Responses to Comments of Anonymous Reviewer #2

1. **This manuscript deals with the important issue of how we best teach hydrology and particularly hydrological modelling. While the presented study certainly might have its value, I am afraid I could not really see this in the presented manuscript. There are no clearly formulated research questions and I found it hard to understand what actually had been done and why. So, after ready the text a few times, I feel more confused.**

Thank you for highlighting the need to clarify our introduction. We will revise the introduction to state the two primary research questions more clearly. We also modify the introduction in accordance with the specific comments of reviewer 3. The primary research questions are:

- Is there consistency in results across the various methodologies for appraising the effectiveness of teaching modalities (i.e., student surveys, student assessments, in-class observations of student-instructor engagements)?
- Do any of the three teaching modalities (i.e., lecture, student-led modeling, student-led design studio) lead to significantly improved student perceptions, student assessment scores, or the frequency of in-class engagements?

Below are a few of the questions that I struggled with:

2. **The results might be heavily influenced by the temporal sequence of the different ways of teaching. Is the study design with one course with a mixture of teaching approaches really suitable to study the differences of the different teaching approaches?**

We understand the reviewer's concern; however, we note that this experimental design is common. Other studies have shifted teaching modalities within a single course rather than conducting the experiment with separate groups of students. For example, many instructors studied the effect of shifts to online teaching as a result of COVID-19 pandemic (e.g. Khalil et al, 2020). Other non-COVID studies have similarly implemented the experimental design that we choose: several teaching modalities implemented in sequence (e.g. Maeng & Kim 2011; Limperos et al, 2015; Setyono 2016).

The primary advantage of this experimental design is that the core group of students is constant. As reviewer 2 mentioned in another comment, variations across groups of students can be large and may confound results. We are intentionally eliminating this variability by exposing the same group of students to different teaching approaches at different times. Though the temporal effect cannot be extracted from our study, we did choose to implement the modifications starting half-way through the course. Starting half-way through reduces the influence of students "warming up" to the instructor and course and gaining confidence as is expected during the first few weeks of any course.

We note that reviewers 1 and 3 requested that we include a discussion of uncertainty in this study. We will include a discussion of sequencing in this section.

<ins>COVID-19:</ins>

Khalil, R., Mansour, A. E., Fadda, W. A., Almisnid, K., Aldamegh, M., Al-Nafeesah, A., ... & Al-Wutayd, O. (2020). The sudden transition to synchronized online learning during the COVID-19 pandemic in Saudi Arabia: a qualitative study exploring medical students' perspectives. *BMC medical education*, *20*(1), 1-10.

<ins>Assessment of teaching changes within a single course:</ins>

Maeng, S., & Kim, C. J. (2011). Variations in science teaching modalities and students' pedagogic subject positioning through the discourse register and language code. *Science Education*, *95*(3), 431-457.

Limperos, A. M., Buckner, M. M., Kaufmann, R., & Frisby, B. N. (2015). Online teaching and technological affordances: An experimental investigation into the impact of modality and clarity on perceived and actual learning. *Computers & Education*, *83*, 1-9.

Setyono, B. (2016, January). Providing variations of learning modalities to scaffold pre-service EFL teachers in designing lesson plan. In *Proceeding of International Conference on Teacher Training and Education* (Vol. 1, No. 1).

3. **Were the authors also the teachers? From the text this seems so but I could not see this clearly stated.**

This detail cannot be disclosed per IRB protocols as it could potentially be used to identify the students who participated in the study. We note that most prior studies using TAR approaches (that were reviewed in our introduction) also did not disclose this information.

4. **The number of participants is low, does this allow drawing conclusions? We all know how variable student populations are and how much the general 'mood' can vary from year to year (often based on a few students who 'set the tone'**

We can interpret this comment as asking two different questions: 1) was the sample size large enough to produce statistically significant results or 2) was a sample of 20 students truly a representative sample of all hydrology students everywhere. We will attempt to answer both:

1) The concern is the possibility of false-negatives or false-positives that result from a population of 20 students. When discussing statistical significance, we always presented p-values for each test rather than just a binary significance report (i.e., the null hypothesis was rejected or failed to be rejected) with respect to an α threshold. The p-value is the probability that a significant result was observed when in fact there was no underlying mechanism, which numerically accounts for sample size. In most cases, results were sensitive at the $\alpha < 0.01$ threshold, which suggests that it was highly improbable that our results were just random chance due to small sample sizes. We

will add more to the discussion to explain results in the context of uncertainty where p-values are in the $0.01 < \alpha < 0.1$ range. Where results were not significant, we followed a similar approach. For example, in the case of paired differences in assessment grades we observed no significant difference. The concern in this case is that a low sample size could have potentially resulted in a false negative (i.e., there was a significant change in assessment scores across teaching methods, but we did not detect one because of a small sample size). Our reported p-value was not a borderline case that could easily change with a larger sample size.

For the response rates (Fig 5), we note that these regressions were not as limited by the class size as individual students could ask more than one question (or none), and the entire "experiment" was repeated across at least three periods for each teaching modality. For the number of engaged students (Fig 6) we similarly were able to repeat this experiment across multiple class periods. We do note that some of these p-values are only sensitive at the $\alpha < 0.1$ and $\alpha < 0.05$ thresholds, which we do mention in Section 3.3, but will discuss in more detail in Section 4.

We agree strongly with the reviewer that year to year variability occurs in this type of course, which was explicitly considered when designing the experiment. If conducted across multiple years, we hypothesize that *cohort* would likely be a significant variable, and an unnecessary complication when attempting to analyze and discuss results. A result may have occurred not because of teaching methods, but as the reviewer says, simply because the "mood" differed. We decided that it was most informative to focus collecting more observations from one consistent group of students and one consistent instructor rather than across groups of students.

2) If the comment is asking if our group of students was truly representative of all hydrology students: this is more difficult to answer. The STEM education studies that we reviewed in our introduction had at most around 150 students. Even in these cases when sample size was larger, it is very unlikely that students all attending school in one geographic area were representative of all students everywhere. The students in this class were from a mixture of concentrations: civil engineering, environmental engineering, natural resources with a focus on water resources, and natural resources with a focus on forestry. I can only say qualitatively that there did not appear to be a subset of students who swayed the opinion of the entire class towards or away from any particular modality. We mention general school demographic data to give a sense of the environment that the students were in, as their responses may also be influenced by the more general academic environment in which they are participating. It is likely that our results are more relevant to students at similar institutions, and we will discuss this further in our discussion.

We note that reviewers 1 and 3 requested that we include a discussion of uncertainty in this study. We will include a discussion of sample size in this section.We will add the following to the discussion to address uncertainty of our results:

"4.4 Variability and Uncertainty of Results

Student responses to pedagogical practices may vary between cohorts of students and institutions. This study analyzed the effect of varying teaching modality on the same cohort of students to eliminate the effect of inter-annual variability of cohorts and instructors. We also note that the general university/learning environment may influence the ways students engage with differing learning modalities and their responses to changes in modality. We provide generalized background demographic data on this institution to provide the ability to compare between similar institutions. We anticipate our results to be more similar to student responses at institutions of similar size and composition.

Like many upper-level courses, this course was enrolled by a relatively small number of students, 24 students. Due to the class size, we designed the study such that our engagement regressions were not limited by class size, as individuals could ask multiple questions. Small class sample sizes are an obstacle that many upper-level course-based experiments may experience, but which is representative of the environment in which these topics are taught. "

5. **What was the return rate of the questionnaires? How many accepted the link to the 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%. All students who agreed to be surveyed also agreed to have grades linked to survey results. We note that several students missed class when surveys were administered so the return rate for individual surveys was lower than the acceptance rate for research participation (18-lecture, 16-modeling, 18-design). The total number of surveys collected and average number of students in class (21.7 students/day on average) across the study period will be added to the manuscript.

6. **The (very good) grades are of course highly influenced by the choice of questions and grading, the numbers alone do not say much**

We understand the reviewer's point, however, we first offer one clarification: none of our conclusions were based on the scores of any single assessment. The test variables were always the paired differences across tests. We will clarify that in the methods. We understand that test difficulty will impact scores. We will make the individual assignments given to students available as part of the supplemental material.

Each assessment has been implemented and refined in this course over a number of years. The average score for each assessment in prior lecture-only offerings (which were not included in our IRB review) were all similar (i.e., no significant differences across assessments). As each individual assessment was approximately of the same difficulty, we believed that paired differences in assessment scores were an objective metric to capture significant improvement (or worsening). We observed no changes in grades, which mirrored previous years using lecture only.

7. **How many questions were there in the questionnaire? Only the five shown in figure 3? I am no expert, but I would assume there are better ways to design questionnaires to get more detailed information.**

It appears that possibly the reviewer did not see that the full questionnaire was included as a supplemental material. The reviewer's general point holds though: the survey was short. The survey design was kept short for two reasons:

1) A longer survey doesn't necessarily result in a better survey. The questions were chosen to align with the grand challenges in hydrology education that have been highlighted in prior research on this topic (Thompson et al. 2012; Ruddel & Wagner, 2015). We will make this point clearer in our revision. Our questions were aimed at understanding: 1) perceived value of the teaching method, 2) interest in the material, 3) interest in further learning, and 4) interest in career development. We included one additional question to control for potentially problematic group dynamics. The survey also included an open-ended feedback section to test for theme saturation. In this field, students could introduce new ideas possibly missed in the survey. Students primarily used this section to reinforce their responses to the previous likert questions. The results from the open feedback section suggested that we achieved theme saturation on the topic with only a few efficient questions.

2) We wanted students to report perceptions immediately after engaging with the material, and not some time later when the material and experience was no longer fresh. We left approximately 5 minutes of class time at the conclusion of each module for students to complete the assignments, which necessitated an efficient survey.

Ruddell, B. L., & Wagener, T. (2015). Grand challenges for hydrology education in the 21st century. *Journal of Hydrologic Engineering*, *20*(1), A4014001.

Thompson, S. E., Ngambeki, I., Troch, P. A., Sivapalan, M., & Evangelou, D. (2012). Incorporating student-centered approaches into catchment hydrology teaching: a review and synthesis. *Hydrology and Earth System Sciences*, *16*(9), 3263-3278.