Supplemental Materials

"Cracking the code: An evidence-based approach to teaching Python in an undergraduate earth science setting"

Ethan C. Campbell*[§], Katy M. Christensen*, Mikelle Nuwer, Amrita Ahuja, Owen Boram, Junzhe Liu[†], Reese Miller, Isabelle Osuna, Stephen C. Riser

School of Oceanography, University of Washington, Seattle, Washington 98195

This file includes:

Figures S1-S3
Tables S1-S6
Supplemental References

* These authors contributed equally to this work.

[§] Corresponding author: Ethan C. Campbell, ethancc@uw.edu, School of Oceanography, University of Washington, Seattle, WA 98195, USA

[†] Current affiliation: Lamont-Doherty Earth Observatory, Columbia University, Palisades, NY 10964, USA

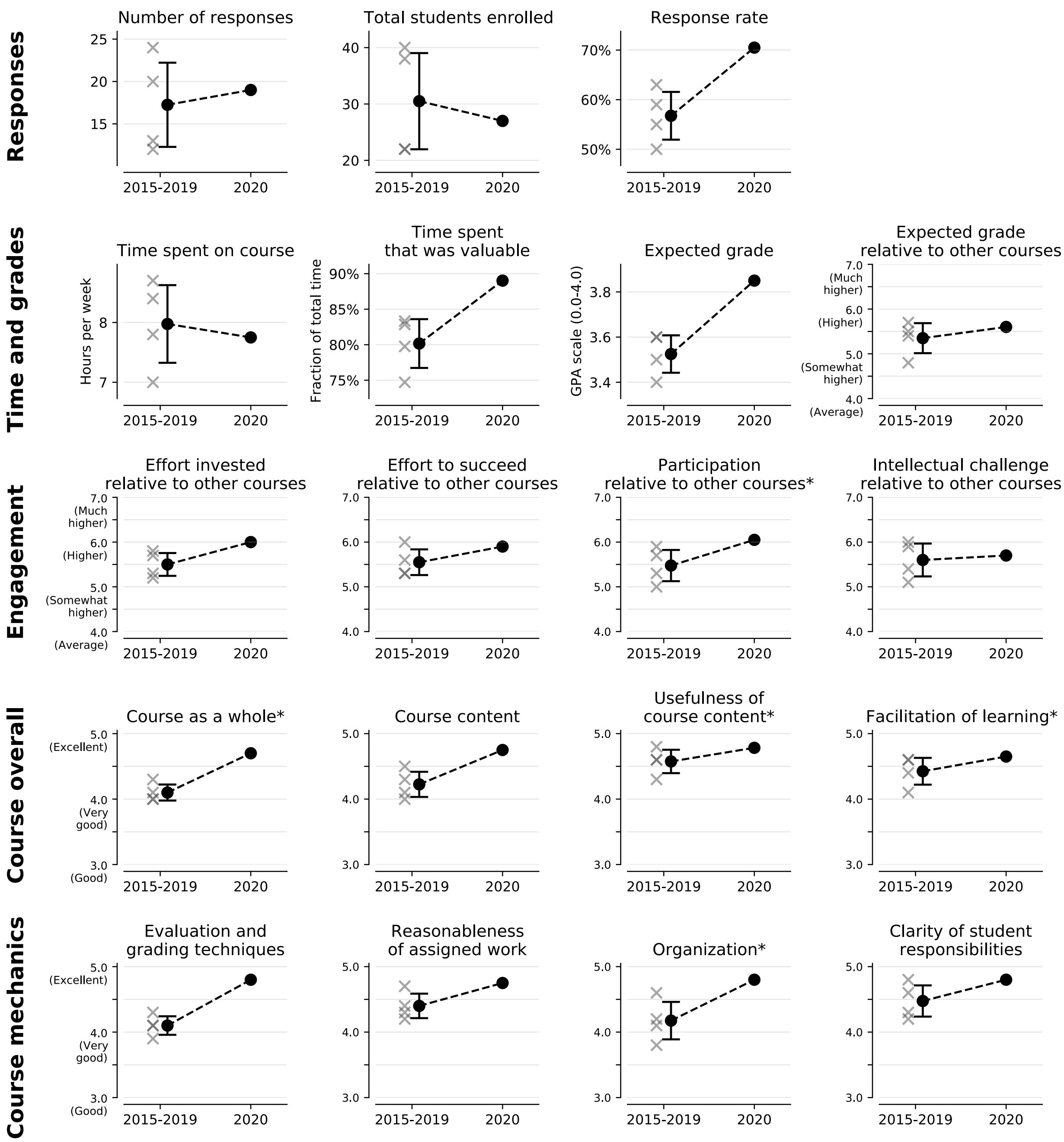

Figure S1. All metrics from anonymous end-of-quarter student evaluations in 2015, 2016, 2017, 2019, and 2020 (see Methods section "Initial, mid-quarter, and end-of-quarter surveys"). Differently worded questions were mapped between years as shown in **Table S2** in the Supplemental Materials. Metrics shown are class medians for 2015, 2016, 2017, and 2019 (gray crosses, except for those in the first row ["Responses"]); 2015-2019 mean or 2020 class median (black points); and 2015-2019 standard deviation (bars). Note that y-axes have been truncated from the full 1-5 scale ("Very poor" to "Excellent") or 1-7 scale ("Much lower" to "Much higher"). Survey questions for which a consistent mapping across years was not possible were excluded; instructor-specific questions are also not shown.

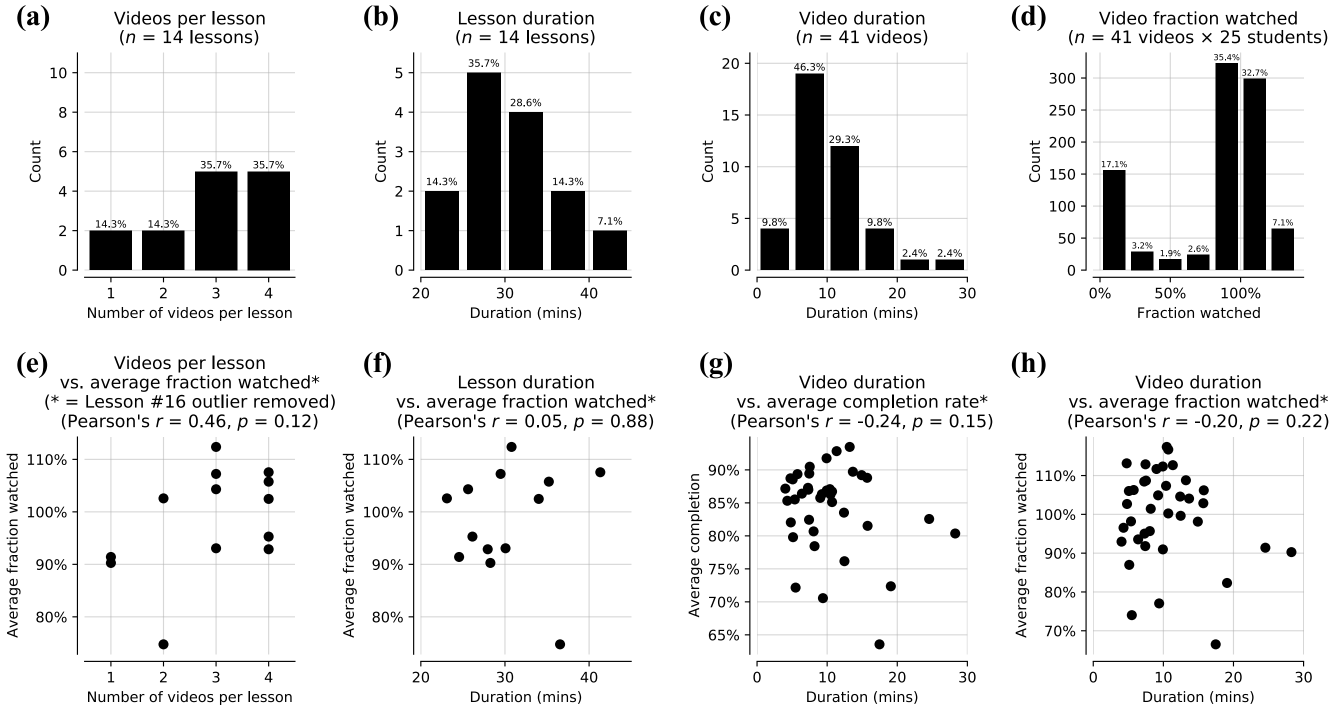

Figure S2. Additional statistics on flipped lesson videos that were posted and viewed on the Panopto platform, based on video-specific metrics obtained from Panopto. Pearson's *r* represents the linear correlation between two variables, which can be considered statistically significant if $p \leq 0.05$. Note that none of the correlations tested in panels **(e)**-**(h)** were significant. **(a)** Distribution of number of videos included per lesson (as the 14 topical lessons were usually split into multiple videos). **(b)** Distribution of the total duration of lessons. **(c)** Distribution of individual video duration. **(d)** Distribution of fraction of each video watched for each student. Fraction watched represents the total minutes that a specific video was viewed by a specific student divided by its duration, and thus can exceed 100% due to rewinds and repeat views. **(e)** Videos per lesson vs. video fraction watched, averaged across all students. Note that the final video lesson (Lesson #16) was excluded as an outlier where indicated with an asterisk (*) due to its lower viewership. **(f)** Lesson duration vs. fraction watched, averaged across all students. **(g)** Video duration vs. completion rate, averaged across all students. Completion rate represents the fraction of a video that was viewed at least once, and thus is capped at 100% for a specific student and video (unlike "fraction watched"). **(h)** Video duration vs. fraction watched, averaged across all students.

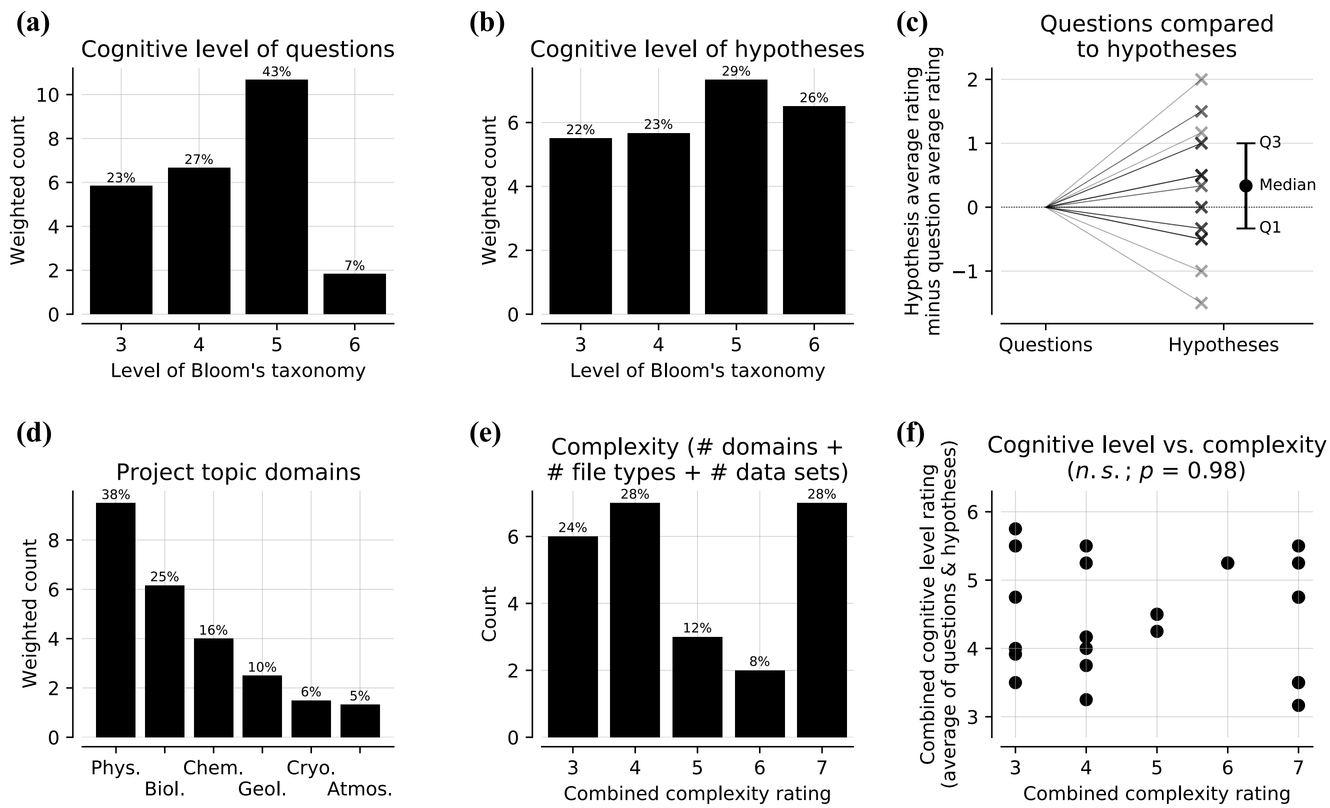

Figure S3. Analysis of students' final project questions and hypotheses based on the cognitive process dimension of the revised Bloom's taxonomy (Krathwohl, 2002). Each student's questions and hypotheses (up to three each per student) were assessed using the rubric and weighting described in **Table 2**, with higher levels of Bloom's taxonomy representing higher-order questioning and prediction. **(a)** Distribution of cognitive level of students' questions. **(b)** Distribution of cognitive level of hypotheses. **(c)** Each student's questions (which were posed first) compared to their hypotheses (posed second), with the median and interquartile range (25%-75%) of change from questions to hypotheses shown at right. Darker lines and crosses reflect more than one students' data. **(d)** Distribution of domains of students' final projects. If a student's project touched multiple domains, each domain was weighted such that, for example, a project spanning three domains would contribute ⅓ of a point to each of the domains' total count. **(e)** Distribution of final project complexity, defined as the number of domains (see panel [d]; classwide minimum: 1, maximum: 3, mean: 1.8) plus the number of data file types used in a student's project (either CSV or netCDF or both; classwide min: 1, max: 2, mean: 1.2) plus the number of distinct data sets analyzed in the project (classwide min: 1, max: 4, and mean: 1.8). **(f)** Average cognitive level of students' questions and hypotheses vs. the combined complexity rating from panel [e]. No significant Pearson's correlation between the two variables exists ($p = 0.98$).

Table S1. Rubric used to assess students' prior coding experience based on their written responses to the Assignment #0 survey during Week 1 of the course. Students were asked: "Do you have prior coding experience, and if so, with what language?" and "How comfortable do you feel using technology?" Responses to the first question were graded subjectively based on word choice on a scale from 1-5, using the keywords in quotes (e.g., "a little") when present. Additional points were awarded to weight responses in favor of prior exposure to Python or similar high-level and/or interpreted languages (MATLAB, Java, R). Points were subtracted to account for less relevant prior experience. Results are presented as the metric "Prior coding experience" in **Fig. 6**.

1	**2**	**3**	**4**	**5**
No experience	Minimal experience (e.g., "a little", "small", "tiny amount")	"Some" or "moderate" experience	Experience	Experience (with full additions)

Additions (maximum total: +1.0)	**Subtractions (maximum total: -0.5)**
+0.5 for one of MATLAB, Java, R	-0.5 if response mentions many years since their previous experience
+1.0 for Python or multiple languages	-0.5 if response mentions that their previous experience was not useful

Note: If no level of coding proficiency was provided, the base number is from the students' "comfort with technology" statement ("Very comfortable": 4; "Fairly comfortable": 2).

Table S2. Mapping of IAS (university-administered) final course evaluation questions from 2015-2019 to 2020. The mapping allows the slightly different evaluations from the two periods to be compared in **Fig. 2** and **Fig. S1** in the Supplemental Materials. Note that all metrics listed are the median of all responses collected for the class.

<table>
<tr><th>Paraphrased question</th><th>Original survey question(s) (2015-2019)</th><th>Original survey question(s) (2020)</th><th>Metric and units</th></tr>
<tr><td>Time spent on course</td><td colspan="2">On average, how many hours per week have you spent on this course, including attending classes, doing readings, reviewing notes, writing papers and any other course related work?</td><td>Hours per week</td></tr>
<tr><td>Time spent that was valuable</td><td colspan="2">From the total average hours above, how many do you consider were valuable in advancing your education?</td><td>Hours per week, expressed as percent relative to response to question above</td></tr>
<tr><td>Expected grade</td><td colspan="2">What grade do you expect in this course?</td><td>GPA scale (0.0-4.0)</td></tr>
<tr><td>Expected grade relative to other courses</td><td colspan="2">Do you expect your grade in this course to be:</td><td rowspan="5">1-7 scale (“Much lower” to “Much higher”)</td></tr>
<tr><td>Effort invested relative to other courses</td><td colspan="2">The amount of effort you put into this course was:</td></tr>
<tr><td>Effort to succeed relative to other courses</td><td colspan="2">The amount of effort to succeed in this course was:</td></tr>
<tr><td>Participation relative to other courses</td><td>Your involvement in course (doing assignments, attending classes, etc.) was:</td><td>Relative to similar courses taught in person, your participation in this course was:</td></tr>
<tr><td>Intellectual challenge relative to other courses</td><td colspan="2">The intellectual challenge presented was:</td></tr>
<tr><td>Course as a whole</td><td>The course as a whole was:</td><td>The remote learning course as a whole was:</td><td rowspan="4">0-5 scale (“Very poor” to “Excellent”)</td></tr>
<tr><td>Course content</td><td colspan="2">The course content was:</td></tr>
<tr><td>Usefulness of course content</td><td>Relevance and usefulness of course content were:</td><td>Average of: “Usefulness of reading assignments in understanding course content was:”, “Usefulness of written assignments in understanding course content was:”, “Usefulness of online resources in understanding course content was:”</td></tr>
<tr><td>Facilitation of learning</td><td>Amount you learned in the course was:</td><td>The effectiveness of this remote course in facilitating my learning was:</td></tr>
</table>

Evaluation and grading techniques	Evaluative and grading techniques (tests, papers, projects, etc.) were:		
Reasonableness of assigned work	Reasonableness of assigned work was:		
Organization	Course organization was:	Organization of materials online was:	
Clarity of student responsibilities	Clarity of student responsibilities and requirements was:		
Instructor’s contribution to the course	The instructor's contribution to the course was:		
Effectiveness of instructor’s teaching	The instructor's effectiveness in teaching the subject matter was:		
Quality of instructor answers and feedback	Average of: “Explanations by instructor were:”, “Instructor's ability to present alternative explanations when needed was:”, “Instructor's interest in whether students learned was:”, “Answers to student questions were:”	Quality/helpfulness of instructor feedback was:	

Table S3. Open-ended questions asked in IAS (university-administered) mid-quarter and final course evaluations in 2020. Students’ anonymous responses are tabulated in **Fig. 3** and are excerpted throughout this study.

Evaluation period	**Question**
Mid-quarter	What is helping you to learn in this course?
	What is hindering your learning in this course?
	What can your instructor do to improve your learning in this course?
Final	Was this class intellectually stimulating? Did it stretch your thinking? Why or why not?
	What aspects of this class contributed most to your learning?
	What aspects of this class detracted from your learning?
	What suggestions do you have for improving this class generally?
	If this course were offered remotely again, what suggestions do you have to improve the student experience?

Table S4. Functions, operators, and methods taught in the course that were used as search terms to assess the complexity of students' final project code. A Python script was used to count instances of each search term in students' project code notebooks, and the number of search terms used at least once (expressed as a percent of all search terms below) is presented as the metric "Python skills used in project" in **Fig. 6**.

Topic	**Search terms**
Basic functions	'len(', 'print(', 'display(', 'range(', 'enumerate(', 'zip(', 'int(', 'float(', 'complex(', 'bool(', 'tuple(', 'type(', 'readline('
Lists	'list(', '.append(', '.extend(', '.insert(', '.remove(', 'del', '.pop(', '.reverse(', '.copy(', '.join(', '.sort('
Strings	'str(', '.lstrip(', '.rstrip(', '.upper(', '.lower(', '.count(', '.replace(', '.split(', '.format('
NumPy	'np.array(', '.dtype', '.astype(', 'np.append(', 'np.insert(', 'np.flip(', 'np.tolist(', '.sum(', '.mean(', '.median(', '.max(', '.min(', 'np.std(', 'np.pi', 'np.e', 'np.inf', 'np.nan', 'np.absolute(', 'np.round(', 'np.sqrt(', 'np.exp(', 'np.sin(', 'np.cos(', 'np.zeros(', 'np.ones(', 'np.full(', 'np.arange(', 'np.linspace(', '.size', '.ndim', '.shape', '.reshape(', ',flatten(', '.transpose(', '.vstack(', '.hstack(', 'np.genfromtxt(', 'np.meshgrid('
Time	'datetime.now()', '.year', '.month', '.day', '.hour', '.minute', '.second', '.microsecond', 'datetime.strptime', 'datetime.strftime', '.total_seconds()', 'timedelta', 'mdates.date2num('
Pandas	'.Series(', '.index', '.values', '.loc[', '.iloc[', 'pd.concat(', 'pd.DataFrame(', '.describe(', '.to_csv(', '.read_csv(', '.read_excel('
Xarray	'.open_dataset(', '.open_mfdataset(', '.attrs', '.isel(', '.sel(', '.item'
SciPy	'stats.linregress(', 'interpolate.interp1d(', 'interpolate.griddata('
Plotting	'.figure(', '.subplots(', '.xlabel(', '.ylabel(', '.set_xlabel(', '.set_ylabel(', '.grid', '.colorbar', '.set_label', '.clabel', '.invert_yaxis', '.gca', '.axes', '.coastlines', '.add_feature(', '.set_extent('
Plot types	'.plot(', '.scatter(', '.hist(', '.contour(', '.contourf(', '.pcolormesh('
Logic	' if ', ' while ', ' for ', ' is ', ' in ', ' not ', ' else:', ' elif ', ' and ', '~', '==', '!=', '>=', '<='

Table S5. List of guiding questions offered to undergraduate student coauthors for structuring their testimonial submissions, which are presented in **Box 1** (see Methods section "Student focus group"). Students were encouraged to address one or more of the questions in their submissions.

1. How did your prior experience with coding (or lack of prior experience) impact your experience with the course? If you have prior coding experience and it was self-taught, what do you see as the benefits of learning scientific programming in a structured environment rather than teaching it to yourself? If your prior coding knowledge was learned from course(s), how did we teach programming that was different and more or less effective than those past course(s)?
2. How did the accessibility elements that we implemented (e.g., captioning, syllabus late policy, extensions, not grading on attendance, breaks during class, virtual office hours, making slide decks available, video optional on Zoom, ability to use chat during class, no course prerequisites, extra credit opportunities, etc.) affect your success in the course?
3. How did the expectations and norms established in the course impact your experience?
4. How did you navigate the course policies we created on collaboration and original work? If you worked with a partner on assignments and/or the final project, what was your experience like? Was it productive/challenging/surprising, and how did the technological tools we used (Colab, Zoom) facilitate it? What advice would you give to professors who are teaching a programming course and want to create opportunities for collaboration?
5. How did the key course elements (recorded videos, in-class activities, assignments, final project, etc.) and technological platforms (Google Colab, Piazza, Zoom, Google Drive/Docs, Canvas) help or hinder your learning?
6. Instead of a textbook, we allowed use of external resources (e.g., documentation websites, Stack Overflow, etc.). How did this compare to having a textbook for the course?
7. How did guidance from the instructors and classmates (via Piazza or in class) help you complete assignments and shape and execute your final project?
8. In what ways did the class help you learn about oceanography sub-disciplines (marine geology, chemistry, physics, biology) or other earth science subjects adjacent to oceanography (e.g., cryosphere, meteorology, climate)? What value do you see in teaching programming in an oceanography curriculum rather than a computer science department?
9. How do you feel this course fit into your overall undergraduate education? How did this course prepare you for future research, like your senior thesis? In what ways do you feel more capable now that you have Python in your arsenal?
10. How do you feel this course shaped your career/life goals or motivation to pursue oceanography or data science during and after college?
11. What was it like taking this class during the pandemic? How does this course compare to other classes you've taken remotely during the pandemic?

Table S6. Grading rubric for students' final research projects. This rubric was provided to students to delineate expectations and evaluation techniques.

Presentation Content				
	Limited (0-50%)	Good (50-75%)	Exceptional (75-100%)	
Background	Topic background is missing or severely lacking in detail.	Topic background is sufficient, but missing some details or lacks coherency.	Topic background is clear, complete, and relevant.	3 points
Questions / Hypotheses	Questions are not well-defined. Hypotheses are not substantiated.	Questions are well-defined. Hypotheses draw on prior knowledge.	Questions are well-defined and pertinent for the topic. Hypotheses draw on prior knowledge and have clear explanations for why they are expected.	2 points
Data Information	Information about the data collection process is missing key details or is inaccurate. The limitations of the data are missing or not realistic.	Information about the data collection process is accurate, but missing some minor details. The limitations of the data are explained.	Information about the data collection process is complete and accurate. Underlying problems and limitations of the data are explained. Use of these data to answer the project questions is justified.	3 points
Data Processing	The student has made errors in processing their data. The student is missing steps.	The student has processed the data correctly. Steps for obtaining, loading, cleaning, and analyzing the data are well-defined.	The student has processed the data correctly and taken precautions to ensure that their results are appropriate. Steps for obtaining, loading, cleaning, and analyzing the data are well-defined.	3 points
Results	Results of the project do not attempt to answer the scientific questions. The data visualizations are not relevant.	Results of the project somewhat answer the scientific questions. Data visualizations are mostly appropriate for the data.	Results of the project answer, or earnestly attempt to answer, the scientific questions. Data visualizations are entirely appropriate for the data.	3 points
Presentation Skills				
Organization	The presentation is not in a logical order and the student makes no effort to guide the audience.	The presentation is organized in a logical order and takes some care to guide the audience.	The presentation is organized in a logical order and shows exceptional attention to guiding the audience.	2 points
Timing	The student far exceeds their allotted time and/or has not made an effort to practice.	The student completes the presentation in somewhat over 5 minutes.	The student completes the presentation within 5 minutes and it is clear that they have practiced.	1 point

Explanation of Ideas / Information	The ideas and information explained in the presentation were not clear and were not relevant.	The ideas and information explained in the presentation were clear and relevant.	The ideas and information explained in the presentation were exceptionally clear, relevant, and coherent.	3 points	Present-ation: 20 points
Code					
Correctness	The student misuses code and does not produce reasonable results.	The student uses some coding techniques/tools learned throughout the quarter. The analysis produces reasonable answers that can be replicated with some effort.	The student properly and efficiently uses the coding techniques/tools learned throughout the quarter. The analysis produces reasonable answers that can be replicated easily.	8 points	
Functionality	The code does not run and has egregious errors.	The code is mostly able to run, but has some (small) errors.	The code runs efficiently with no errors.	5 points	
Tidiness	The code breaks proper etiquette and should not be shared with others.	The code mostly follows proper coding etiquette. The organization is somewhat lacking and would need review before sharing.	The code follows proper coding etiquette. It is organized and commented effectively so that it can easily be shared with another person.	6 points	
Perseverance	The student has made no effort to work through problems and hurdles.	The student has made some effort to work through problems.	The student has made a gallant effort to work through problems and documented in their code their best understanding of the problems they are facing.	5 points	
Plots					
Plot Clarity	The plots are unclear and do not make sense in the context of the project.	The plots are mostly clear and show some thought from the students about ways to present their data.	The plots are extremely clear and are effective tools to help the audience understand the results/analysis.	5 points	
Colormaps	The colormaps are not appropriate for the data being shown.	The colormaps are appropriate for the data being shown.	The colormaps are appropriate for the data being shown and take into account colorblindness, and perceptual accuracy.	3 points	
Proper Labels	The plots are missing most/all labels or have improper labels.	The plots are labeled with general accuracy and completion.	The plots are labeled extremely accurately in a way that guides the audience through the figure.	5 points	
Creativity	The student made no effort to create original plots.	The student has made some effort to create original plots.	The student has created original plots that show the data/analysis in an extremely effective manner.	3 points	Code: 40 points

Supplemental References

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory Into Practice*, *41*(4), 212–219.