AGU ADVANCING EARTH AND SPACE SCIENCE

Earth and Space Science

TECHNICAL REPORTS: METHODS
10.1029/2021EA001896

Labeling Poststorm Coastal Imagery for Machine Learning: Measurement of Interrater Agreement

Evan B. Goldstein[1], Daniel Buscombe[2], Eli D. Lazarus[3], Somya D. Mohanty[4], Shah Nafis Rafique[4], Katherine A. Anarde[5], Andrew D. Ashton[6], Tomas Beuzen[7], Katherine A. Castagno[8], Nicholas Cohn[9], Matthew P. Conlin[10], Ashley Ellenson[11], Megan Gillen[12], Paige A. Hovenga[11], Jin-Si R. Over[13], Rose V. Palermo[12], Katherine M. Ratliff[14], Ian R. B. Reeves[15], Lily H. Sanborn[12], Jessamin A. Straub[9], Luke A. Taylor[3], Elizabeth J. Wallace[16], Jonathan Warrick[17], Phillipe Wernette[17], and Hannah E. Williams[18]

[1]Department of Geography, Environment, and Sustainability, University of North Carolina at Greensboro, Greensboro, NC, USA, [2]Marda Science LLC, Contracted to USGS Pacific Coastal and Marine Science Center, Santa Cruz, CA, USA, [3]Environmental Dynamics Lab, School of Geography and Environmental Science, University of Southampton, Southampton, UK, [4]Department of Computer Science, University of North Carolina at Greensboro, Greensboro, NC, USA, [5]Department of Civil, Construction, and Environmental Engineering, North Carolina State University, Raleigh, NC, USA, [6]Department of Geology and Geophysics, Woods Hole Oceanographic Institution, Woods Hole, MA, USA, [7]Department of Statistics, University of British Columbia, Vancouver, BC, Canada, [8]Center for Coastal Studies, Provincetown, MA, USA, [9]U.S. Army Engineer Research and Development Center, Field Research Facility, Duck, NC, USA, [10]Department of Geological Sciences, University of Florida, Gainesville, FL, USA, [11]College of Engineering, Oregon State University, Corvallis, OR, USA, [12]MIT-WHOI Joint Program in Oceanography/Applied Ocean Science & Engineering, Cambridge and Woods Hole, MA, USA, [13]U.S. Geological Survey, Coastal and Marine Science Center, Woods Hole, MA, USA, [14]Earth and Ocean Sciences, Duke University, Durham, NC, USA, [15]Department of Geological Sciences, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA, [16]Department Earth, Environmental, and Planetary Sciences, Rice University, Houston, TX, USA, [17]U.S. Geological Survey, Pacific Coastal and Marine Science Center, Santa Cruz, CA, USA, [18]Water Engineering, Faculty of Engineering and Physical Sciences, University of Southampton, Southampton, UK

Key Points:

- We measure agreement among coastal scientists labeling the same sets of poststorm images
- Coastal scientists agree more when rating landforms, less when labeling inferred processes
- Iterating on questions, providing documentation, and using smaller image sizes all increase agreement

Correspondence to:
E. B. Goldstein,
ebgoldst@uncg.edu;
@ebgoldstein

Citation:
Goldstein, E. B., Buscombe, D., Lazarus, E. D., Mohanty, S. D., Rafique, S. N., Anarde, K. A., et al. (2021). Labeling poststorm coastal imagery for machine learning: Measurement of interrater agreement. *Earth and Space Science*, *8*, e2021EA001896. https://doi.org/10.1029/2021EA001896

Received 26 JUN 2021
Accepted 26 AUG 2021

Abstract

Classifying images using supervised machine learning (ML) relies on labeled training data—classes or text descriptions, for example, associated with each image. Data-driven models are only as good as the data used for training, and this points to the importance of high-quality labeled data for developing a ML model that has predictive skill. Labeling data is typically a time-consuming, manual process. Here, we investigate the process of labeling data, with a specific focus on coastal aerial imagery captured in the wake of hurricanes that affected the Atlantic and Gulf Coasts of the United States. The imagery data set is a rich observational record of storm impacts and coastal change, but the imagery requires labeling to render that information accessible. We created an online interface that served labelers a stream of images and a fixed set of questions. A total of 1,600 images were labeled by at least two or as many as seven coastal scientists. We used the resulting data set to investigate interrater agreement: the extent to which labelers labeled each image similarly. Interrater agreement scores, assessed with percent agreement and Krippendorff's alpha, are higher when the questions posed to labelers are relatively simple, when the labelers are provided with a user manual, and when images are smaller. Experiments in interrater agreement point toward the benefit of multiple labelers for understanding the uncertainty in labeling data for machine learning research.

Plain Language Summary

After hurricanes and storms, pictures taken from a plane can be used to observe how the coast was impacted. A single flight might take thousands of pictures. If a computer could automatically analyze the pictures, then a person would not need to look at them one-by-one. To teach a computer to analyze images, we need many pictures and many labels that describe what is visible in each picture. But where do we get those labels? Typically, a coastal scientist labels the pictures by sorting them into folders or typing codes into a spreadsheet. But does every coastal scientist label pictures the same way? Some labeling questions are easy to answer, and scientists mostly agree ("Is this image all water?"). Other labeling questions are harder to answer, and cause disagreement ("Was there damage to buildings?"). This paper is about how well scientists agree when labeling the same pictures,

and how we can improve agreement among scientists. We try some experiments and offer a few ideas on how to improve agreement. We suggest writing very clear questions, using smaller images, and having a comprehensive manual. It turns out that having a manual with examples—and reading the manual!—really helps.

1. Introduction

Successful image classification using supervised machine learning (ML) relies on the availability of many images and associated classifying labels (e.g., Sun et al., 2017). This requirement has led to the development of large labeled image data sets (e.g., ImageNet; Deng et al., 2009), which are used to train models for a variety of target tasks and as a ML benchmark. Given changing attitudes in the Earth and environmental sciences toward greater accountability and transparency to the public and scientific funders, data are increasingly being shared publicly (e.g., Stall et al., 2018, 2019). However, there is a paucity of imagery that is sufficiently labeled for machine learning applications, which has hindered its wider adoption. For example, there is a growing repository of coastal images from satellite (e.g., Landsat, Sentinel), aerial platforms (e.g., Madore et al., 2018), oblique cameras (e.g., Holman & Stanley, 2007), Structure-from-Motion camera systems (e.g., Conlin et al., 2018; Sherwood et al., 2018; Warrick et al., 2017), and community science (e.g., Harley et al., 2019). Labels or annotations are often added later, specific to a particular study. In some situations, labels can be assigned automatically—e.g., merging time-stamped images with time-stamped sensor data (e.g., Buscombe & Carini, 2019; Buscombe et al., 2020). But most of the time, labeling cannot be done programmatically and instead requires human interpretation (e.g., Ellenson et al., 2020; Liu et al., 2014; Buscombe & Ritchie, 2018; Morgan et al., 2019; Yang et al., 2021).

Manual labeling is time-consuming and tends to be regarded as unglamorous, despite the acknowledged importance among practitioners that labeled data has for supervised machine learning (e.g., Sambasivan et al., 2021). Publications and ML curriculum rarely discuss in detail the practice of data annotation and labeling and how to report these details (e.g., Geiger et al., 2020; Sambasivan et al., 2021). Therefore, important labeling details—such as how class sets were decided, what class sets were tried and discarded, and labeling errors—can go unreported. As a result, searching for guidance in the published literature typically does not help. A reader is often left to infer (rather than be explicitly told) the questions asked during labeling, what the possible answers were, whether examples were provided to labelers, which tools/software were used, how many people labeled each image, etc. (Geiger et al., 2020).

Information about the data-labeling process is critical for understanding how data sets used in ML are constructed, and for exploring ML model limitations and strengths. ML models can only be as accurate as the data used to train the model. That is, errors in input label data become irreducible model errors. Providing details on the annotation process helps researchers to understand and correct potential sources of error and biases that are otherwise inherited by further data analysis, and ultimately the ML model (e.g., Sambasivan et al., 2021). Recognizing the importance of understanding data provenance—both for the samples (e.g., images) and the labels—has led to a call for increasing care in annotation practices (Paullada et al., 2020) and more detailed metadata for ML data sets (e.g., Gebru et al., 2018; Holland et al., 2018).

Annotation commonly relies on just a single labeler developing a gold-standard data set. But as Aroyo and Welty (2015) pointed out, a single labeler for a given classification scheme is often not enough. Having multiple labelers is expensive (in time and money) but has many benefits (Aroyo & Welty, 2015). Beyond random error detection, such as an errant mouse click, disagreement among labelers can enable insight into the labeling practice (e.g., vagueness in the label definitions, subjective annotation categories, or questions), as well as insight into individual examples within the data set (e.g., cryptic features). Involvement of multiple labelers also allows for calculating interrater agreement (also referred to as interrater reliability) to quantitatively examine the level of agreement among several labelers, normalized by the agreement expected by chance (e.g., Cohen, 1960; Gwet, 2014; Hallgren, 2012).

We generated multiple labels for 1,600 aerial images of U.S. coastal environments after three hurricanes (Florence, 2018; Isaias, 2020; Michael, 2018) to explore the process of labeling imagery for ML applications in Earth and environmental science. All labels are released in Goldstein et al. (2021). The labels correspond to images from the large repository of Emergency Response Imagery (ERI) collected by the National

Geodetic Survey Remote Sensing Division of the National Oceanographic and Atmospheric Administration (National Geodetic Survey, 2021). This imagery aids in recovery efforts, and the rapid assessment of the physical, ecological, economic, and societal impacts of storms along coastlines and among coastal communities (e.g., Madore et al., 2018; Overbeck et al., 2015). For each image, we posed a fixed set of questions regarding visible storm impacts, and each image was labeled by between two and seven coastal scientists from a collective group of 22 coastal scientists familiar with poststorm imagery.

The goals of this study were to critically examine the labeling process for Earth and environmental science imagery, discover baseline interrater agreement scores for the resulting labels, uncover potential pitfalls during the labeling process, and suggest potential practices for future labeling efforts. Our effort here is focused exclusively on the data side of ML—investigating data labeling practices that are rarely discussed in papers and viewed as subordinate to work building ML models (e.g., Sambasivan et al., 2021). Toward this end, we created a custom annotation interface that runs in a web browser and conducted three labeling experiments. We developed a set of questions that explored evidence in the imagery, as perceived by the labeler, of storm-driven physical processes and storm impacts to the built environment. We also posed questions for general, multiuse image categorization unrelated to specific storms but specific to aerial poststorm imagery. First, participating coastal scientists labeled images through the annotation interface. Second, in an effort to increase agreement scores, we refined the questions posed to labelers and developed a detailed labeling manual of definitions and visual examples. We then asked participants to label a new bank of images, informed by the manual. Third, we quartered the original images, and participants labeled the smaller tiles individually. We then computed and compared interrater agreement scores for all three experiments. We use these labeling experiments and their resulting scores to quantitatively examine disagreement and its causes. Results from this study can inform future labeling work, which will continue to be necessary as ML use in the Earth and environmental science continues to rise (e.g., Goldstein et al., 2019; Karpante et al., 2018; Razavi et al., 2021; Yu & Ma, 2021). We conclude the paper with a discussion of future directions for labeling.

2. Methods

2.1. Imagery

We used poststorm aerial imagery from three hurricane events that impacted sections of the U.S. Atlantic and Gulf Coasts (Figure 1). Hurricane Florence (2018) formed on August 31 and dissipated on September 18 (Stewart & Berg, 2019); the imagery used in this study was taken on September 17 over coastal North Carolina (ERI flight 20180917a). Hurricane Michael (2018) formed on October 7 and dissipated on October 16 (Beven et al., 2019); we used imagery from October 11 over the Florida Panhandle (ERI flight 20181011a). Hurricane Isaias (2020) formed on July 30 and dissipated on August 5 (Latto et al., 2021); we used imagery from August 4 over the North and South Carolina coastlines (ERI flight 20200804a). All imagery is in the Joint Photographic Experts Group (JPEG) format [both Tagged Image File Format (TIFF) and JPEG imagery are provided by NOAA] and can be downloaded directly from NOAA (https://storms.ngs.noaa.gov; National Geodetic Survey, 2021) or using the package developed by Moretz et al. (2020).

The 1,600 specific images labeled in this study were selected through an existing active learning system for developing an ML model to automatically detect washover deposits—the sedimentary features left on land surfaces by elevated coastal water levels (e.g., Hudock et al., 2014; Lazarus, 2016; Lazarus et al., 2021; Price, 1947)—in unlabeled imagery from the three reconnaissance flights following Hurricanes Florence, Michael, and Isaias, respectively. Here, we briefly review the active learning system, but refer readers to Goldstein et al. (2020b) for more details. Initially, several of the coauthors labeled 388 random poststorm images from NOAA flight 20180917a after Hurricane Florence (Goldstein et al., 2020a). These 388 images were labeled using an interface similar to one described in this study but were not included in the 1,600 images analyzed here. Goldstein et al. (2020b) used these 388 labeled images (179 washover and 209 no washover) to train a model to detect washover presence. VGG16 (Simonyan & Zisserman, 2014) was used as the feature extractor—initialized with ImageNet weights—and joined to one fully connected layer (with 50% dropout). An additional 140 images from Hurricane Michael (70 washover, 70 no washover) were labeled by two coastal scientists and used to test the trained VGG16 model. The F1 score for the trained model was 0.92.

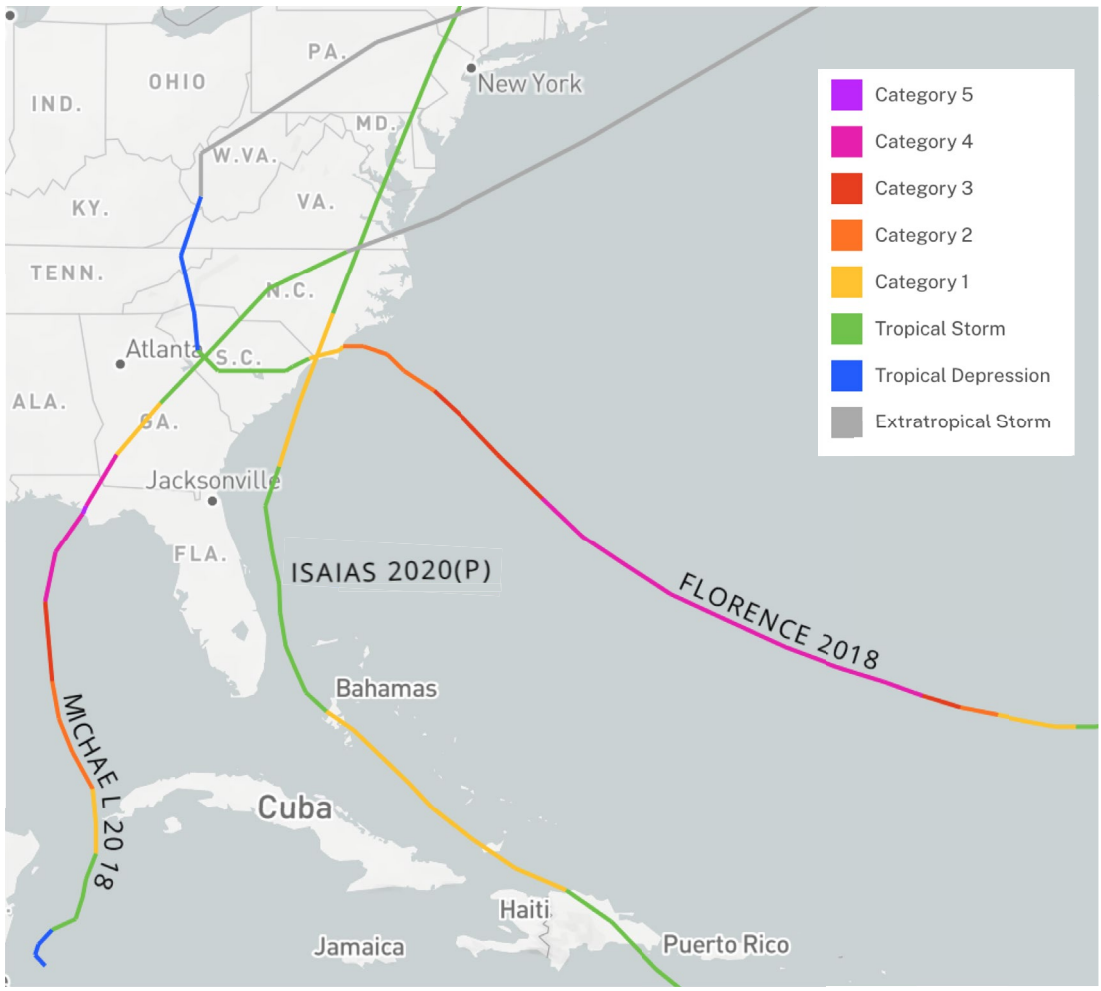

Figure 1. Storm tracks for Hurricane Florence (2018), Hurricane Michael (2018), and Hurricane Isaias (2020). Note the Isaias track is preliminary. Modified from https://coast.noaa.gov/hurricanes.

To complete the active learning process (e.g., Settles, 2011), the unlabeled images with highest model classification uncertainty were provided to human labelers for manual annotation. We repeated this active learning process several times to select the 1,600 images examined in this study. We acknowledge that this method of selecting images for manual labeling is imperfect because model classification uncertainty is computed as the distance of the model classification probabilities—$p(y|X,\theta)$ where y, X, and θ are labels, images, and model parameters, respectively—from a threshold. ML systems that make decisions based on a threshold (or maxima, or minima) are prone to amplifying biases (e.g., Peterson et al., 2019; Zhao et al., 2017). However, in our case, the model improved significantly upon each iteration as evidenced by training/validation metrics, and the use of interpretability techniques to visualize which parts of the image were important for detection (e.g., Selvaraju et al., 2017).

2.2. Labeling Interface

All image labeling was performed using a custom open-source interface (Figure 2; Rafique et al., 2021). Labelers accessed the site through a web browser, and all labelers answered the same set of questions for identical series of images (Table 1). The labeling interface had an option to display a full-size version of the image in a new tab, with zooming and panning enabled for closer inspection. The questions were designed to yield classification results for studies on washover, terrain analysis, storm impacts, and damage assessment. Questions were either formatted as radio buttons with two options (i.e., "yes" or "no") or checkboxes

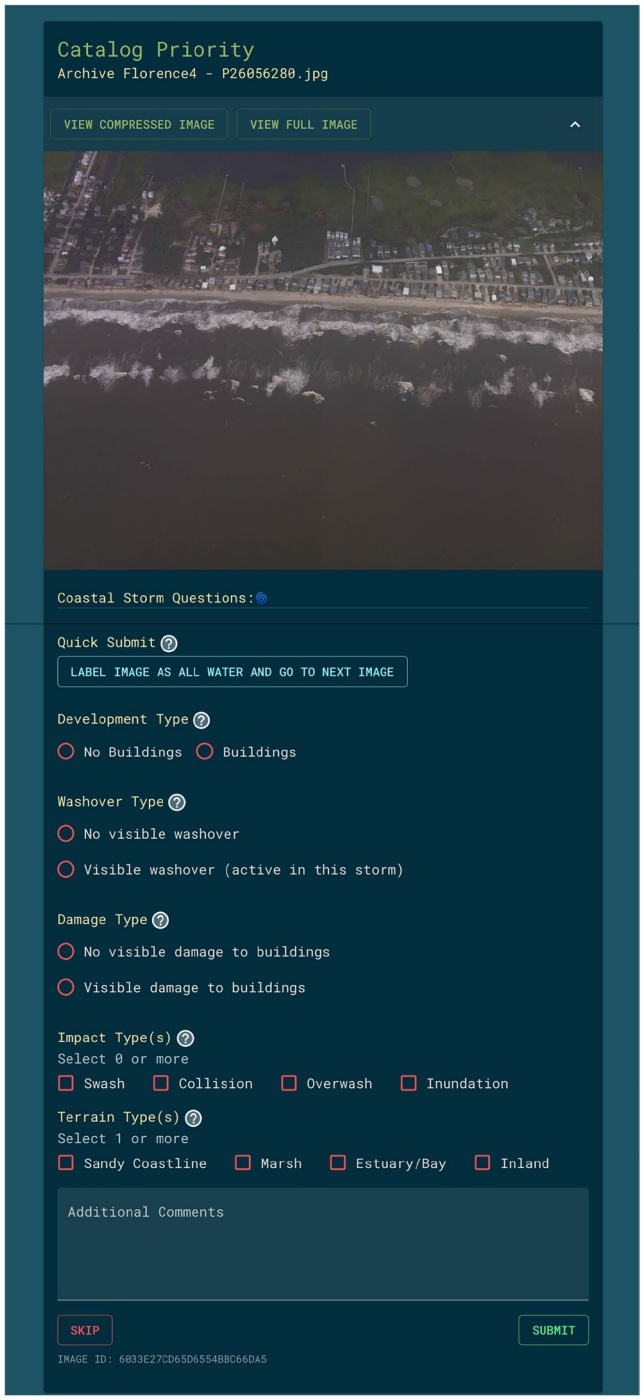

Figure 2. A screenshot of the labeler interface (Rafique et al., 2021) from a web browser showing an image and questions from Experiment 2.

(e.g., "check one or more"). Labelers were required to answer all questions (i.e., they could not advance to the next image unless the questions were complete) apart from two: a question that asked labelers to select the evident level of storm impact according to a well-known qualitative scale (Sallenger, 2000), and a free-text field in which to enter any additional notes or comments. The storm-impact scale question (Sallenger, 2000) is based on the elevation of the dune relative to the total water level during the storm: the swash regime if the total water level remains below the dune, the runup regime if the total water level exceeds the base of the dune, the overwash regime when the waves overtop the dune (or berm if there is no dune), and the inundation regime when the water level is high enough to entirely submerge a barrier island. The storm-impact question was not required because this question is not relevant to all images (e.g., inland images). A labeler could temporarily skip an image, but the skipped image would be appended to their image queue. Each labeler labeled groups of 100 images, and labeled each image only once.

2.3. Labeling Experiments

Here, we report the results of three labeling experiments, 1,600 images and 6,600 labels in total. Each image was labeled by between two and seven coauthors. All labelers have received university-level education and specialized in some aspect of coastal science, and some coauthors participated in more than a single experiment. Note that all labelers are coastal scientists, but their specific areas of expertize may not be interpretation of poststorm aerial images. For the first experiment, 13 coauthors labeled 900 images, yielding 4,500 labels. For the second experiment, 12 coauthors labeled 600 images, yielding 1,700 labels. Each image in the second experiment was labeled by between two and four coauthors.

Initially, coauthors were not issued detailed instructions prior to labeling—we refer to this phase as the first experiment. After receiving feedback on question ambiguity, we developed a second experiment that included slightly modified questions (Table 1) and a comprehensive instruction manual, which can be found in Rafique et al. (2021). At present, the instruction manual has three parts: first, a description of project goals; second, specific instructions for using the labeler (i.e., how to navigate the menu system); and third, a sequential discussion of each question, the answer options, the rationale behind that particular question, and some visual examples of a given classification. An excerpt from this third part of the manual, regarding the goal of the "washover type" question, reads as follows:

> *"Goal: Price (1947) coined the term "washover" to refer to sediment that is deposited beyond the beach as a result of elevated water levels. This deposit is the result of overwash (the process). Washover deposits can be deposited in a variety of identifiable contexts, such as beyond a berm into a marsh or developed area, or beyond a dune crest into the back dune area. Defining washover can be a difficult task that requires care and attention (and zooming in). Note that currently, finding and defining washover is the primary focus of the labeler.*
>
> *We are interested in finding washover deposits that were active recently, presumably in the storm that caused the flight to be commissioned by NOAA. So this question is asking: do you see washover deposits that you think were active during the past storm event? You will likely need to zoom in and inspect the*

Table 1
Questions in the Labeler Interface

Question Experiment 1	Type	Answer options	Modification for Experiments 2 and 3
Quick submit	Single button	"Label image as all water and Go to Next image" (this skips all subsequent questions and serves next image)	None
Development type	Radio buttons (2 options)	"Undeveloped; Developed"	Answer switched to "Buildings; No buildings"
Washover type	Radio buttons (2 options)	"Visible washover, No visible washover"	One answer option changed to:"Visible washover (active in this storm)"
Damage type	Radio buttons (2 options)	"No visible damage to infrastructure; Visible damage to infrastructure"	Answers changed to: "No visible damage to buildings; Visible damage to buildings"
Impact type(s)	Checkboxes (4 options, 0 required)	"Swash, Collision, Overwash, Inundation"	None
Terrain type(s)	Checkboxes (4 options, 1 required)	"Sandy coastline, Marsh, River, Inland"	Answers changed to: "Sandy coastline, Marsh, Estuary/Bay, Inland"

image closely. For example, by hitting those buttons above the image to look at the compressed version or full resolution version in new tabs.

When there is beachfront development (buildings) in the scene, beach-access points and spaces between buildings often allow for overwash and washover deposits to form.

We have observed that a new user can easily pick out large washover deposits. But smaller deposits are a bit harder to spot. Here are some examples of small, cryptic deposits just to show you some examples. To reiterate, you will likely need to zoom in to confirm the presence/absence of these deposits."

The NOAA images cover a large area, so many coastal features can appear in a single image and each feature can be relatively small (especially when the entire image is viewed). Labelers were prompted by the instructions to zoom in to examine images, but washover deposits and other impact features can be small and/or cryptic. To examine the effect of image size on labeler agreement, we performed a third experiment. We randomly selected 25 images that were previously labeled in Experiment 2. These images were split into quarters, and four labelers annotated all 100 image quarters. The four labelers in Experiment 3 did not necessarily label the original 25 images during Experiment 2.

2.4. Interrater Agreement Calculation

We computed the interrater agreement using both percent agreement—the percent of images where labelers agree on the assignment of a given label—and Krippendorff's alpha (Krippendorff, 1970), which is a numerical score that accounts for chance agreement, allows for any number of observers, allows for missing data, and is suitable for nominal, ordinal, interval, and ratio data (Hayes & Krippendorff, 2007). Alpha = 1 is perfect agreement, while alpha = 0 is no agreement between raters beyond chance; values less than zero indicate systematic lack of agreement (i.e., less than chance). Krippendorff's alpha is more stringent than computing the percent agreement, but less readily interpretable. Note that Krippendorff's alpha scores do not translate directly to percent agreement, so an alpha = 0.9 does not mean that raters agree for 90% of images and disagree for the other 10%. By providing both metrics we hope to give a richer understanding of the data. All data analysis was done with the R programming language (R Core Team, version 3.6.1, 2019) using RStudio (RStudio Team, 2021) and the IRR package (Gamer et al., 2019).

We used Krippendorff's alpha primarily for its ability to accommodate missing data and many labelers, which is advantageous compared to other classic metrics such as Cohen's kappa (and other derivatives; e.g., Hallgren, 2012). Krippendorff's alpha is computed as:

Table 2
Percent of Images From Each Experiment With At Least 1 "Yes" Label for the Category

Question	Experiment 1 (n = 900; %)	Experiment 2 (n = 600; %)	Experiment 3 subset (n = 25; %)	Experiment 3 quadrants (n = 100; %)
Buildings?	70	59	36	19
Washover?	46	73	64	19
Damage?	28	24	12	7
Swash?	54	45	40	29
Collision?	54	65	48	26
Overwash?	45	73	64	18
Inundation?	12	15	16	12
No impact?	77	27	44	80
All water?	10	11	16	42
Sandy coastline?	70	81	80	50
River?	25	1	—	—
Marsh?	52	62	—	32
Inland?	30	4	—	21
Bay/Estuary?	—	41	—	33

$$\alpha = 1 - \frac{D_o}{D_e}$$

where D_o is the observed disagreement and D_e is the expected disagreement due to chance. Observed and expected disagreement were computed by converting the observations into a reliability matrix and tabulating the coincidence matrices for all values; we refer the reader to Krippendorff (2011) for more details about the computation method.

Computing agreement for binary questions is more straightforward, but we asked three questions that allowed for multiple answers (i.e., multiple labels): impact type, terrain type, and storm-impact scale. We broke these questions into a series of binary questions. For example, if we marked images as having both "collision" and "overwash," then this was tabulated as being a "yes" in both the collision question and the overwash question, and a "no" for the Swash and Inundation questions.

3. Results

All images used in each experiment were labeled for multiple categories. The percent of images from each experiment with at least one "yes" label from a labeler, for each category, are shown in Table 2. We provide descriptive statistics for the 1,600 labeled images in Table 2 to report the distribution of categorical labels for the images in each dataset.

3.1. Interrater Agreement Results

The interrater agreement statistics for Experiments 1 and 2 can be seen in Table 3 and Figure 3. Recall that the key differences between Experiments 1 and 2 were the refinement of the questions posed to labelers, and the provision of a manual with examples. As a group, the questions regarding storm-impact scale had the lowest alpha (and agreement), while the questions regarding the presence of buildings, damage, sandy coastline, and "all water" had the highest agreement. Percent agreement and Krippendorff's alpha tended to increase for all questions in Experiment 2, except for the storm-impact questions regarding washover (decrease in alpha and % agreement), inundation (decrease in alpha and % agreement), and swash (decrease in alpha), and the question regarding the presence of marsh (decrease in alpha).

Table 3
Interrater Agreement Results for Experiments 1 and 2

Question	Experiment 1: percent agreement (%)	Experiment 1: Krippendorff's alpha	Experiment 2: percent agreement (%)	Experiment 2: Krippendorff's alpha
Buildings?	81	0.78	92	0.89
Washover?	66	0.51	64	0.47
Damage?	78	0.48	88	0.75
Swash?	51	0.25	59	0.14
Collision?	51	0.26	54	0.31
Overwash?	60	0.37	63	0.45
Inundation?	88	0.09	86	0.13
No impact?	54	0.56	92	0.85
All water?	95	0.82	98	0.91
Sandy coastline?	89	0.87	95	0.89
River?	77	0.39	—	—
Marsh?	51	0.27	53	0.26
Inland?	84	0.74	97	0.45
Bay/Estuary?	—	—	68	0.39

In general, these results suggest that adjusting the questions (e.g., buildings, damage) and providing an instruction manual tended to raise agreement scores. Some questions still had low scores, despite greater detail in the user manual and the inclusion of examples (e.g., washover).

The interrater agreement statistics for Experiment 3 can be seen in Table 4 and Figure 4. Agreement scores tended to rise when using quadrants relative to full-size tiles, except for damage (decrease in alpha), no impact (decrease in % agreement), inundation (decrease in alpha and % agreement), sandy coastline (decrease in % agreement), and "all water" (decrease in % agreement). We admit that this is a small experiment, but the results suggest that scores improve when using imagery that is small enough—and therefore manageable enough—that labelers do not need to zoom in to see fine-scale features.

3.2. Agreement and Disagreement

We can further investigate labeler agreement to gain intuition about Krippendorff's alpha using the largest fully crossed group of images in the data set: a set of 300 images from Experiment 1, each checked by seven labelers. Each panel of Figure 5 corresponds to a different question and reports the number of images that received a given number of "yes" labels (with seven labelers, this ranges from 0 to 7). Complete agreement for a given label across the entire image bank would appear as counts in only the two end-member columns: a feature is either present in an image, and all seven labelers note it as such, or the feature is absent, and none of the seven labelers mark it as present. (Technically, were a particular feature evident in all 300 images, or absent from all 300 images, complete agreement could manifest as a count of 300 in either the 0 or 7 column.) Therefore, populations of counts in columns other than 0 and 7 reveal disagreement among raters. For example, the "washover?" question yielded 169 images for which all raters agreed that no washover was present (hence a count of 169 in the "0" column); 42 images were labeled by only one person as "yes, washover" (i.e., six people labeled these as "no washover"), yielding a count of 42 in the "1" column; 25 images were labeled by two people as "yes, washover" (i.e., five people labeled these as "no washover"), resulting in a count of 25 in the "2" column, and so on.

In Experiment 1, the question with the highest Krippendorff's alpha—alpha = 0.87 (Table 3)—addresses the presence of a sandy coastline. Out of 300 images, 248 have perfect agreement—107 images were labeled by all seven labelers as "no" (107 images in the 0 column, meaning complete agreement for those images no sandy coastline is evident) and 141 images were labeled as "yes" (141 images in the 7 column, meaning complete agreement for those images a sandy coastline is evident). Fifty-two images drew between 1 and

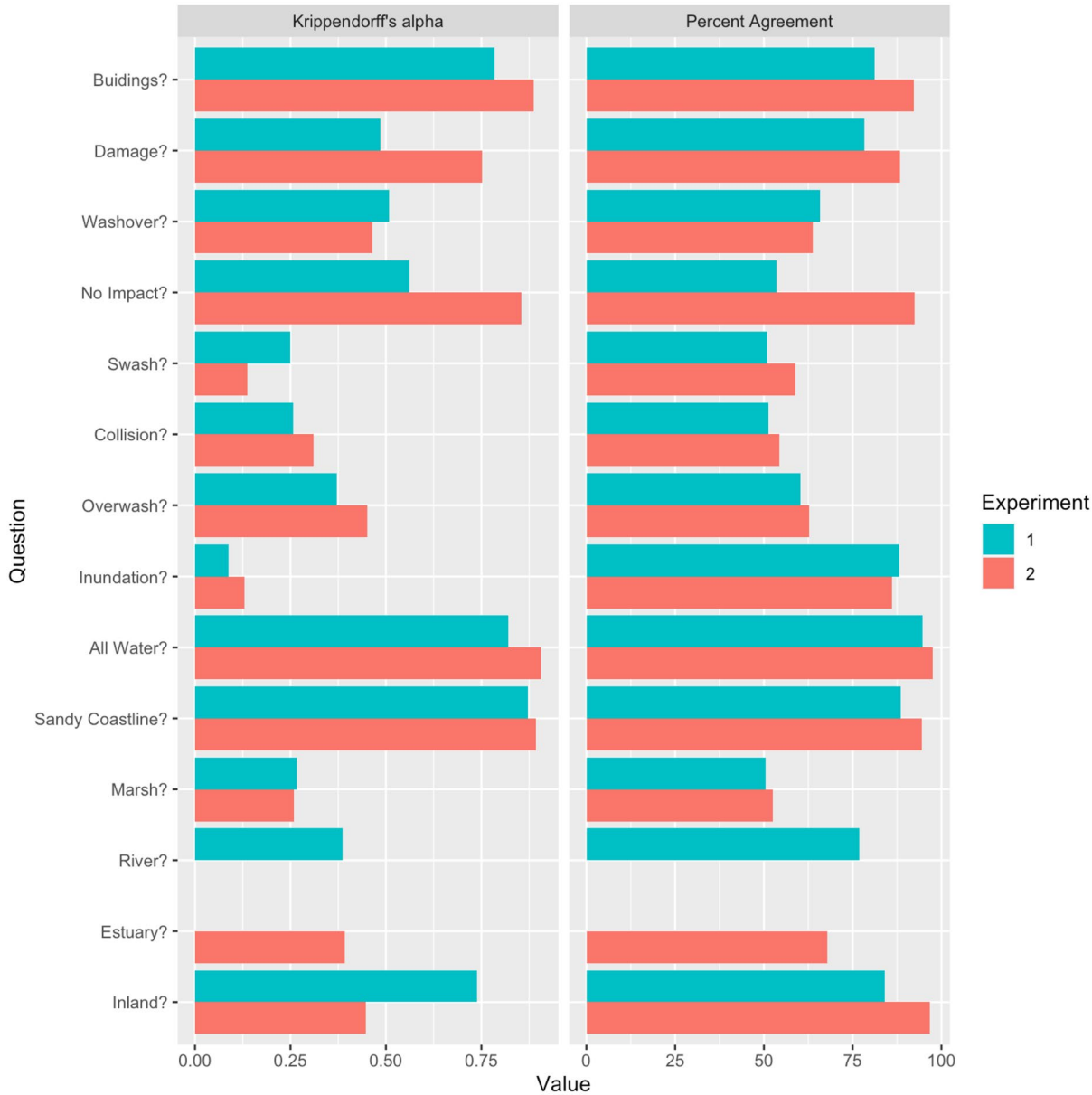

Figure 3. Krippendorff's alpha (left) and percent agreement (right) per question (rows) for Experiments 1 and 2 (shaded bars, as indicated by the legend).

6 labels for sandy coastlines. In this case, 83% of the images have perfect agreement for sandy coastlines, while 17% show disagreement, from minor (one labeler disagreeing) to more major (half of labelers disagreeing with the other half).

We examine agreement further by looking at specific images. In Figure 6, we highlight two consensus examples from Experiment 1: an image where all seven labelers answered "no" to the washover question, and an image where all seven labelers answered "yes" to the washover question. These examples caused no disagreement among labelers. In contrast, Figure 7 highlights six images that resulted in a range of responses—none of these six images received a unanimous (consensus) score. Upon detailed inspection, all of these images can be interpreted as having features that look like washover deposits, but may not have been labeled as such because of the cryptic nature of the deposit, ambiguity in the question (such as the lack of a manual during Experiment 1), lack of zooming, operator or interface error (errant mouse click, click not registering), or perhaps other reasons.

Table 4
Interater Agreement Results for Experiment 3

Question	Experiment 3 subset: percent agreement (%)	Experiment 3 subset: Krippendorff's alpha	Experiment 3 subset quads: percent agreement (%)	Experiment 3 subset quads: Krippendorff's alpha
Buildings?	88	0.81	95	0.90
Washover?	64	0.44	86	0.61
Damage?	96	0.86	96	0.75
Swash?	64	0.04	73	0.38
Collision?	60	0.29	75	0.32
Overwash?	56	0.33	85	0.52
Inundation?	92	0.69	88	0.20
No impact?	84	0.77	83	0.77
All water?	92	0.75	90	0.89
Sandy coastline?	84	0.74	82	0.80
River?	—	—	—	—
Marsh?	—	—	72	0.49
Inland?	—	—	84	0.54
Bay/Estuary?	—	—	70	0.43

4. Discussion

We have presented metrics for interrater agreement but have thus far not provided any interpretation of the metrics beyond noting that values closer to 1 (Krippendorff's alpha) and 100% (percent agreement) are better. We are unaware of a standard, justified interpretation of interrater agreement (e.g., McHugh, 2012; Monarch, 2021). Krippendorff (1970) pointed to values above 0.8 being ideal and values between 0.67 and 0.8 allowing for tentative conclusions to be drawn, which are often suggested as the guidelines for general application of alpha scores (e.g., Monarch, 2021). Landis and Koch (1977), who examine Kappa statistics that have the same range and align with Krippendorff's alpha for more restrictive cases of no missing data (e.g., Zapf et al., 2016), provide, in their words, "arbitrary" divisions: 0 or below is poor agreement, 0–0.2 is slight, 0.2–0.4 is fair, 0.4–0.6 is moderate, 0.6–0.8 is substantial, and 0.8–1 is almost perfect.

Defining what is acceptable agreement likely depends on the task and the domain. Furthermore, disagreement is likely inevitable and not a bad thing—initial disagreement is an opportunity, as Aroyo and Welty (2015) mention, to help refine any study through insight into the features and the questions posed. The multilabeler classification experiment we present here comprises a mix of easy questions and difficult questions. A key benefit of this multiquestion study design is that it allowed us to compare the difficulty of questions. Interrater agreement convolves information about the study questions and the labelers' collective, agreed-upon understanding of the words used in the questions. Examining agreement allowed us to understand which questions needed to be clarified and also to understand which classes will likely be difficult for future ML. As a result, measuring agreement and adjusting the labeling process in response to the agreement scores will ultimately lead to this data set being more useful.

4.1. Improving Agreement

Through this study we have gained insight into how to improve agreement. First, many labelers remarked about the vagueness of some questions. For example, in Experiment 1, we asked about development and infrastructure. These terms are vague and need to be well-defined. In Experiment 2, we refined these questions to ask specifically about buildings. The instruction manual also provided labelers with clarifications of these definitions through text and picture examples. As a result, interrater agreement for these questions improved.

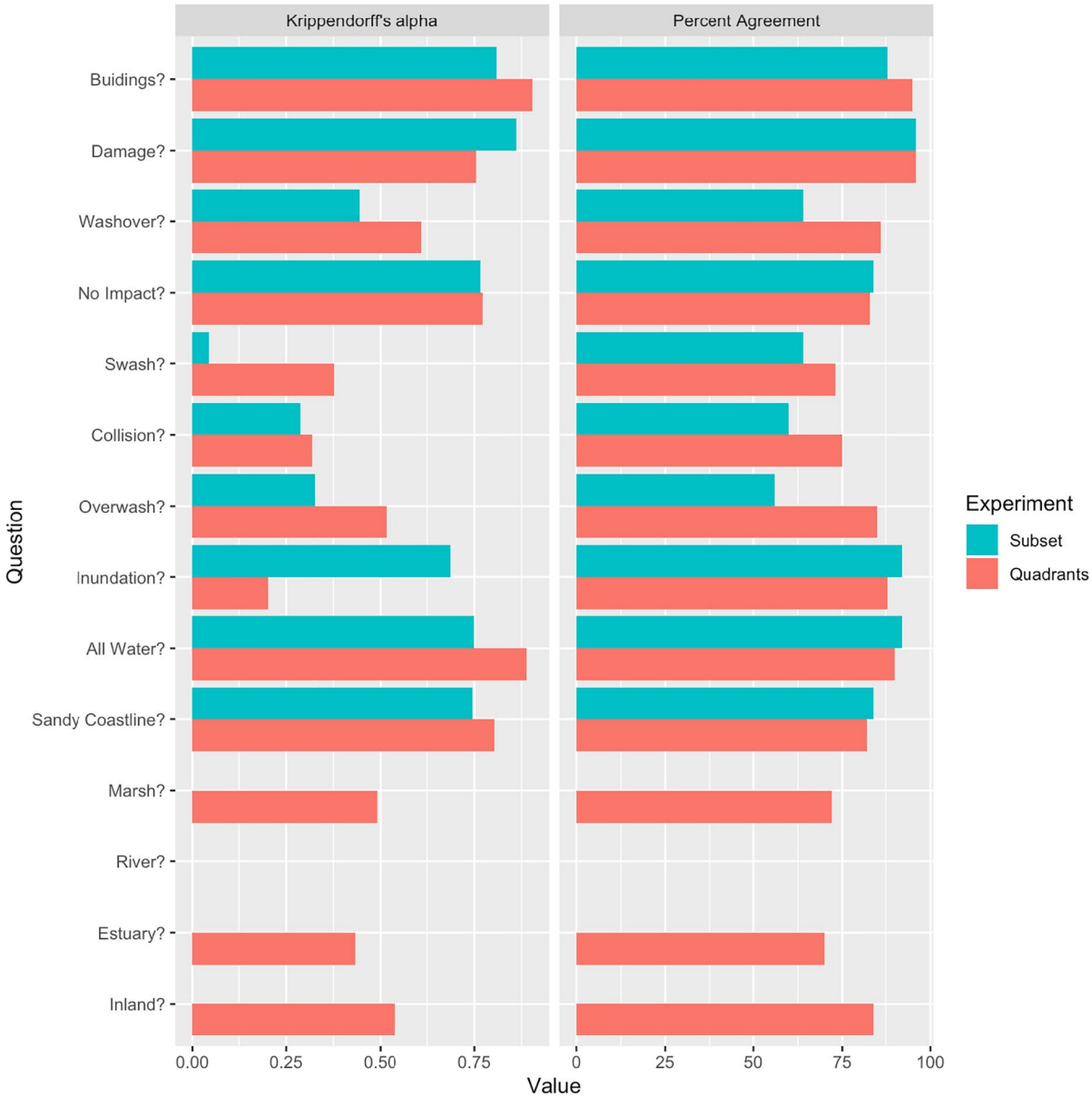

Figure 4. Krippendorff's alpha (left) and percent agreement (right) per question (rows) for Experiment 3 (shaded bars, as indicated by the legend).

Even with better clarity to the questions and a user manual, there was still disagreement among labelers on questions that seem straightforward, such as whether an image shows buildings, or is all water. The agreement metrics provided for these categories provide a good baseline for what to expect for seemingly straightforward questions—percent agreement of 92% and 98%, alpha of 0.89 and 0.91, respectively (metrics from Experiment 2). The lack of perfect agreement for these two categories indicates how a small sliver of land in the corner of an image with water can be missed or confuse a labeler and how the color and texture of a building roof can be cryptic.

Detecting some features in the image can also be difficult because of its vantage—especially small features that do not require large spatial context for identification. Building damage, for example, may be less obvious from a mostly overhead perspective. To mitigate this issue, the instruction manual cued labelers to look for debris around houses, blue tarps over roofs, and/or flooding that might impact houses. Instructions that list specific aspects of the image to look for appear to have resulted in increased agreement.

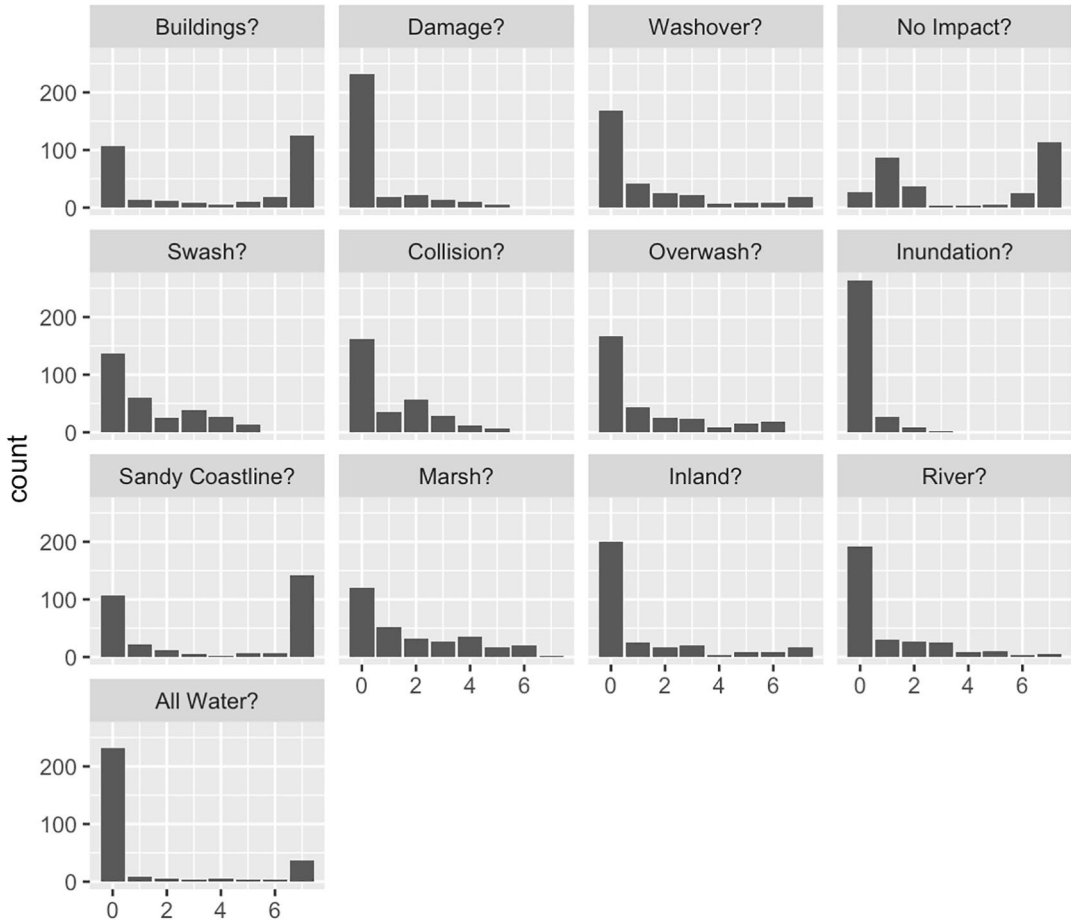

Figure 5. Results from the 300 images in Experiment 1 labeled by seven people. Each panel corresponds to a question and reports the number of images that received a given number of "yes" labels—with seven labelers, this ranges from 0 to 7, corresponding to 0/7 "yes" labels and 7/7 "yes" labels. Disagreement among labelers is seen when populations exist in columns other than 0 and 7.

In general, detecting visible features—buildings, sandy coastlines, washover deposits—had higher agreement than inferring the physical processes that previously occurred in the scene (e.g., swash, collision, inundation). For example, it is difficult to determine from an aerial image if dunes were just recently eroded, or if a region was inundated in the past several days. Occasionally there are clues—visible scarps, flooded areas, sand deposits that stretch from beach to bay—but these are not always present. Compounding this problem, the implementation of the Sallenger (2000) storm-impact regime question was also poorly

Figure 6. Examples of unambiguous "no washover" (0 labels; left) and unambiguous "washover" (7 labels; right) from NOAA ERI imagery.

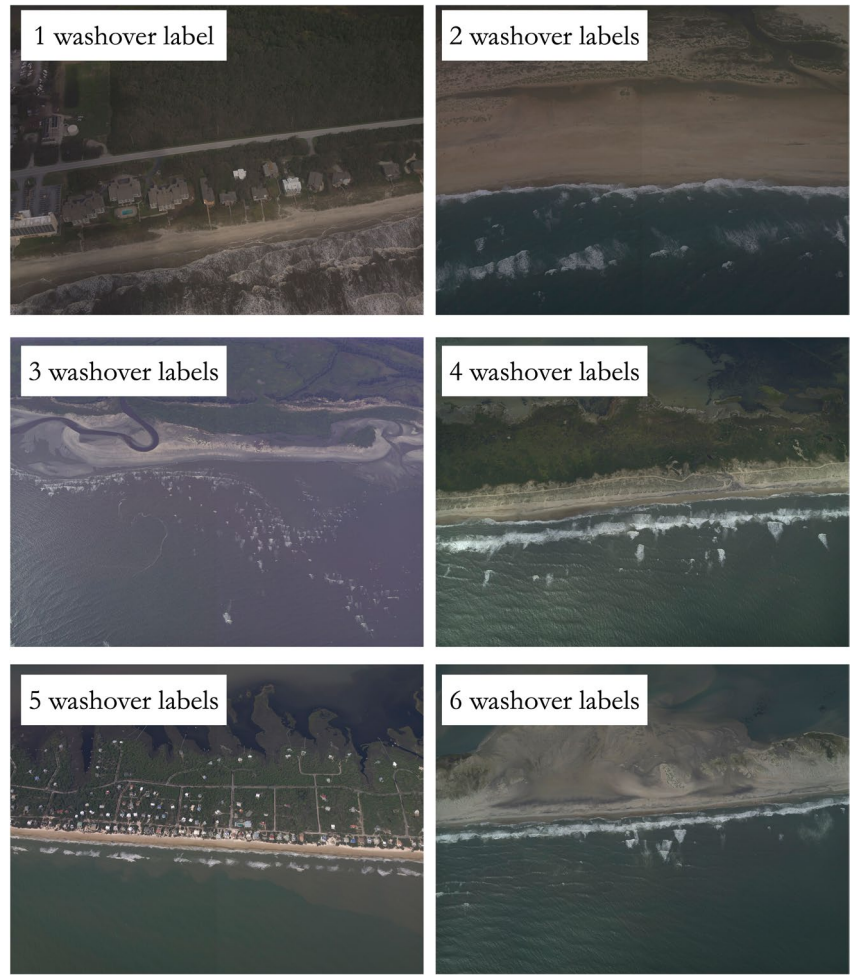

Figure 7. Examples of disagreement for the washover question from NOAA ERI imagery. Each image received a differing number of labels (out of a possible 7) indicating the presence washover.

designed, which is especially evident in the results of Experiment 1: this question did not require an answer before the labeler could advance to the next image and many labelers naturally saw this as an opportunity to pass over a difficult question. Originally, we made this choice because the reconnaissance images include a variety of settings, including inland and river scenes, where the question of the Sallenger (2000) storm-impact scale is not necessarily relevant. But this optionality allowed labelers to skip the impact question even where the question was relevant, such as an image of a sandy coastline. We encoded all skipped answers as "no impact"—i.e., a labeler skipping the impact question registered a label for "no impact." This tended to reduce alpha for the "no impact" in Experiment 1 for which there was no manual to explain that this question was not optional and to give guidance on how to answer the question. Even with a user manual with explanation and examples, this question was perhaps too difficult or too subjective. The interrelated nature of storm impacts also presents a problem with this question—e.g., the Sallenger (2000) scale is based on the elevation of the total water level relative to the dune crest, so if there was overwash, collision/swash must have previously happened at some point in the storm. As a result, agreement scores were low for these images, and labelers expressed reluctance to assign a past process to an image. This points to future work, where instead of asking about processes, we might ask only about features in an image and then infer process from the answers. For example, the agreement scores for "washover visible in image" were consistently higher than the process "overwash." So one could imagine asking if an image contained a recently active dune scarp instead of asking if collision had occurred. Alternate storm-impact scales could also be used to refine questions (e.g., Leaman et al., 2020), as well as the inclusion of different impact features (e.g., Over et al., 2021).

The images we used were very large, thousands of pixels in each dimension. Consequently, small features spanning only a few tens of pixels (such as individual buildings) can be easy to miss. In Experiment 3, where we used image quadrants, labelers remarked that small washover deposits and blue roof tarps (indicative of storm damage) were easier to identify and did not require labelers to zoom/pan. From an ML perspective, smaller imagery can be advantageous—models can be trained with images at native resolution (the full NOAA images require downsizing to avoid memory constraints). The major disadvantages to using smaller imagery, for both labeling and modeling, are labeling time and image context. Using image quadrants leads to an increase in the time and button clicks required to classify the same amount of area covered by the full-size labeled images. Some larger features such as washover deposits can only be identified with larger context, and some buildings only appear damaged when viewed in the context of buildings in adjacent areas. Ultimately, we chose to label the majority of imagery at the native size and resolution for three reasons: (a) to avoid further image processing tasks such as splitting and deciding on a common size for all images that is amenable to all questions; (b) to keep the number of individual images to label manageable; and (c) to avoid the disadvantages over the use of small tiles described above. As a result, in the instruction manual, we encourage annotators to zoom in and look for specific features. Future iterations of the labeling interface could focus on an interactive zoom mechanism that would permit users to look for detail in the image while still retaining the context that the full image provides.

4.2. Reconciling Disagreements

We did not go through the process of reconciling label disagreements for this study; instead, we released all the labels (Goldstein et al., 2021). We suggest four potential methods for reconciliation. First, class labels could be based on simple majority. Second, examples with conflicting labels could be relabeled or removed to develop a standard specific to the study at hand. A third option for reconciliation would be to use a model strategy where the label disagreement was incorporated in the model training, such as modification to a loss function or some other mechanism (e.g., Shin et al., 2020). Lastly, a separate rational strategy could be developed and implemented: e.g., for work focused on the presence of washover we have visually inspected disagreement and found that the features can be cryptic—so if one person finds a washover, the image likely has a washover (Figure 7), and a false positive rate would be demonstrably lower than a false negative rate. Hence one could use a rule that each image where at least one person finds washover is considered as an example of washover. This rule also highlights the tradeoff between accuracy and time—asking for labelers to zoom in (and expecting errors because of small features) versus asking labelers to look at more images if we were to split full-size images into tiles. Users of these image labels for ML should be aware of these tradeoffs during labeling. Combining multiple conflicting labels into a single class creates noisy or uncertain labels (e.g., an image with 3 “washover” labels, and 1 “no washover” label being labeled as “washover”). Labeling errors of this type are common in benchmark data sets (e.g., Northcutt et al., 2021) and models can still be trained to reach human performance even with label errors (e.g., He et al., 2015; Shankar et al., 2020). For example, Goldstein et al. (2020) trained a model for predicting washover deposits based on the rule presented above: if one person finds washover, the image likely has washover. Model results demonstrated that washover was able to be identified, suggesting that the noisy, uncertain labels from this data labeling exercise can be used to develop a model capable of generalizing.

4.3. Agreement, Disagreement, and Ground-Truth

We have thus far discussed labeler agreement and disagreement between labelers, but even perfect agreement is different from an actual ground-truth. Lack of reference data sets from the field is a common problem for remotely sensed Earth and environmental science data sets (e.g., Karpatne et al., 2018). As mentioned previously, for the questions that focus on features (e.g., buildings, sandy coastlines, washover deposits), some measure of agreement might be sufficient as a proxy for ground-truth in a given study. Using agreement as a proxy for ground-truth—trusting labels more when agreement is high—is less straightforward for questions focused on process (e.g., “was there inundation”). For example, as a worst-case scenario, this could lead to false positives where labelers all agree on the wrong label. Interpreted labels of processes could be more prone to disagreement and more prone to error—although we have no evidence that this is the case.

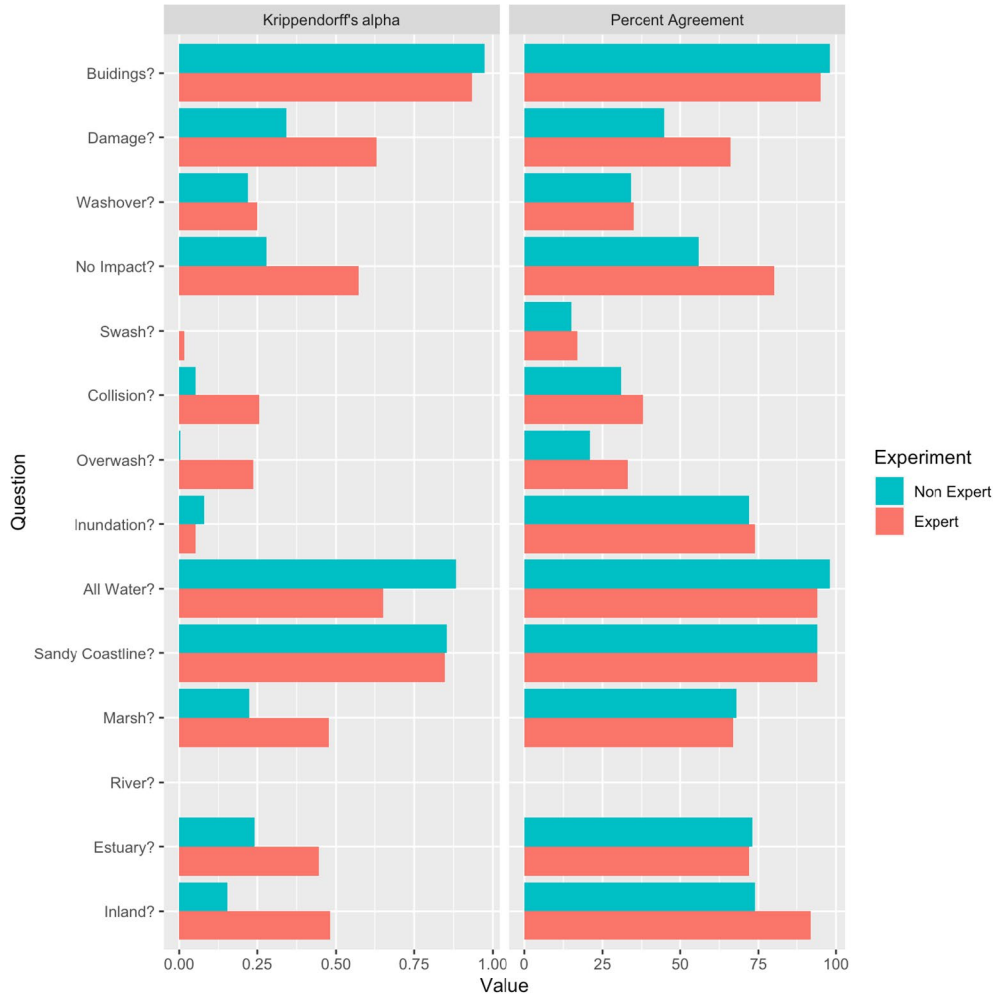

Figure 8. Krippendorff's alpha (left) and percent agreement (right) per question (rows) for noncoastal versus coastal labelers (shaded bars, as indicated by the legend).

However, labels of process are potentially more useful to researchers as they could be used as a proxy for ground-truth observations to test numerical models of storm impacts.

One option is to try to fuse these images with other existing data sources to develop ground-truth data sets. For example, linking images with colocated field observations (e.g., Anarde et al., 2020; Coogan et al., 2019; Reeves et al., 2021), numerical model experiments (e.g., Gharagozlou et al., 2020), poststorm damage surveys (e.g., Kennedy et al., 2020; Zhai & Peng, 2020), or social media messages (Mohanty et al., 2021). Ground-truth data can be used to assess whether the human labels are able to identify past processes from images.

4.4. Nonexpert Labelers

Thus far we have relied on trained coastal experts to label images. As a pilot study, four people who are not trained as coastal scientists labeled 100 images that were previously labeled by four coastal scientists as part of Experiment 2 (i.e., with the use of an instruction manual). The interrater agreement statistics for this pilot study can be seen in Figure 8. Agreement was lower for experts when asked about the presence of buildings, but higher for experts when asked about building damage and the presence of washover. The storm-impact scale questions again had the lowest overall agreement, with higher agreement scores for experts (except alpha for inundation). Questions focused on the landscape yielded mixed results, with most scores equal in percent agreement between experts and nonexperts. Alpha scores for landscape questions

were mixed—experts' scores were higher for some questions (marshes, estuary, and inland), worse for the "all water" question, and identical to nonexperts for "sandy coastlines." It is not possible to be conclusive with the results from this pilot of 100 images. These results do suggest that—with a comprehensive manual—it might be possible to obtain good results from nonexperts for labeling poststorm images.

4.5. Next Steps for Labeling Studies in Earth and Environmental Science

Labeling tasks will likely continue to be necessary in tandem with increased adoption of ML use for Earth and environmental science studies (e.g., Goldstein et al., 2019; Karpante et al., 2018; Razavi et al., 2021; Yu & Ma, 2021). Correctly labeled data are also critical for building trustworthy models, which is a current focus for environmental ML research (e.g., McGovern et al., 2020). We recommend four areas for future labeling research in Earth and environmental science, especially where labels are discrete and there is no objective gold standard against which to compare. First, we do not track the time each labeler takes to annotate an image, nor the button presses they make (if they choose to look at the image at full or compressed size in a different tab, if they zoom, etc.). The time it takes to label an image might correlate with agreement among labelers. Second, eye tracking could be employed to understand where users are looking (or not looking) to make decisions about annotations. Eye-tracking data could also be used in ML workflows (e.g., Karargyris et al., 2021) that could be adapted for image classification. Third, labelers could be prompted to identify the region within each image that was used as a basis for their decision. This could be done graphically (with a bounding box), via the text box, or by recording audio. Fourth, all labelers in this study are coastal scientists. Other labeling projects have had success using crowd-sourced labels from people who are not subject matter experts (e.g., Morgan et al., 2019; Whittemore et al., 2020). We have presented a pilot study with labels from nonexperts that resulted in metrics that are directly comparable with those from experts, but further work involving more people and more imagery is needed. Preconceptions may be different between experts and nonexperts, and a lack of preconceptions could sometimes be advantageous. Interrater agreement is a way in which we can assess what classes are identifiable by both experts and nonexperts, if expert labels are truly required for a task, or if a well-written instruction manual is enough to train nonexperts to label images correctly.

Data Availability Statement

All data used for this study are available on Zenodo via Goldstein et al. (2021). The code for the labeling interface is available on Zenodo and GitHub via Rafique et al. (2021). The code for the data analysis is available on Zenodo and GitHub via (Goldstein, 2021).

Acknowledgments
We thank the editor, two reviewers, and Chris Sherwood for feedback on this work. The authors gratefully acknowledge support from the U.S. Geological Survey (G20AC00403 to EBG and SDM), NSF (1953412 to EBG and SDM; 1939954 to EBG), Microsoft AI for Earth (to EBG and SDM), The Leverhulme Trust (RPG-2018-282 to EDL and EBG), and an Early Career Research Fellowship from the Gulf Research Program of the National Academies of Sciences, Engineering, and Medicine (to EBG). U.S. Geological Survey researchers (DB, J-SRO, JW, and PW) were supported by the U.S. Geological Survey Coastal and Marine Hazards and Resources Program as part of the response and recovery efforts under congressional appropriations through the Additional Supplemental Appropriations for Disaster Relief Act, 2019 (Public Law 116-20; 133 Stat. 871).

References

Anarde, K., Figlus, J., Sous, D., & Tissier, M. (2020). Transformation of infragravity waves during hurricane overwash. *Journal of Marine Science and Engineering*, *8*(8), 545. https://doi.org/10.3390/jmse8080545

Aroyo, L., & Welty, C. (2015). Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, *36*(1), 15–24. https://doi.org/10.1609/aimag.v36i1.2564

Beven, J. L. II, Berg, R., & Hagen, A. (2019). *National Hurricane Center Tropical Cyclone Report*, Tech. Rep. AL142018, pp. 86. National Hurricane Center. https://www.nhc.noaa.gov/data/tcr/AL142018_Michael.pdf

Buscombe, D., & Carini, R. J. (2019). A data-driven approach to classifying wave breaking in infrared imagery. *Remote Sensing*, *11*(7), 859. https://doi.org/10.3390/rs11070859

Buscombe, D., Carini, R. J., Harrison, S. R., Chickadel, C. C., & Warrick, J. A. (2020). Optical wave gauging using deep neural networks. *Coastal Engineering*, *155*, 103593. https://doi.org/10.1016/j.coastaleng.2019.103593

Buscombe, D., & Ritchie, A. C. (2018). Landscape classification with deep neural networks. *Geosciences*, *8*(7), 244. https://doi.org/10.3390/geosciences8070244

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Conlin, M., Cohn, N., & Ruggiero, P. (2018). A quantitative comparison of low-cost structure from motion (SfM) data collection platforms on beaches and dunes. *Journal of Coastal Research*, *34*(6), 1341–1357. https://doi.org/10.2112/JCOASTRES-D-17-00160.1

Coogan, J. S., Webb, B. M., Smallegan, S. M., & Puleo, J. A. (2019). Geomorphic changes measured on Dauphin Island, AL, during Hurricane Nate. *Shore and Beach*, *87*(4), 15. https://doi.org/10.34237/1008742

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 248–255). https://doi.org/10.1109/CVPR.2009.5206848

Ellenson, A. N., Simmons, J. A., Wilson, G. W., Hesser, T. J., & Splinter, K. D. (2020). Beach state recognition using argus imagery and convolutional neural networks. *Remote Sensing*, *12*(23), 3953. https://doi.org/10.3390/rs12233953

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2019). *IRR: Various coefficients of interrater reliability and agreement*. R package version 0.84.1. Retrieved from https://CRAN.R-project.org/package=irr

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé, H. III, & Crawford, K. (2018). *Datasheets for datasets*. arXiv:1803.09010. https://arxiv.org/abs/1803.09010

Geiger, R. S., Yu, K., Yang, Y., Dai, M., Qiu, J., Tang, R., & Huang, J. (2020). Garbage in, garbage out? Do machine learning application papers in social computing report where human-labeled training data comes from? In *FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 325–336). https://doi.org/10.1145/3351095.3372862

Gharagozlou, A., Dietrich, J. C., Karanci, A., Luettich, R. A., & Overton, M. F. (2020). Storm-driven erosion and inundation of barrier islands from dune-to region-scales. *Coastal Engineering*, *158*, 103674. https://doi.org/10.1016/j.coastaleng.2020.103674

Goldstein, E. B. (2021). Post-storm image label agreement code (v1.0). *Zenodo*. https://doi.org/10.5281/zenodo.5212158

Goldstein, E. B., Buscombe, D., Lazarus, E. D., Anarde, K., Ashton, A. D., Beuzen, T., et al. (2021). Labels for emergency response imagery from Hurricane Florence, Hurricane Michael, and Hurricane Isaias (Version 1.4) [Data set]. *Zenodo*, https://doi.org/10.5281/zenodo.5172799

Goldstein, E. B., Coco, G., & Plant, N. G. (2019). A review of machine learning applications to coastal sediment transport and morphodynamics. *Earth-Science Reviews*, *194*, 97–108. https://doi.org/10.1016/j.earscirev.2019.04.022

Goldstein, E. B., Lazarus, E., Beuzen, T., Williams, H., Limber, P., Cohn, N., et al. (2020a). *Labels for Hurricane Florence (2018) emergency response imagery from NOAA*. Figshare. https://doi.org/10.6084/m9.figshare.11604192.v1

Goldstein, E. B., Mohanty, S. D., Rafique, S. N., & Valentine, J. (2020b). An active learning pipeline to detect hurricane washover in post-storm aerial images. *AI for Earth Sciences Workshop at NeurIPS 2020*. https://doi.org/10.31223/X5JW23

Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, *8*(1), 23–34. https://doi.org/10.20982/tqmp.08.1.p023

Harley, M. D., Kinsela, M. A., Sanchez-Garcia, E., & Vos, K. (2019). Shoreline change mapping using crowd-sourced smartphone images. *Coastal Engineering*, *150*, 175–189. https://doi.org/10.1016/j.coastaleng.2019.04.003

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, *1*(1), 77–89. https://doi.org/10.1080/19312450709336664

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1026–1034). https://doi.org/10.1109/iccv.2015.123

Holland, S., Hosny, A., Newman, S., Joseph, J., & Chmielinski, K. (2018). *The dataset nutrition label: A framework to drive higher data quality standards*. arXiv:1805.03677. https://arxiv.org/abs/1805.03677

Holman, R. A., & Stanley, J. (2007). The history and technical capabilities of Argus. *Coastal Engineering*, *54*(6–7), 477–491. https://doi.org/10.1016/j.coastaleng.2007.01.003

Hudock, J. W., Flaig, P. P., & Wood, L. J. (2014). Washover fans: A modern geomorphologic analysis and proposed classification scheme to improve reservoir models washover fans: A modern geomorphologic analysis and classification. *Journal of Sedimentary Research*, *84*(10), 854–865. https://doi.org/10.2110/jsr.2014.64

Karargyris, A., Kashyap, S., Lourentzou, I., Wu, J. T., Sharma, A., Tong, M., et al. (2021). Creation and validation of a chest X-ray dataset with eye-tracking and report dictation for AI development. *Scientific Data*, *8*(1), 1–18. https://doi.org/10.1038/s41597-021-00863-5

Karpatne, A., Ebert-Uphoff, I., Ravela, S., Babaie, H. A., & Kumar, V. (2018). Machine learning for the geosciences: Challenges and opportunities. *IEEE Transactions on Knowledge and Data Engineering*, *31*(8), 1544–1554. https://doi.org/10.1109/TKDE.2018.2861006

Kennedy, A., Copp, A., Florence, M., Gradel, A., Gurley, K., Janssen, M., et al. (2020). Hurricane Michael in the area of Mexico Beach, Florida. *Journal of Waterway, Port, Coastal, and Ocean Engineering*, *146*(5), 05020004. https://doi.org/10.1061/(asce)ww.1943-5460.0000590

Krippendorff, K. (1970). Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, *30*(1), 61–70. https://doi.org/10.1177/2F001316447003000105

Krippendorff, K. (2011). *Computing Krippendorff's alpha-reliability*. Retrieved from https://repository.upenn.edu/asc_papers/43

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159–174. https://doi.org/10.2307/2529310

Latto, A., Hagen, A., & Berg, R. (2021). *National Hurricane Center Tropical Cyclone Report, Hurricane Isaias* (Tech. Rep. AL092020). National Hurricane Center. Retrieved from https://www.nhc.noaa.gov/data/tcr/AL092012_Isaac.pdf

Lazarus, E. D. (2016). Scaling laws for coastal overwash morphology. *Geophysical Research Letters*, *43*. https://doi.org/10.1002/2016GL071213

Lazarus, E. D., Goldstein, E. B., Taylor, L. A., & Williams, H. E. (2021). Comparing patterns of hurricane washover into built and unbuilt environments. *Earth's Future*, *9*, e2020EF001818. https://doi.org/10.1029/2020EF001818

Leaman, C. K., Harley, M. D., Splinter, K. D., Thran, M. C., Kinsela, M. A., & Turner, I. L. (2020). A storm hazard matrix combining coastal flooding and beach erosion. *EarthArXiv*. https://doi.org/10.31223/X5Q592

Liu, S. B., Poore, B. S., Snell, R. J., Goodman, A., Plant, N. G., Stockdon, H. F., et al. (2014). USGS iCoast—Did the coast change? Designing a crisis crowdsourcing app to validate coastal change models. In *CSCW Companion '14: Proceedings of the Companion Publication of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing* (pp. 17–20). https://doi.org/10.1145/2556420.2556790

Madore, B., Imahori, G., Kum, J., White, S., & Worthem, A. (2018). *NOAA's use of remote sensing technology and the coastal mapping program*. In Oceans 2018 MTS/IEEE Charleston (pp. 1–7). https://doi.org/10.1109/OCEANS.2018.8604932

McGovern, A., Bostrom, A., Ebert-Uphoff, I., He, R., Thorncroft, C., Tissot, P., et al. (2020). Weathering environmental change through advances in AI. *Eos, Transactions American Geophysical Union*, 101. https://doi.org/10.1029/2020EO147065

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, *22*(3), 276–282. https://doi.org/10.11613/bm.2012.031

Mohanty, S. D., Biggers, B., Sayedahmed, S., Pourebrahim, N., Goldstein, E. B., Bunch, R., et al. (2021). A multi-modal approach towards mining social media data during natural disasters—A case study of Hurricane Irma. *International Journal of Disaster Risk Reduction*, *54*, 102032. https://doi.org/10.1016/j.ijdrr.2020.102032

Monarch, R. (2021). Human-in-the-loop machine learning: Active learning and annotation for human-centered AI (pp. 1–456). New York, NY: Manning Publications.

Moretz, M. C., Foster, D., Weber, J., Chowdhury, R., Rafique, S. N., Goldstein, E. B., & Mohanty, S. D. (2020). psi-collect: A Python module for post-storm image collection and cataloging. *Journal of Open Source Software*, *5*(47), 2075. https://doi.org/10.21105/joss.02075

Morgan, K. L., Plant, N. G., Srockdon, H., & Snell, R. J. (2019). ICoast—Did the coast change? Storm-impact model verification using citizen scientists. In *Coastal sediments 2019* (pp. 1424–1438). World Scientific. https://doi.org/10.1142/9789811204487_0124

National Geodetic Survey. (2021). *Emergency response imagery*. Retrieved from https://storms.ngs.noaa.gov
Northcutt, C. G., Athalye, A., & Mueller, J. (2021). *Pervasive label errors in test sets destabilize machine learning benchmarks*. arXiv:2103.14749.
Over, J. R., Brown, J. A., Sherwood, C. R., Hegermiller, C., Wernette, P. A., Ritchie, A. C., & Warrick, J. A. (2021). A survey of storm-induced seaward-transport features observed during the 2019 and 2020 hurricane seasons. *Shore and Beach*, *89*(2), 31–40. https://doi.org/10.34237/1008924
Overbeck, J. R., Long, J. W., Stockdon, H. F., & Birchler, J. J. (2015). Enhancing evaluation of post-storm morphologic response using aerial orthoimagery from Hurricane Sandy. In *Coastal sediments 2015*. World Scientific. https://doi.org/10.1142/9789814689977_0250
Paullada, A., Raji, I. D., Bender, E. M., Denton, E., & Hanna, A. (2020). *Data and its (dis) contents: A survey of dataset development and use in machine learning research*. arXiv:2012.05345
Peterson, J. C., Battleday, R. M., Griffiths, T. L., & Russakovsky, O. (2019). Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 9617–9626). https://doi.org/10.1109/iccv.2019.00971
Price, W. A. (1947). Equilibrium of form and forces in tidal basins of coast of Texas and Louisiana. *AAPG Bulletin*, *31*(9), 1619–1663. https://doi.org/10.1306/3D933A3A-16B1-11D7-8645000102C1865D
Rafique, S. N., Goldstein, E. B., & Mohanty, S. D. (2021). Coastal image labeler: Production with Scribbler (Version 2.1). *Zenodo*. https://doi.org/10.5281/zenodo.4973582
Razavi, S. (2021). Deep learning, explained: Fundamentals, explainability, and bridgeability to process-based modelling. *ESSOAr*. https://doi.org/10.1002/essoar.10506045.1
R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/
Reeves, I. R. B., Goldstein, E. B., Anarde, K., & Moore, L. J. (2021). Remote bed-level change and overwash observation with low-cost ultrasonic distance sensors. *Shore and Beach*, *89*(2), 23–30. https://doi.org/10.34237/1008923
RStudio Team. (2021). *RStudio: Integrated Development Environment for R*. Boston, MA: RStudio, PBC. Retrieved from http://www.rstudio.com/
Sallenger, A. H. Jr. (2000). Storm impact scale for barrier islands. *Journal of Coastal Research*, *16*(3), 890–895. https://www.jstor.org/stable/4300099
Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P. K., & Aroyo, L. M. (2021). "Everyone wants to do the model work, not the data work": Data cascades in high-stakes AI. In *CHI '21: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Vol. *39*, pp. 1–15). https://doi.org/10.1145/3411764.3445518
Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 618–626). https://doi.org/10.1109/iccv.2017.74
Settles, B. (2011). From theories to queries: Active learning in practice. In *Active learning and experimental design workshop in conjunction with AISTATS 2010* (pp. 1–18). JMLR Workshop and Conference Proceedings.
Shankar, V., Roelofs, R., Mania, H., Fang, A., Recht, B., & Schmidt, L. (2020). Evaluating machine accuracy on imagenet. In *International Conference on Machine Learning* (pp. 8634–8644). PMLR.
Sherwood, C. R., Warrick, J. A., Hill, A. D., Ritchie, A. C., Andrews, B. D., & Plant, N. G. (2018). Rapid, remote assessment of hurricane Matthew impacts using four-dimensional structure-from-motion photogrammetry. *Journal of Coastal Research*, *34*(6), 1303–1316. https://doi.org/10.2112/JCOASTRES-D-18-00016.1
Shin, W., Ha, J. W., Li, S., Cho, Y., Song, H., & Kwon, S. (2020). *Which strategies matter for noisy label classification? Insight into loss and uncertainty*. arXiv:2008.06218.
Simonyan, K., & Zisserman, A. (2014). *Very deep convolutional networks for large-scale image recognition*. arXiv:1409.1556.
Stall, S., Yarmey, L. R., Boehm, R., Cousijn, H., Cruse, P., Cutcher-Gershenfeld, J., et al. (2018). Advancing FAIR data in Earth, space, and environmental science. *Eos, Transactions American Geophysical Union*, *99*. https://doi.org/10.1029/2018EO109301
Stall, S., Yarmey, L. R., Cutcher-Gershenfeld, J., Hanson, B., Lehnert, K., Nosek, B., et al. (2019). Make scientific data FAIR. *Nature*, *570*, 27–29. https://doi.org/10.1038/d41586-019-01720-7
Stewart, S. R., & Berg, R. (2019). *National Hurricane Center Tropical Cyclone Report: Hurricane Florence*. National Hurricane Center. Retrieved from https://www.nhc.noaa.gov/data/tcr/AL062018_Florence.pdf
Sun, C., Shrivastava, A., Singh, S., & Gupta, A. (2017). Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 843–852). https://doi.org/10.1109/iccv.2017.97
Warrick, J. A., Ritchie, A. C., Adelman, G., Adelman, K., & Limber, P. W. (2017). New techniques to measure cliff change from historical oblique aerial photographs and structure-from-motion photogrammetry. *Journal of Coastal Research*, *33*(1), 39–55. https://doi.org/10.2112/JCOASTRES-D-16-00095.1
Whittemore, A., Ross, M. R., Dolan, W., Langhorst, T., Yang, X., Pawar, S., et al. (2020). A participatory science approach to expanding instream infrastructure inventories. *Earth's Future*, *8*, e2020EF001558. https://doi.org/10.1029/2020EF001558
Yang, X., Pavelsky, T. M., Ross, M. R. V., Januchowski-Hartley, S. R., Dolan, W., Altenau, E. H., et al. (2021). Mapping flow-obstructing structures on global rivers. *ESSOAr*. https://doi.org/10.1002/essoar.10507070.1
Yu, S., & Ma, J. (2021). Deep learning for geophysics: Current and future trends. *Reviews of Geophysics*, *59*, e2021RG000742. https://doi.org/10.1029/2021RG000742
Zapf, A., Castell, S., Morawietz, L., & Karch, A. (2016). Measuring inter-rater reliability for nominal data—Which coefficients and confidence intervals are appropriate? *BMC Medical Research Methodology*, *16*, 93. https://doi.org/10.1186/s12874-016-0200-9
Zhai, W., & Peng, Z. R. (2020). Damage assessment using Google street view: Evidence from Hurricane Michael in Mexico Beach, Florida. *Applied Geography*, *123*, 102252. https://doi.org/10.1016/j.apgeog.2020.102252
Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K. W. (2017). *Men also like shopping: Reducing gender bias amplification using corpus-level constraints*. arXiv:1707.09457.