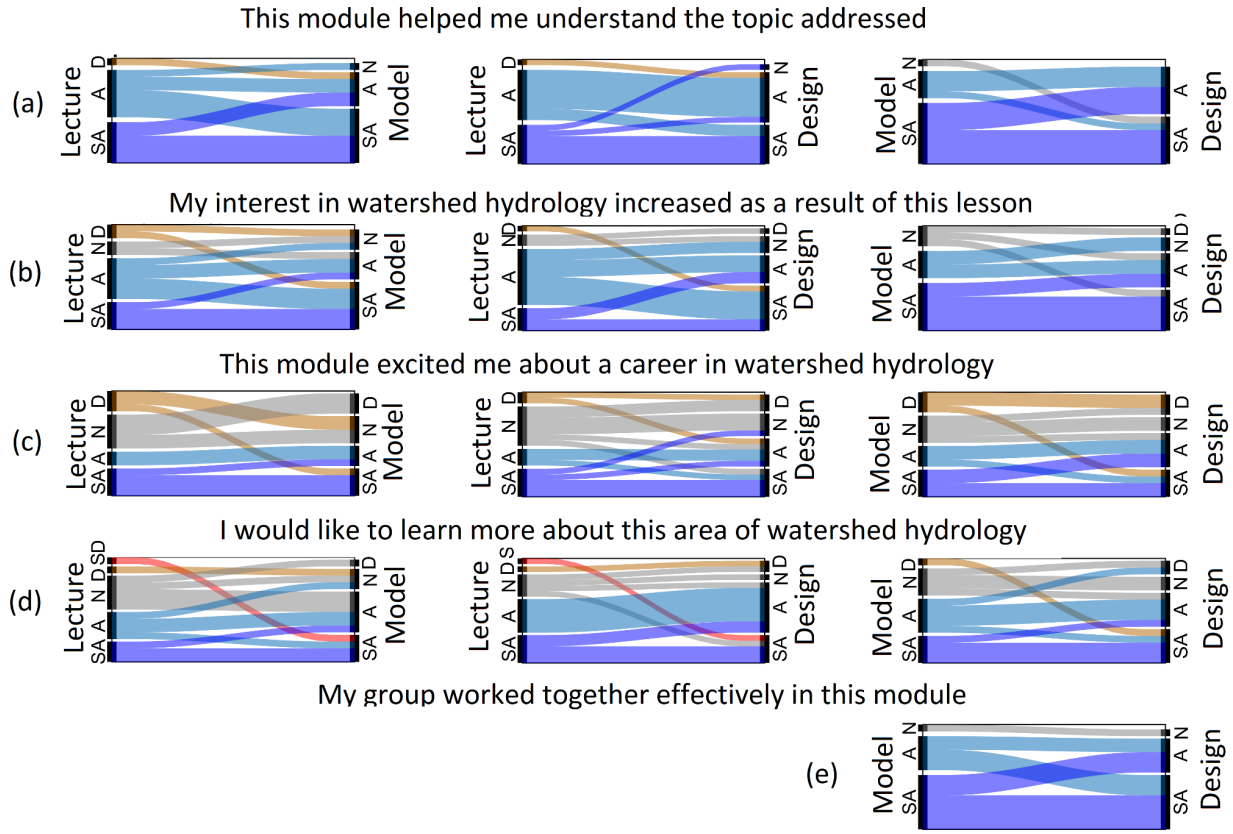
Out of the 20 that agreed to participate in the research, the lecture survey was completed by 18 students, the model survey by 16, and design by 18 due to absences. We will clarify this in the revised text.

15. This way of visualizing student preference seems to lose information, because the preference of any individual student cannot be tracked across the three teaching modes. Separating comparative preference between Lecture and Model on the one hand, and Model and Design on the other hand means it is unknown how student perception of Lecture compares to their perception of Design. A possible way to connect responses from an individual across all three teaching modalities could be a Parallel Coordinate Plot, with the three course types on X, student response on Y.

Per the reviewer's suggestion, we tried a parallel axis plot and it masks more information than it shows. The figure below is the result of the first question on Fig 4. This graph occurs because the data are all discrete so many lines overlap. The alluvial diagrams, while imperfect, show the proportions of students.



We have modified the alluvial diagram to contain another column showing lecture-design variations. We note that the total number of reported outcomes varies slightly as the alluvial diagram will only show results when individual students completed both surveys. Two students did not complete the lecture survey, four did not complete the model survey, and two did not complete the design survey. We will include this figure in the revised manuscript.



16. Assuming each dot represents an observation of cumulative number of questions asked, then it's unclear to me why these plots do not show a cumulative progression

of number of questions asked, but rather bounce up and down around a generally upward trend.

See for example Fig. 5a, red dots:

- No questions for the first 15 minutes

- 10 questions asked at minute 19

- < 10 questions asked between minutes 30 and 37 or so

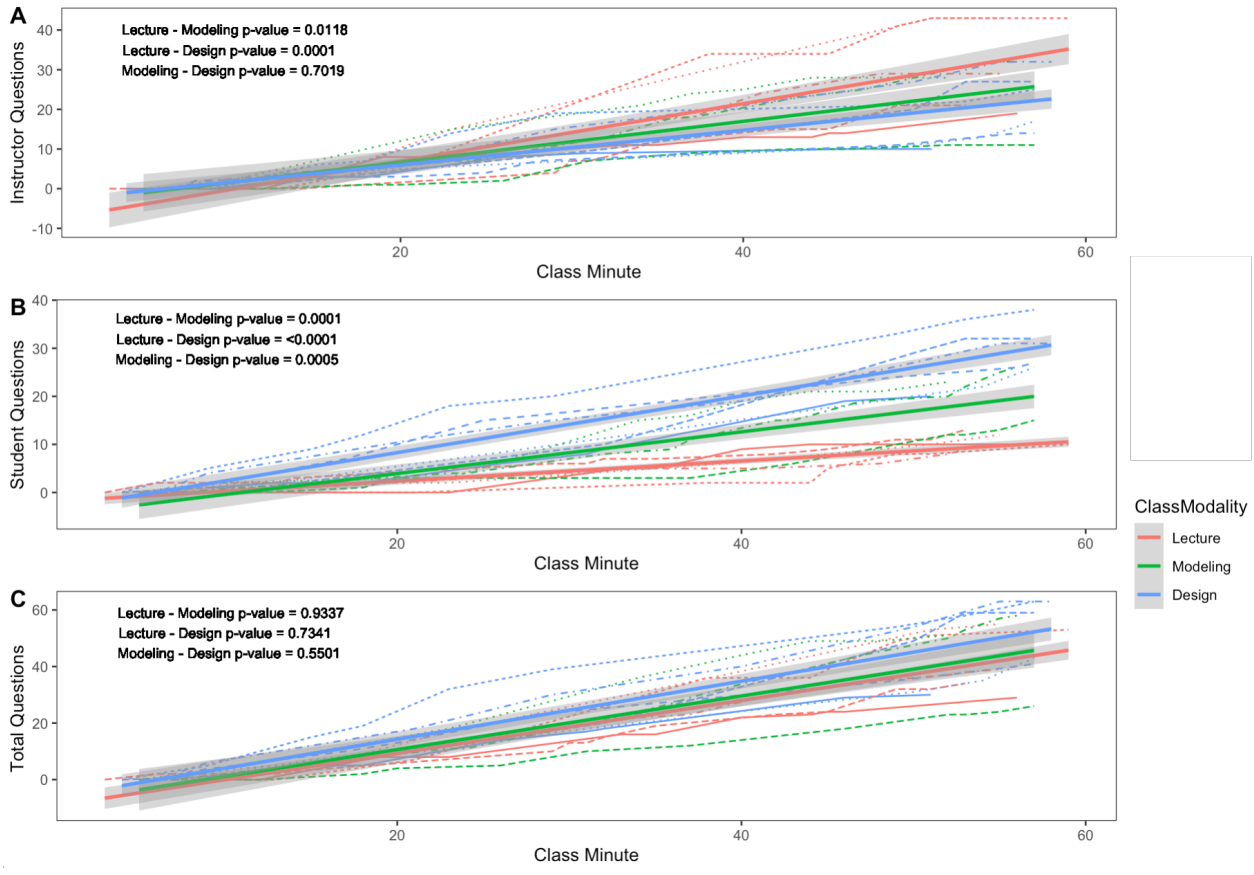
- A jump to 30+ questions asked around minutes 38, 39

- Back to 10 questions asked at minute 40

- etc.

I expect I'm missing something here but it's not clear to me what.

These figures show the overlapping cumulative questions across multiple class periods. We will replace the data for each class period, previously represented by points with lines distinct for each class period so that the cumulative questions are presented more clearly and the variability between classes clearer. The proposed modification is below. We will also clarify this in the the figure caption.



17. **Line 231-233: Is there a record of the types of questions asked? Having run tutorials myself I can imagine that at least a number of the questions asked during the modeling and design courses were along the lines of "when is the assignment deadline?" and "where can I find the assignment data?" Were such questions filtered out so that the graph shows only questions related to understanding/working with the presented material?**

We did not record questions asked by students live in class, but we can confirm that they were a mix of high level questions about hydrological processes as well as more routine questions. We considered any engagements to be worth counting, even if questions were only to clarify something related to an assignment. Therefore, the engagements were not filtered based on content. This detail will be added to the methods. The result shown on Figure 6 demonstrates that we were getting only 15/25 students engaged during the design module, which engaged the most individual students. We speculate that if students are left wondering about a particular detail of an assignment, they may not be fully capable of absorbing the more detailed hydrology concepts that are being presented. We also find (qualitatively) that once a student breaks the ice by asking a question out loud, they become more likely to ask subsequent questions in class, which may be of a more detailed nature.

18. **Similarly, group work tends to be noisy compared to class lectures. Where one question and one answer suffices to convey that information to all students present in a lecture scenario, the same question may need to be repeatedly answered if asked by separate groups in a group work scenario. Where such duplicates filtered out?**

Duplicate questions were not filtered out because we were interested in the total unique engagements, with each unique student counting as an engagement. However, upon receipt of duplicate questions the instructor addressed the class as a whole or wrote relevant information on the board, reducing further duplicates.

19. **To my understanding:**

    - The lecture focused on surface runoff, storm sewer design and storm water best management practices.

    - The model exercise focused on a field trip outside, optimization of infrastructure placement, model building and the running of simulations.

    - The design project focused on evaluating the performance of infrastructure on campus, data collection, data access, GIS analysis and basic modeling.

    Between the three modalities, not only the teaching style changed, but also the topics the student were tasked to work on. Therefore I think this statement as is, is not supported by the result of this experiment. Unless the topics can be considered

irrelevant for student engagement (which I don't think can be said), I believe this statement should read:

"By covering a broader range of topics and teaching styles, ..."

The reviewer's interpretation is correct. With respect to the rest of the course, these three modalities covered the same general topic: stormwater management. We do agree with the point the reviewer is making and will adopt the recommended wording in the revised text.

20. **Line 303: The reader might benefit from having a short list of examples of what such out-of-classroom experiences can be.**

Thank you for highlighting this point of interest. The 'experiences' to which we are referring are out-of-classroom pre-lecture preparations by the students that enable, if completed, maximum in-person learning. We will clarify this in the text as follows:

"Recent research on the application of student-led learning in science curricula suggests improved classroom experiences can possibly be negated by lack of out-of-classroom preparation by students (Akçayır & Akçayır, 2018; Chen et al., 2018)."

21. **Line 304-305: If I understand the methodology correctly, this is based on student performance on three written assessments (i.e. tests), taken after each of the three modalities concluded. Lectures transfer knowledge, but modeling exercises and design projects transfer additional skills (data handling, debugging, handling group dynamics, etc). Did these written tests account for transferable skills gained during the three different teaching styles? Students may have acquired extra skills that are not evident in Fig. 7.**

The longer-format exams were focused on knowledge. The three short format assessments included questions that accounted for an understanding of model functionality and required data manipulation. The assessments were designed with the goal of testing the most relevant skills of each section. We will make the assessments available as part of the supplemental material.

22. **Line 323: Relative compared to what other fields?**

Thank you for identifying the lack of clarity of this phrase. We intended to reference the cited study, which found most hydrology educators teach from independent material, with fewer than 20% using community-generated education materials, which presents both advantages and disadvantages. We will modify the phrasing so that rather than comparing hydrology to other fields, we focus on variation within the field.

23. **Line 325: Referencing should be consistent. These two publications are by the same person.**

Thank you for catching this citation error. It will be corrected and all other references checked.

24. **I don't think Section 4.2 in its current shape adds a lot to this paper. The listed benefits of decentralized curricula seem debatable to me (why would a 1000 different hydrology professors scattered across the world be able to update their teaching materials more quickly when they all work individually?) and the suggestion for centralized education evaluation criteria comes out of left field.**

Thank you for identifying this discussion section as unclear. We will further explain the advantages of a decentralized curriculum to clarify its inclusion in this manuscript. And will include the questions outlined in comment 25 as questions that will need to be addressed if we are to centralize the curriculum. We will modify the explanation of advantages of decentralized curriculum as follows:

"Heterogeneity in hydrologic curricula suggests that the field is well positioned to adapt to relevant case studies highlighting local and regional water issues, which may increase student interest in the field by grounding examples in familiar locations. Similarly, a decentralized hydrology curricula can also be rapidly updated to meet changing social or environmental conditions, advances in the field, or changing technologies by avoiding bureaucratic delays in implementing changes (Merwade & Ruddell, 2012)."

25. **Line 344-345: If this highlighted sentence is the main point of this section than I think the authors should expand on it. Key questions that this statement invites are:**

    - Who in the community or how should the community decide what these evaluation criteria are?

    - What aspects of education should these criteria cover?

    - Why does the hydrologic community need to invest effort into this instead of relying on findings from the broader STEM research into effective education?

    I'm not convinced that these questions can be easily answered.

More than proposing a complete centralization of the curriculum, we acknowledge that a centralized curriculum would allow for widespread use the pedagogies that have critically evaluated. A more useful tool would be a set of practices tested in the hydrological sciences that courses could be built from with regional and local examples, retaining the flexibility of a decentralized curriculum with the robustness of a centralized methodologies. We will rephase the sentence highlighted here to include this perspective, and include the questions outlined above as necessary considerations for forming a centralized hydrology-focused pedagogical best practices list. We will revise the highlighted section as follows:

“We therefore propose that the advancement of a field reliant on decentralized curricula should strive to adopt a set of centralized, and evaluated pedagogical best practices that can be applied across a diverse curriculum, retaining both the flexibility of a decentralized curriculum and allowing implementation of tested teaching practices. Establishment of such a set of practices requires consideration of who is included in this teaching and learning community, practices to be evaluated, and valuation within the field of hydrology and academia at large of time spent on evaluating teaching practices.”

26. **Line 350: Recommendations 1 (make sure the right computational infrastructure is ready to go) and 2 (take care when forming student groups when a number of students are absent) seem good recommendations to me, but also quite general (and possibly not very surprising) ones. Recommendation 3 relates to the way this experiment was designed and not to Student-Led Modalities, which does not fit the title of this section. The authors might consider splitting this section in two or rephrasing the current title.**

Thank you for citing this inconsistency. We will modify the title of this section to be “Practical Recommendations”, and also include a discussion of pros and cons as suggested and outlined below in our response to comment 28.

27. **Line 353-355: This seems contradictory to me, or did none of the students work on Apple machines? Also, although I realize that these terms are often used interchangeably colloquially, "PC" (personal computer) should probably be replaced by "Windows" (the specific operating system that imposes certain conditions on the compiled executable). [follow-up] thinking about this further, is the intended meaning of these sentences something like the following?**

    "Our hydrologic modeling modality did not specify any requirements for student laptops. However, US EPA SWMM, which we employed in this exercise is currently only offered as a compiled executable for Windows (though source code is also available). This led to a complication for students without access to a laptop with Windows, or the knowledge to readily compile source code themselves, resulting in a delayed start ..."

Thank you for identifying this terminology error. We will clarify this in the text as recommended and correct the use of “pc.” Three students did use Mac laptops. Two needed to pair with other students to complete the exercise. One successfully ran EPA SWMM on a Windows emulator, but this required substantial effort for both the student and instructor.

28. **Line 389-390: This recommendation might be coupled to a discussion item, where the authors outline the practical implications of using a higher diversity of teaching methods. Teaching is only part of most academics' workloads, and prepping and teaching the same material in three different ways takes longer than doing so in only**

one way. Some discussion of the (authors' experiences of the) pros and cons of offering multiple teaching styles could be a very helpful addition to the paper.

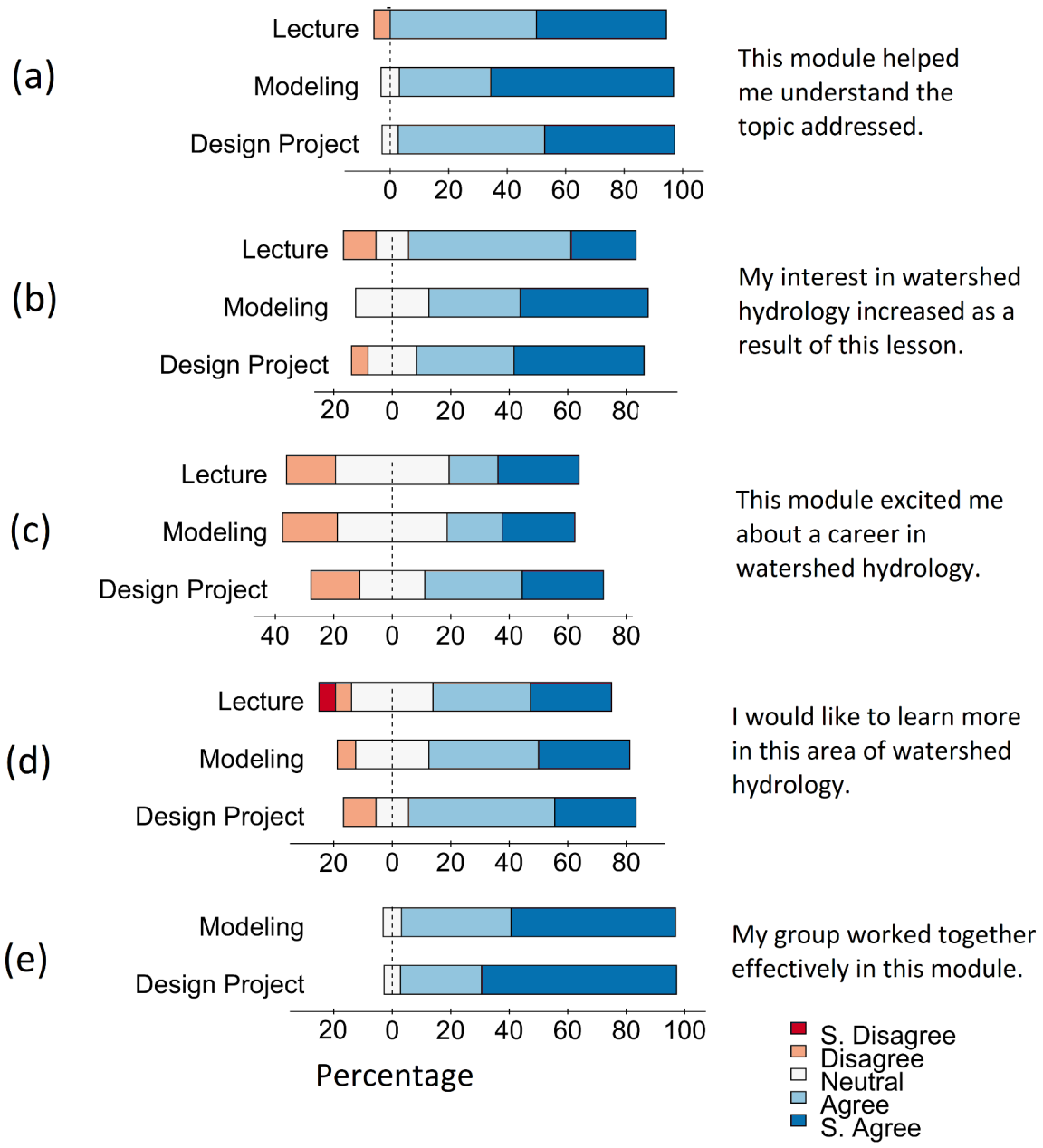
We will modify the title of section 4.3 to be inclusive of the discussion recommended here to "Practical Recommendations", and will include the following discussion to highlight the points recommended:

"We recognize that modifying existing curricula to include a wider variety of teaching modalities may be initially time intensive for instructors, and formally evaluating those changes additionally intensive. However, along with others (e.g. Wagener et al., 2007), we recommend an expansion of assessing pedagogical outcomes in hydrology to ensure implementation of best practices."

"And lastly, a data-rich experimental design such as the one discussed in this experiment with student engagement and participation data, survey responses, and de-identified grade analysis, may be infeasible for a single instructor to accurately collect alone in a traditional classroom setting. Classroom data collection by a third party enabled the depth of data collection required for this analysis and allowed the instructor to focus on preparation of teaching materials. Collaborators in university teaching and learning centers may already be trained on collecting this type of data and could support research to reduce time required of the instructor. Although the time required to (1) generate novel teaching materials and (2) evaluate novel practices is non-negligible, we believe the impact on students and ability to make data-driven decisions is meaningful."

29. **Using red for agreement feels intuitively strange to me. The authors might consider flipping the color scheme**

We will change the color scheme from having red for agreement in Figure 3. Thank you for this recommendation.


(a)
Lecture
Modeling
Design Project
0 20 40 60 80 100
This module helped me understand the topic addressed.
(b)
Lecture
Modeling
Design Project
20 0 20 40 60 80
My interest in watershed hydrology increased as a result of this lesson.
(c)
Lecture
Modeling
Design Project
40 20 0 20 40 60 80
This module excited me about a career in watershed hydrology.
(d)
Lecture
Modeling
Design Project
20 0 20 40 60 80
I would like to learn more in this area of watershed hydrology.
(e)
Modeling
Design Project
20 0 20 40 60 80 100
Percentage
My group worked together effectively in this module.
S. Disagree
Disagree
Neutral
Agree
S. Agree