Check for updates

Earth and Space Science

RESEARCH ARTICLE
10.1029/2023EA003364

Special Collection:
Advances in Machine Learning for Earth Science: Observation, Modeling, and Applications

Increasing the Reproducibility and Replicability of Supervised AI/ML in the Earth Systems Science by Leveraging Social Science Methods

Christopher D. Wirz[1], Carly Sutter[2], Julie L. Demuth[1], Kirsten J. Mayer[1], William E. Chapman[1], Mariana Goodall Cains[1], Jacob Radford[1,3,4], Vanessa Przybylo[2], Aaron Evans[2], Thomas Martin[5], Lauriana C. Gaudet[6], Kara Sulia[2], Ann Bostrom[7], David John Gagne II[1], Nick Bassill[2], Andrea Schumacher[1], and Christopher Thorncroft[2]

[1]NSF National Center for Atmospheric Research, Boulder, CO, USA, [2]University at Albany, SUNY, Albany, NY, USA, [3]NOAA Global Systems Laboratory, Boulder, CO, USA, [4]Colorado State University, Fort Collins, CO, USA, [5]NSF Unidata, Boulder, CO, USA, [6]The Weather Company, Andover, MA, USA, [7]University of Washington, Seattle, WA, USA

Key Points:

- We provide a rigorous hand labeling procedure to improve the replicability and reproducibility of supervised machine learning (ML)
- Our case study and step-by-step guide clearly outline how the procedure can be applied
- The procedure is an actionable path forward for addressing ethical considerations and goals for ML development in Earth systems science

Correspondence to:
C. D. Wirz,
cdwirz@ucar.edu

Citation:
Wirz, C. D., Sutter, C., Demuth, J. L., Mayer, K. J., Chapman, W. E., Cains, M. G., et al. (2024). Increasing the reproducibility and replicability of supervised AI/ML in the Earth systems science by leveraging social science methods. *Earth and Space Science*, *11*, e2023EA003364. https://doi.org/10.1029/2023EA003364

Received 19 OCT 2023
Accepted 17 JUN 2024

Author Contributions:

Conceptualization: Christopher D. Wirz, Carly Sutter, Julie L. Demuth, Kirsten J. Mayer, William E. Chapman, Mariana Goodall Cains, Ann Bostrom
Data curation: Carly Sutter
Formal analysis: Carly Sutter
Funding acquisition: Julie L. Demuth, Kara Sulia, Ann Bostrom, David John Gagne II, Nick Bassill, Christopher Thorncroft
Investigation: Christopher D. Wirz, Carly Sutter, Mariana Goodall Cains,

© 2024. The Author(s).
This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

Abstract Artificial intelligence (AI) and machine learning (ML) pose a challenge for achieving science that is both reproducible and replicable. The challenge is compounded in supervised models that depend on manually labeled training data, as they introduce additional decision-making and processes that require thorough documentation and reporting. We address these limitations by providing an approach to hand labeling training data for supervised ML that integrates quantitative content analysis (QCA)—a method from social science research. The QCA approach provides a rigorous and well-documented hand labeling procedure to improve the replicability and reproducibility of supervised ML applications in Earth systems science (ESS), as well as the ability to evaluate them. Specifically, the approach requires (a) the articulation and documentation of the exact decision-making process used for assigning hand labels in a "codebook" and (b) an empirical evaluation of the reliability" of the hand labelers. In this paper, we outline the contributions of QCA to the field, along with an overview of the general approach. We then provide a case study to further demonstrate how this framework has and can be applied when developing supervised ML models for applications in ESS. With this approach, we provide an actionable path forward for addressing ethical considerations and goals outlined by recent AGU work on ML ethics in ESS.

Plain Language Summary Artificial intelligence and machine learning can make it hard to do science in a way that can be repeated. This can mean redoing a study in the exact same way to see if you can get the same or similar results (reproducibility) or trying to use the same study design on a new problem to see if the results are the same or similar (replicability). These types of scientific repetitions is important for developing robust knowledge, but is hard to do with certain types of machine learning that rely on data that were categorized by researchers. The researchers have to make decisions and categorize their data, which the machine learning algorithm then uses as a guide to make its own decisions. Generally, there is not enough information shared by the researchers about how these decisions were made to repeat the science or evaluate how good it is. In this paper, we provide a way to address these shortcomings. The approach and example we offer illustrates how to (a) create a rulebook that can be shared for how to make decisions and (b) quantitatively measure how consistent the researchers are at using that rulebook to make their decisions.

1. Introduction

The roles and prevalence of artificial intelligence (AI) and machine learning (ML) have rapidly increased in Earth systems science (ESS) research. ESS applications of ML range from detection (e.g., Prabhat et al., 2021) to predictive applications (e.g., Chapman et al., 2022; Ham et al., 2019; Mayer & Barnes, 2022; Weyn et al., 2021). Many important developments in the ML-ESS space in particular have been made through supervised ML (e.g., Beucler et al., 2021; Lam et al., 2022; Pathak et al., 2022). Supervised ML is an ML approach that requires a model to be trained using a set of labeled data (e.g., Hastie et al., 2009). Although some of these labels can be defined through mathematical heuristics, for example, by using climate or weather indices (e.g., the Madden-Julian Oscillation index, (Wheeler & Hendon, 2004) or anomaly/normalization methods (e.g., Ham et al., 2019; Mayer & Barnes, 2021, 2022), some of them, especially image classification tasks, require a human to assign the labels or to "hand label" the data (e.g., Biard & Kunkel, 2019; Prabhat et al., 2021). We use the term

23335084, 2024, 7, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2023EA003364, Wiley Online Library on [02/06/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Jacob Radford, Vanessa Przybylo, Aaron Evans
Methodology: Christopher D. Wirz, Carly Sutter, Julie L. Demuth, Mariana Goodall Cains, Vanessa Przybylo
Supervision: Julie L. Demuth, Kara Sulia, Nick Bassill, Christopher Thorncroft
Visualization: Christopher D. Wirz, Kirsten J. Mayer
Writing – original draft: Christopher D. Wirz, Julie L. Demuth
Writing – review & editing: Christopher D. Wirz, Carly Sutter, Julie L. Demuth, Kirsten J. Mayer, William E. Chapman, Mariana Goodall Cains, Jacob Radford, Vanessa Przybylo, Aaron Evans, Thomas Martin, Lauriana C. Gaudet, Kara Sulia, Ann Bostrom, David John Gagne II, Nick Bassill, Andrea Schumacher, Christopher Thorncroft

"hand labeling" broadly to refer to the process of humans manually labeling the training data for supervised machine learning. We note this generally covers approaches for categorizing samples or identifying features "by hand." Such approaches allow humans to be explicitly involved in defining and identifying the examples for the algorithm to "learn" from and then have the algorithm complete tasks across large data sets that would be impractical or infeasible for the team of humans to label on their own (Nasteski, 2017).

Current practices of hand labeling for supervised ML typically involve a member (or subgroup) of the research team who makes labeling decisions based on their personal judgment. This labeling is often done with little or no documentation about how the phenomenon or feature to be labeled should be defined and thus how labels should be assigned, nor how ambiguous cases should be handled. The process outlined here is likely familiar to many researchers who have used hand labeling and represents a status quo (e.g., Raji et al., 2021; Sambasivan et al., 2021), with several examples in ESS (e.g., Bond et al., 2007; Gagne et al., 2009; Sobash et al., 2023). However, given the reliance on personal judgment, which may stem from human intuition, there are no measures of how consistent multiple labelers were and no clear documentation on the process and logic used to assign labels.

The general approach used by these studies and others is problematic because it "black-boxes" essential methodological detail about what the supervised ML model was trained on and the quality of those data. The black-boxing of the labeling process limits the reproducibility of the training data and thus the extent to which the research can be reproduced, replicated, and evaluated by researchers or potential users. From our experience, not having these details clearly documented with a training procedure makes it difficult for research teams to expand their training sets or handle personnel changes and turnover on the labeling team. Furthermore, this "status quo" approach to hand labeling has the potential to introduce undiagnosable error into the training data, which can propagate into the ML model and its output, as we discuss later on. Such data bias issues are one key element of broader conversations about and approaches to improving ethics and responsibility of ML models (McGovern et al., 2022).

In this paper, we introduce quantitative content analysis (QCA) as an approach to increase the reproducibility and replicability of hand labeling for training supervised ML in ESS applications. With this approach, we provide an actionable path forward for addressing ethical considerations and goals outlined by recent AGU work on ML ethics in ESS (Stall et al., 2023). QCA is a method used in social science to systematically and objectively categorize data using a standardized set of rules, known as a "codebook," together with assessments of reliability (Coe & Scacco, 2017). This method provides a more systematic approach for hand labeling training data and a reporting process for subsequent evaluation and documentation (Franzosi, 2008). In this paper, we identify potential problems with current approaches to hand labeling and describe how QCA can help address them. We then provide a general overview of the QCA approach, followed by an explanation of applying QCA in our work developing a supervised ML model for New York State Department of Transportation (NYSDOT) to assess roadway conditions. We use this case study to further demonstrate how this framework has been and can be applied when developing supervised ML models for ESS applications.

2. Potential Problems With Hand Labeling Approaches for Supervised ML Models

In ML development and research, there are generally acknowledged issues with inadequate reporting and documentation on how key decisions were made in model development, which inhibit other researchers' ability to meaningfully reproduce, replicate, and evaluate others' models (e.g., Gundersen & Kjensmo, 2018; Liu et al., 2021). Although providing code is a helpful practice for addressing these points, code does not communicate everything. Historically, code has not been shared for the majority of papers (e.g., Hutson, 2018), and it does not generally convey anything about how the labeling was done nor if it was done well (Haibe-Kains et al., 2020). Making the labeled data sets publicly available is desirable, but it fails to provide the important details about how the data set was created. Similarly, journal publications are generally the way researchers share their methodologies and details on their decision making, but articles vary in what is reported with many key details often excluded.

In this section, we outline two interconnected areas that are relevant to current approaches for hand labeling training data for supervised ML models and the lack of reporting and communicating the procedures used. The first area of this section relates to problems with *reproducibility* and *replicability*, which have broader scientific implications. *Reproducibility* refers to the ability to generate the same (or similar) results of a study using the same approach,

23335084, 2024, 7, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2023EA003364, Wiley Online Library on [02/06/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

code, and data as the original study, and *replicability* refers to finding similar results as a previous study in a new study employing the same or similar approach but not necessarily the same code or data (NASEM, 2019). Both reproducibility and replicability are important for advancing science in that they are ways to evaluate the results presented by a single study and, in some instances, to determine how generalizable a study's findings are or are not. The second area of this section pertains to the ability of other researchers and potential users to evaluate how supervised ML models were developed, which focuses more on the extent to which it is possible for experts and potential users to meaningfully evaluate ML models and the influence of this evaluability (or lack thereof) on their perceptions and use of models. To be clear, we do not claim these problems are true for all hand labeling campaigns in ESS, but rather we are highlighting problematic areas that some labeling efforts have or may run into.

2.1. Problems With Reproducibility and Replicability of Hand-Labeled Supervised ML Models

Issues with methodological transparency and reporting relate to concerns about the reproducibility and replicability of ML (Gundersen & Kjensmo, 2018; Haibe-Kains et al., 2020; Hutson, 2018; Peng, 2011; Tatman et al., 2018). For some, reproducibility has been a major concern broadly for research using computational and ML approaches (e.g., NASEM, 2019; Peng, 2011; Stodden et al., 2016) and specifically within ESS (e.g., Bush et al., 2021; Nüst & Pebesma, 2021). There is a growing body of literature that is starting to address these concerns across the ML community (Gundersen et al., 2022; Gundersen & Kjensmo, 2018; Pham et al., 2021; Zhuang et al., 2021). However, in the ESS domain there is a strong need for improvement on how the reproducibility and replicability of ML models are addressed in practice. From the computational and ML perspective, many ESS authors focus largely on reporting and sharing the data, code, workflows, and details on the computing environment to increase reproducibility and replicability of such studies. However, for supervised ML models trained using hand labels to be more reproducible and replicable, detailed information about how hand labels were defined and assigned must also be included.

Given the importance of the hand labels to the model and its performance (Northcutt et al., 2021), knowing precisely how the labels were assigned is essential for new researchers to recreate the workflow when attempting to reproduce the work and for efforts to replicate the results (Bond, 2015; Bond et al., 2007). For reproducibility, insufficient documentation on labeling procedures inhibit researchers from being able to generate labeled data from the raw data in the same or even similar way. Different interpretations and perceptions of the labeling task can introduce errors and biases that will propagate into further errors and differences in the resulting model. If the labels are included with the training data, relabeling the data would not be needed to assess the reproducibility of the model architecture. However, if the labeling process is considered part of the research being reproduced, then it should also be evaluated in a reproducibility experiment. This raises issues of data reproducibility, or concerns about being able to effectively reproduce the data that were used for a given study (Boulbes et al., 2018; Mobley et al., 2013; Pawlik et al., 2019; Xu et al., 2022), for the ML research community.

Prior research has addressed similar problems but uses slightly different concepts and approaches, such as the quality of data annotations (Aroyo & Welty, 2015; Welty et al., 2019) or crowd sourced data (Daniel et al., 2018). All this work comes together to demonstrate the importance of data quality (Sambasivan et al., 2021), especially for supervised ML development (Inel & Aroyo, 2019). The reproducibility of the data can affect the ML models in two ways. First, models with the same architecture that use the same training data set that has been labeled differently can lead to differences in the model output, which are affected by the quality of the different labeling (Shamsaliei et al., 2022). Second, because there are many sources of irreproducibility (Gundersen et al., 2022), models with the same architecture trained with different training data sets aimed to represent the same concepts may also yield different results. To this end, reproducibility, especially in the context of data, are essential considerations for ML in ESS research and the main motivation of this paper.

For replicability, if the labeling process is not appropriately documented then efforts to apply the labeling approach to new domains or data sets will be limited in their ability to claim the research does or does not generalize to the new context. For ML applications, replicability has been defined as "obtaining consistent results across studies aimed at answering the same scientific question using new data or other computational methods" (Adali et al., 2022, p. 5). Putting the definition in the context of our case study, replicability would apply to taking the same model architecture and training it on data from a different state or getting consistent results for the same state using a different model architecture. However, what counts as "consistent" and the limits of replicability, as well as reproducibility, for ML is still largely an active and evolving area of research (NASEM, 2019). Relatedly,

23335084, 2024, 7, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2023EA003364, Wiley Online Library on [02/06/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

the lack of reproducibility also then negatively impacts the replicability of that research. For example, if a study applied a vague, poorly documented labeling scheme and model architecture to a new case and found different results than the initial study, it would be difficult to assess whether the issue was differences in how labels were assigned or whether the research overall did not generalize. The ability to reproduce and replicate ML models is a key part of scientific advancement that requires effective methods to facilitate in practice.

2.2. Problems With Being Able to Evaluate the Development of Supervised ML Models

When key documentation and reporting of the hand labeling process outlined above is lacking, other experts and potential users cannot meaningfully evaluate the supervised machine learning. Being able to evaluate the quality and design of a supervised ML model is essential because assigning labels can be a subjective process in which even experts vary greatly in their execution of the task (Bond, 2015; Bond et al., 2007). Such variation in expert labeling has been quantified in applications of supervised ML for identifying atmospheric rivers, which required post hoc methods to account for the variation in the model development (O'Brien et al., 2020).

Furthermore, rigorous and well-documented hand labeling procedures are essential for evaluating and assessing quality because this phase can introduce biases that affect the overall model. Researchers have demonstrated how the often subconscious decisions and biases of developers lead to harmful errors that are then concretized and hidden behind what some perceive to be "objective" algorithms (e.g., Barocas & Selbst, 2016; Bechmann & Bowker, 2019). As McGovern et al. (2022) demonstrated, issues with biased or faulty training labels are ones in which AI can lead to problems in environmental science-related applications. Put simply, the adage "garbage in, garbage out" describes the importance of high quality hand labeling for training supervised ML models (Chase et al., 2022; Filipiak et al., 2023). Generally, inconsistent and poorly defined hand labeling will lead to worse and more inconsistent performance, whereas more consistent hand labeling for training models will yield better and more consistent performance (Northcutt et al., 2021). Although performance is not the only factor that influences users' trust and use of ML models (Wirz, Demuth, et al., 2022), it nonetheless can play an important role, and having rigorous approaches to enhance model performance with thorough and systematic documentation can further be helpful.

2.3. Systemic Pressures That Exacerbate Problems With the Reproducibility and Replicability of Supervised ML Models

The issues of reproducibility and replicability of supervised ML models are exacerbated by pressure to move at an incredibly fast pace to keep up with ML advances, which often outweighs incentives for taking time to focus on quality labeling. Trying to stay at the cutting edge of ML is a herculean task for many academics, especially when competing with other labs and private sector companies with more resources and flexibility. This is reflected to some degree by the recent calls to pause large scale AI experiments (Bengio et al., 2023), which aim to counteract the ways competition has prompted the quick publishing of models over focusing on the safety and ethics of these models. Some research and development teams might not be aware that individual labels are an important aspect of reproducibility and replicability of supervised ML and, if they are aware, may lack the networks, time, and resources needed to pursue high-quality collaborative labeling. These points, paired with the often time-intensive and tedious nature of hand labeling thousands of examples, often lead researchers to focus on getting the task done rather than on being deliberate. Public sector and academic researchers are often left under-resourced and pressured for time as they attempt to compete with the private sector on ML research (Togelius & Yannakakis, 2023). Such pressures and stresses have led to a system that is vulnerable to prioritizing speed over robustness and quality.

An excellent demonstration of the importance of such efforts is a study conducted by Lebovitz et al. (2021) that demonstrated how expert users from the medical domain wanted to know more information about the process used for hand labeling data and about the expertise of the labelers when deciding whether or not to explore the use of ML models in a professional context. These experts needed to be able to critically evaluate the labeling process and the specific labels that were used to train the ML models in order to meaningfully evaluate the model and its performance. The medical experts expected to have this information *before* using the ML model. In the next section we outline how QCA can help address these limitations and provide an approach for more robust development in ESS moving forward. This example demonstrates how, in the context of hand labeling training data, approaches that are consistent and well-documented are important for deploying applications of ML models.

23335084, 2024, 7, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2023EA003364, Wiley Online Library on [02/06/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

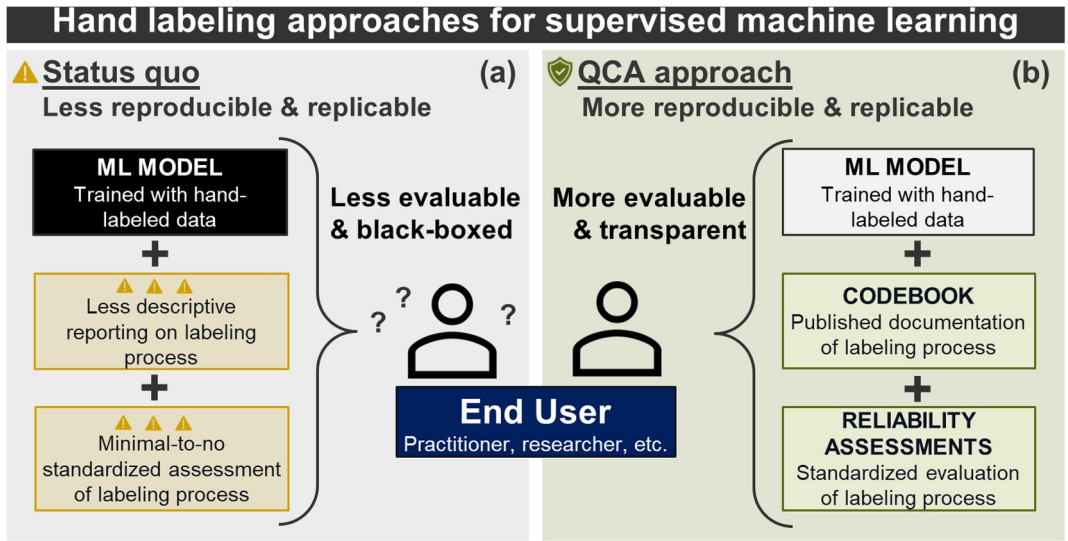


Figure 1. Conceptual diagram of how QCA addresses weaknesses of the hand labeling status quo (a) and can help make supervised ML more reproducible and replicable (b).

3. How QCA can Address Limitations and Problems of Training Supervised ML Models

QCA is a systematic categorization that uses a "codebook" to document and standardize all decisions made in the hand labeling or *coding* process and then uses a statistical evaluation to quantify the *reliability* of the labelers or "coders." Quantitative content analysis terminology, such as coders and coding, are similar to words that carry different meanings in the ML community. Here, we use the term "labelers" instead of coders and "labeling" instead of coding for clarity, but we provide the QCA terminology for those who are interested in exploring that literature (see Neuendorf (2016) and Krippendorff (2018) for general QCA instruction and background).

Content analysis has broadly been discussed and applied as a method in social science research for decades (e.g., Gerbner, 1958; Kaplan, 1943). Over time, methods for QCA have been debated, evaluated, and refined into a largely standard approach that involves the development of a codebook and reliability assessments (e.g., Krippendorff, 2018; Neuendorf, 2016). Interestingly, over 80 years ago, Berelson (1952) noted the applicability of content analysis to the natural sciences, "To the extent that historians, students, of literature, lawyers, economists, anthropologists, and *even natural scientists* (emphasis added) deal with the materials of communication–and all of them do, to some extent–content analysis procedures may be useful" (Berelson, 1952, p. 9).

QCA has the potential to address the limitations described above with supervised ML in the ESS by providing clear, detailed documentation on the decision-making process via a codebook and by quantifying the reliability of the hand labelers. The codebook and reliability metrics can be published with the ML model to provide more transparency in the underlying data set used to train the models (Figure 1). This approach expands efforts to address human bias in labeling data (e.g., Inel & Aroyo, 2019; Welty et al., 2019) by leveraging QCA as a structured way of ensuring reliability among labelers that has a strong grounding in social science. We further contribute to this space by providing a detailed, step-by-step process that is tailored specifically for ESS researchers.

In this section, we provide a high-level overview of the core components of QCA—i.e., of codebooks, sampling data for labeling, and reliability metrics—and summarize several points for ML and ESS scientists to consider before beginning the labeling process. This section is meant to introduce readers generally to the process and main ideas, and then we provide a concrete example from our work in Section 4.

3.1. Codebooks: QCA Provides a Systematic and Transparent Approach for Hand Labeling

The first major component of QCA is a *codebook*, which is a document that clearly articulates the labeling process and standardizes the decisions to be made for assigning each label. The development of the codebook is iterative

23335084, 2024, 7, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2023EA003364, Wiley Online Library on [02/06/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

and generally time-intensive because concretizing *exactly* what to label data and how requires externalizing detailed information and adjudicating opinions on how labeling should be done.

Importantly, developing a codebook requires a great deal of definitional work, meaning time spent defining and identifying exactly what concepts and phenomena are of interest, what the set of possible labels are, and how each label is identified (Benoit, 2014). The difficulty of creating a codebook and the codebook's complexity depends on the task at hand. The phenomena of interest may be either *manifest* or *latent*. *Manifest* content refers to concrete concepts that are clearly present or absent, whereas *latent* content refers to more abstract concepts that cannot be measured directly but instead are identified by several indicators (Riffe et al., 2018). For instance, if a goal is to hand-label the presence or absence of water in typically dry streambeds based on high-quality webcam images, "water" is a relatively manifest phenomenon, and thus the codebook would require minimal information about how to identify it. Conversely, if a goal is to hand-label a tornadic signature using radar data, "tornadic signature" is a latent phenomenon, and thus the codebook would require more explanations, examples, and rules for what indicators (e.g., hook echo, gate-to-gate velocity couplet) hand-labelers should look for to identify it. In both cases, the first step is to develop the codebook to articulate all the information that one should need to label the data as well as whether specialized knowledge may be needed of labelers or if the codebook is general enough that anyone should be able to use it.

Although individual codebooks vary, they should all define the *unit of analysis*, or what exactly is being labeled, as well as *how labels are to be assigned* to each of these "units." If you were training your model on images of the streambed, the unit of analysis would be images of the streambed for your hand labeling. Units will vary greatly from model to model and must be clearly articulated. The codebook should provide a list of possible labels, descriptions of each label, and how each label should be identified, using examples when possible. Each option should be listed and described in sufficient detail, providing examples when possible. Labeling options must be *exhaustive*, which means there must be a label for each sample, even if this means having an "other" label option. The codebook also must clarify if labels are *mutually exclusive*, meaning only one label can be assigned to each sample, or if multiple labels can be assigned to each one. Note, in the rest of the paper we will refer to "units of analysis" as "samples" to be consistent with ML terminology. An additional option when labeling is to use a *labeling certainty metric* that labelers can use to rate their certainty in assigning a specific label that can be further factored into the model training process (e.g., O'Brien et al., 2020).

In sum, the goal of developing the codebook is to explicate the purpose of what is being coded for, why, and how. In practice, this will mean that, ideally, someone newly joining the project could easily assign labels consistent with the original team by using the codebook. The output of this process is a document, the codebook, that can be published with the model or associated paper (see Figure 1b), which increases the reproducibility and replicability of the work, as well as the ability to evaluate it. For an example of a published codebook, see Wirz, Cains et al. (2022). It is important to allow sufficient time to develop the codebook because the first draft rarely, if ever, sufficiently captures the needed detail for easy and consistent hand labeling across a team. Typically when the first draft of the codebook is put to the test, the team finds ambiguities in the wording or that it does not capture some part of the labeling process. The labeling team should then work together to resolve disagreements and confusion by adding clarifications, rules, and more examples to the codebook as needed until the codebook and labeling team are *reliable* (see Section 3.3).

3.2. Sampling: QCA Requires a Deliberate and Documented Process for Samples to Label

Sampling, or the process by which sets of samples are drawn from a population of interest, is a key factor for training supervised ML models, and thus it is also important for the QCA hand labeling process. Just as the approach by which samples are selected when training an ML model affects the model's performance, the sampling for reliability trials will also affect the subsequent model. As we will discuss in the next Section 3.3, you will need sets of samples to test your codebook and assess the reliability of your labeling process and team. These reliability trial samples should not be part of your final data sets for model training or evaluation unless the samples are labeled again by one of the certified labelers after the group has reached inter-coder reliability, so that samples that have been reviewed and discussed for codebook development do not bias the training data (Neuendorf, 2016).

Sampling techniques can vary significantly and may involve random selection or purposeful methods, among other possibilities. The choice of sampling technique differs by project, and thus there is not one

23335084, 2024, 7, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2023EA003364, Wiley Online Library on [02/06/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

standard approach to follow. However, we generally recommend *random sampling* to better capture a more full range of variance and potential types of samples from your data, to represent the true variance of the underlying data the model will ultimately be analyzing. However, from the model training perspective, random sampling is problematic for many geospatial ESS applications because autocorrelation between neighboring data points can artificially inflate skill (e.g., Meyer et al., 2018, 2019; Valavi et al., 2019). To this point, we note that the sampling process for the reliability trials does *not* have to be the same approach used for the model training sampling. For example, one could use a random sampling approach to achieve reliability and then use a different sampling approach to generate the samples for the labeling team to then label for training the model.

Another sampling option for the reliability trials is *purposive sampling*, which may be necessary if your classes are highly unbalanced (your data are not evenly distributed across the different classes but rather have some classes with a much larger proportion of the data than others). This means specifically targeting certain types of samples from the training set to supplement your random sampling. However, it is crucial to be very deliberate about how this sampling is done and to report the details in your model documentation, as it can influence the results of your reliability trial. Note that how any undersampling or oversampling is done may introduce bias in the labeling process, so work to select these targeted samples to have as much variation as possible.

3.3. Reliability: QCA Provides Standardized Measures of How Consistent the Labelers Are at the Labeling Task

In QCA, reliability has been broadly defined as how *stable* individual labelers are (i.e., consistent over time) and how *reproducible and replicable* the labeling scheme is overall (Krippendorff, 2018). In other words, reliability is a way for us to assess how "good" our labeling really is and whether or not others should be able to repeat what we have done. Reliability in this context also refers to how consistently the labelers are able to apply the codebook when labeling. Although this may seem like a relatively straightforward point, the concept of reliability and how it should be assessed have been some of the most widely discussed topics in the QCA literature (e.g., Hayes & Krippendorff, 2007; Krippendorff, 2004; Lacy et al., 2015; Lombard et al., 2004; Lovejoy et al., 2016). These important debates are highly technical and are the foundation for leading practices in social science research. In this section, we synthesize the main points from this literature and adapt them in recommendations and considerations for ESS applications.

3.3.1. Reliability Trials

The main idea for reliability in QCA is that there must be a quantitative evaluation of agreement among the entire labeling team to assess their consistency with each in applying the codebook *before* they officially label any data. The evaluation is done through what are called *reliability trials,* which involve each member of the labeling team independently labeling the same set of practice samples (remember these samples are ideally drawn from data that will not be used for training the model) and then applying statistical methods to assess the reliability among the labelers. Independence is key here because these assessments are meant to represent what labelers will do if and when they are on their own labeling data to be used for training later on. The exact nature of this task will vary depending on what the unit of analysis is and the complexity of the labeling process. Nonetheless, you must set a number of samples the team will all label and generate these samples using the sampling approach you designated in the codebook (see Section 3.2). When applying QCA for hand labeling in ESS, we recommend (a) selecting the number of samples that a labeler could reasonably label in roughly 1 hr and (b) randomly sampling when possible (e.g., with approximately balanced classes where data are roughly evenly distributed across all classes) and sampling purposefully otherwise (e.g., with unbalanced classes).

Our 1 hr rule is a loose standard that generalizes well to a range of labeling tasks, but it also comes from our experiences and considerations for label quality. Assigning too many samples to label independently runs the risk of labelers either feeling rushed to complete the task or experiencing *coding fatigue* partway through the task. Both rushing and fatigue negatively affect the quality of the labeling, which tends to result in the need for more reliability trials and, in turn, more rushing and fatigue. Conversely, having too few samples to label makes it challenging to meaningfully assess reliability of the labeling. Although future work may develop more concrete guidelines and leading practices, we provide this standard for research and development teams to use as a starting point and adapt

23335084, 2024, 7, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2023EA003364, Wiley Online Library on [02/06/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

as needed for their respective projects. However, a key point here, and for QCA in general, is to record and report whatever procedure was used in the documentation and publications associated with the final model.

Once you have a target number of samples for your reliability trial, prepare the samples for everyone in the reliability trial to label. If you were randomly sampling and determined that 50 samples was the right number to label for the reliability training, you would pull a random subset of 50 samples from the full unlabeled data set. Then each labeler would use the codebook to label their own copy of these 50 samples *independently*, that is, on their own, without talking to one another nor seeing each others' labels. After this process, the labeling team should assess their sample for how often each of the labels were used and discuss whether or not the sample gave sufficient practice for all categories. If some labels were not used or only came up a few times, consider a *purposive* sampling approach for the next trial (see Section 3.2 for considerations on purposive sampling).

3.3.2. Reliability Measures and Reporting

After completing a reliability trial, the next step is to empirically evaluate the trial results. There are multiple statistical measures that can be applied to the results of the trial to assess reliability, which has led to debates about which are the best to use. Krippendorff's alpha (KA) is widely considered to be the gold standard reliability metric (Hayes & Krippendorff, 2007; Krippendorff, 2004, 2018; Lacy et al., 2015; Neuendorf, 2016). KA is a strong option because it works with a range of data types (e.g., categorical, ordinal), can incorporate any number of coders, and allows for missing data (e.g., unlabeled samples). See Hayes and Krippendorff (2007) for an overview of common approaches and justification for using KA.

There are easy and free tools available for calculating intercoder reliability measures. We recommend using the materials created by Freelon, which provide helpful documentation on how to prepare and analyze the data, along with an online tool to run the calculations (Freelon, 2010, 2013). Using these tools, you will then get a value for each of the reliability measures for your given reliability metric. Generally, the goal is 0.8 KA or higher (Krippendorff, 2004), but above 0.7 can be justified if the research is exploratory or an especially tricky, latent concept (Lacy et al., 2015). Some argue that your exact threshold should be related to the stakes of the hand labeling and that there is *some* flexibility in these values (Krippendorff, 2004) and not a set standard for cutoffs (Neuendorf, 2016). Importantly, note that in order to calculate these measures, your labeling team and reliability trials must include at least two labelers. Even if the second labeler does not do any labeling after the final reliability trial, the second person is needed to meaningfully assess reliability.

If your results are above the thresholds outlined here or that you set and justified for your project, then you are ready to begin labeling your training data. It is important to emphasize that, after the labeling team has reached acceptable reliability, each person is "certified" to label samples on their own. This is because *the labelers have demonstrated consistent application of the codebook—that is, are reliable and consistent human "instruments"—and thus do not need to label in teams or label the same samples as a cross check.* It is important to document and report the number of trials and information about the trials, your reliability thresholds, and your final reliability scores. Providing this information in publications and documentation is essential, as it allows others to evaluate your labeling process. If, however, your reliability results are not above the thresholds, you will need to revise, review, and retest through a process we discuss in more detail in Section 4.1.

Lastly, we recommend making a plan and policy if the labeling will take place over longer than a few months or will require adding in new team members. For longer projects, we recommend checking for *drift* (i.e., labelers who were once reliable are no longer so, often after a long break in labeling) among the labelers by conducting another reliability trial with the full team every few months to make sure the team is still labeling consistently. We did this for the NYSDOT case study project we describe below to ensure we were still calibrated after a few month pause on labeling. When adding new labelers, we recommend the new member work with one current member to be trained on the codebook and labeling process before conducting a reliability trial with both the new and current member. Ideally, these two labelers can reach reliability without needing to change the codebook. If they are able to do this, the new labeler is sufficiently reliable and is ready to label images. If not, then the labeling team must come back together and assess the situation. If the inability to reach reliability is because of points not reflected in the codebook, this should be addressed in an updated version of the codebook. If the issue is a lack of essential background knowledge or experience, this should also be clarified in an updated version of the codebook. In the next section we take these high-level ideas and concretize them with an example from our work.

23335084, 2024, 7, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2023EA003364, Wiley Online Library on [02/06/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

4. How to Apply QCA to Hand Labeling for the Training of Supervised Machine Learning Models: A Case Study

Thus far, we have outlined the need for QCA for hand labeling supervised ML training data, and we have provided a broad overview of the core concepts underlying the method. In this section, we build on this foundation and describe the process, or the *how*, more concretely. We provide a step-by-step process and suggest leading practices using an example from our team's work (led by co-author Sutter). Our goal is for this section to provide the basis for others to design and implement QCA for their own hand labeling projects. We begin with an overview of the case study project's goals and background to situate the reader.

The NYSDOT has a network of about 2400 live traffic cameras throughout the state that are used by NYSDOT specialists as a method of determining road surface conditions during the winter. However, monitoring cameras is a time-consuming process, and, given the large number of camera images, the NYSDOT is interested in using ML methods to automate this process. The camera network and NYSDOT's current winter road condition classification are publicly available at 511ny.org (NYSDOT, 2023).

The goal of this project was to use the camera images to predict a specific set of road conditions, and thus a fully-supervised machine learning approach was developed. This required a large hand-labeled data set of camera images for each road condition class. Specifically, a convolutional neural network was trained to predict road surface conditions from a set of labeled road camera images. The xCITE lab at the University of Albany's Atmospheric Sciences Research Center has an archive of the camera images starting in January 2022 (xCITE, 2023). From the xCITE archive of images, we pulled samples from good quality, well-lit cameras for hand-labeling. The New York State Mesonet (NYSM), a network of weather stations (Brotzge et al., 2020), was also used to help sample images by pulling examples during varying weather conditions and to provide additional meteorological information for the hand labelers during the image labeling process.

4.1. Steps for Implementing QCA for Hand Labeling to Train Supervised ML

In this section, we provide six steps for implementing QCA into a supervised ML hand labeling project. Each step is listed in the headings of Table 1 coupled with guiding questions for labeling teams to consider for each step. We detail each step below in the context of our NYSDOT image labeling case study to make the process clearer. The guiding questions and associated descriptions are guidelines and templates to build from and thus will likely need to be modified and adjusted to fit each team's project. Regardless of whether any adjustments are made to our suggested guidelines, it is important to document the process and ensure it is clearly communicated in publications associated with the resulting model. We also note there is a great deal of work that must take place before these steps begin, such as problem identification, defining the scope of the project, assessing the feasibility of the approach, etc. We do not cover this phase here, but rather provide a guide for those who have already determined hand labeling is needed to train their supervised ML model. Our aim with this case study is that it provides a concrete application but has accompanying recommendations and commentary to make the steps more generalizable and easier to transfer to other projects.

4.1.1. Step 1: Identify Goals and Requirements for Training the Model

The ML model for the case study was designed to use images labeled as one of a given set of road surface conditions (RSCs). The images come specifically from the NYSDOT 511NY cameras, which are positioned along roadways throughout the state. Our unit of analysis for labeling was a single image. We anticipated needing several thousand images, so we brought together a team of six labelers to distribute the labeling work.

4.1.2. Step 2: Develop the Initial Codebook With Label Descriptions and Rules

For this effort, the team must develop the codebook by determining and articulating all possible labels and how they should be applied. This project was designed to help NYSDOT officials by providing an automated way to assess road surface conditions in their operational setting. To this aim, the leads of this project met with the NYSDOT to establish how the model could be used operationally, which translated to what labeling classes we needed. Our partners at the NYSDOT identified four key categorizations of RSCs that were important to them for roadway management: *dry, wet, snow,* and *severe snow*. These categories are currently used operationally (with a modification in the ML model to exclude icy conditions which cannot be identified in images) and served as the

Table 1
Overview of the Steps for Implementing QCA for Hand Labeling and Guiding Questions to Consider Throughout the Approach

Step 1	**Identify goals and requirements for training the model**
Guiding questions 	• What is the model intended to do? • What are the inputs needed to train the model? • What is your unit of analysis for labeling? • How many of these inputs or samples do you anticipate needing? • How many labelers do you need to meet your goals?
Step 2	**Develop the initial codebook with label descriptions and rules**
Guiding questions 	• What is the full suite of potential labels needed to train the model? • Are the labels mutually exclusive or can multiple labels be assigned per unit? • What are good examples of each label to use as a reference when labeling? • How should unclear or challenging samples be labeled? • What are the rules for assigning each label and how do these rules interact with one another?
Step 3	**Prepare an approach for reliability assessment**
Guiding questions 	• How many samples should be included in each reliability trial? • How should these samples be sampled? • How will labelers log and share their labels for the trials? • What metrics for reliability will be used and with what threshold? • How will reliability metrics be calculated?
Step 4	**Launch a reliability trial to evaluate the codebook and labelers**
Guiding questions 	• Do the labelers need a training session before the independent labeling? • Do all labelers understand that the trials are independent and know what is expected of them? • Do the labelers have sufficient time to complete the trial without feeling rushed or becoming fatigued? • Did the labelers reach the established standards for reliability? If, **No**, move to *Step 5*; If, **Yes**, move to *Step 6*.
Step 5	**Review disagreements and refine the codebook as needed**
Guiding questions 	• What parts of the codebook were clear and which were unclear? • Are there any themes or patterns in the disagreement? • What could be added or removed from the codebook to resolve ambiguities or disagreements? • Would any samples from the trial be helpful examples for the codebook? • Are there any new rules needed to help clarify the labeling? **Repeat** *Step 4.*
Step 6	**Finalize the codebook and reliability documentation**
Guiding questions 	• Is all of the associated information for the reliability trials documented? • Is all the information used by labelers for the final reliability trial clearly represented in the codebook? • Are there plans for publishing the codebooks and reliability information along with the model in its documentation and/or associated publications? • If this is a longer-term project, is there a plan for how new labelers will be added to the group or for assessing how consistent labelers are over time?

23335084, 2024, 7, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2023EA003364, Wiley Online Library on [02/06/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

23335084, 2024, 7, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2023EA003364, Wiley Online Library on [02/06/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

foundation for our labeling scheme (NYSDOT, 2023). For our set of labeling classes, we augmented these categories with two more that we want the model to predict: *poor visibility* and *obstructed*. We added these two categories to help identify when conditions blocked the ability to see the road, which may end up being a helpful tool for NYSDOT.

Poor visibility covered cases like fog and dense precipitation blocking the camera's view, whereas *obstructed* covered issues with the camera like the image being over-exposed or when the camera was not facing the roadway. The final category for labeling we included was *uncertain*, not to be included in model training. This label was reserved for cases the codebook did not cover or where the labeling task was extremely difficult. We used this label to identify weak points in the codebook and instances where discussion was needed. In our trials, this category was very rarely used. We determined and developed our codebook so that the labels were mutually exclusive and only one label could be assigned to each image.

With the suite of labeling options identified, we then developed a preliminary codebook that defined each of these classes, provided examples, and instructions for how and when to apply each label. We provide a brief overview of how we defined each label with an associated example in Table 2. Our final codebook and the images we coded for the reliability trial are available at Sutter et al. (2023), but note this is the *final* version that took five rounds of reliability trials and discussions to solidify. Our initial version was much simpler and less comprehensive. As we shall discuss in the next sections, we added and clarified many rules to achieve reliability and strengthen the codebook. For example, we added the use of NYSM data (e.g., time since last precipitation, current snow depth, temperature, etc.), to aid in our labeling decision-making, and these details are reflected in the final codebook. The final version reflects the updated rules and guidelines that we used to achieve reliability.

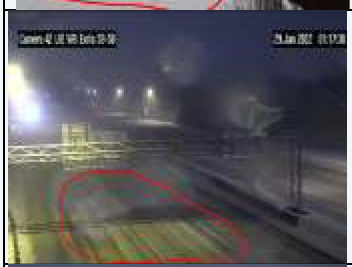
4.1.3. Step 3: Prepare an Approach for Reliability Assessment

For our reliability trials we used approximately 50 images per trial, which satisfied the guideline we described in Section 3.3.1 of about one focused hour of labeling. We found this number to be sufficient for assessing reliability while also not forcing labelers to rush or become fatigued. We used both random and purposive sampling for these trials, but largely used purposive sampling because random sampling resulted in samples that were primarily *dry*. This is intuitive given the majority of the time roads are dry, but not ideal for assessing reliability on the other classes. To this aim, the lead developer of the model (Sutter, who was also a labeler) strategically sampled from time periods within the data set that were likely to have different forms and intensities of precipitation. The other team members were unaware of how this sampling was done and how many images were selected from each purposive sampling effort to avoid priming and anchoring biases. This process ensured more examples from each label type, as well as the transitions between them (i.e., from *wet* to *dry* or from *wet* to *snow*). Ideally, the researcher doing the purposive sampling should not also be a labeler because the sampling process may bias their labeling. However, if the sampling information is kept hidden from the rest of the team, biases should be realized during the reliability trials.

Our approach to labeling and logging our respective labels was very simple. We compiled a spreadsheet that contained a column with links to each of the sampled images and one column for the labeler to list their label. To make calculating reliability easier, we numbered the labels (from 1 to 7) and instructed the labeling team to enter in the number associated with the label they assigned. The spreadsheet also contained optional columns for notes/comments for the labeler to flag any examples they wanted to discuss with the full group after the trial (see step 5). We then duplicated the spreadsheet and made a version for each coder.

We established a KA of 0.8 or higher as our threshold for reliability. We followed this standard from social science naively thinking the task of labeling road surfaces would be straightforward, however, had we realized the complexity of this process we likely would have set that threshold at 0.7. While KA was our determinate reliability statistic, we also calculated percent agreement, Fleiss Kappa, and Cohen's Kappa because they are automatically calculated in the online tool created by Freelon, 2010, 2013. Using this tool was incredibly simple. We copied and pasted each labeler's final labels into one csv file and uploaded to Freelon's site to get the calculated values.

Table 2
Overview of the Road Survey Condition Label Classes, Definitions, and Examples for the NYSDOT Case Study

Example	Label and definition
	**Severe Snow** Pavement is not visible (or minimally visible) in wheel paths, or there are no wheel paths within the snow/slush build-up. Surface cues: Road surface is mostly fully covered with snow (rule of thumb min ~75% covered).
	**Snow** Pavement is visible in wheel paths within the snow/slush. This includes areas with wind blown snow. Surface cues: Snow and/or slush is visible on roads, but the road is not severely covered. There is often a presence of stark pavement streaks/tire lines. In very light slush cases, slush may appear more like a textured wet road.
	**Wet** The segment is wet with liquid water from rain or fully melted snow. Surface cues: Road is visibly wet; common indicators include glare on wet pavement, visible puddles, or car headlights during the day. Also includes roads that are less saturated but still appear to be wet. Surface is darker than usual or has significant dark splotches.
	**Dry** The entire length of the segment appears clear and dry. There are no stretches or spot locations that indicate snow, ice, slush or wet pavement. Surface cues: Roads that are visibly dry and do not include indicators from the other RSC classes. This includes cases where there may be some dark splotches on the road that indicate some wetness, but the majority of the road is dry.
	**Poor visibility** Either the road is not visible at all or the road may be slightly visible but the RSC is not apparent due to weather-related visibility: dense fog, visible precipitation streaks are dense and obstructing the view of road, or blizzard-like conditions with low visibility.
	**Obstructed** Image is too dark, over-exposed, unfocused, fully obstructed, too far from the road, not facing the road, etc.

Note. The red outlines indicate the area of focus for labeling decisions.

23335084, 2024, 7, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2023EA003364, Wiley Online Library on [02/06/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

23335084, 2024, 7, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2023EA003364, Wiley Online Library on [02/06/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

4.1.4. Step 4: Launch a Reliability Trial and Evaluate Reliability With the Drafted Codebook

Before beginning each trial, we reviewed the codebook and any changes we had made as a group, which was followed by time for questions and clarifications. We reviewed a few images together (using examples from past trials if possible) and discussed how we would label them as a group to make sure we were all up to speed and understood the codebook. The process of reviewing and testing the codebook as a group is what we refer to as a *spin-up session*. The spin-up session can be a great way to assess potential issues with the codebook and make sure all labelers are well-calibrated before a reliability trial. We then reviewed the timeline, reminded labelers this was to be done individually (without consulting one another or anything beyond the codebook), and agreed upon a timeline that worked for the whole labeling team. After all labelers completed the reliability trial, we calculated reliability measures using the process detailed in *Step* 3. If the results met our threshold, we moved to *Step* 6. If they were not, as was the case for the first 4 trials, we moved to *Step* 5 to work on strengthening the codebook.

4.1.5. Step 5: Review Disagreements and Refine the Codebook as Needed

If a trial does not meet the reliability threshold, we recommend one person review the aggregated results to identify problematic images and any patterns in the disagreements. Then the full labeling team should meet to first offer their impressions and challenges from the labeling process, then they should review the difficult cases identified ahead of time. We repeated this step several times as we faced new examples and challenging images to label. Through this process, the labeling team worked to clarify wording, add/remove examples, and add/refine rules in the codebook as needed. The goal here is to address *patterns* rather than update the codebook for *every* disagreement. For example, an important point to address would be two labelers consistently applying the same rule differently throughout the labeling process. Addressing an issue that affects many labels is the goal, rather than making new rules that only apply to one rare image type. Finding this balance can be challenging and will likely take some practice to find. The overall goal is to make sure the codebook is general and covers the whole suite of potential samples to label well, rather than hyper-specializing it to capture each image in the entire data set with a unique rule. The latter approach can be tempting, but often results in a codebook that is too long and dense to be applied easily. Essentially, the goal is to avoid "overfitting" your codebook, just like with an ML model. After this process, repeat *Step* 4. For our team, it took four rounds of labeling, discussing agreements and disagreements, updating the codebook and then retesting before we were able to move on to *Step* 6 after completing *Step* 4 successfully.

4.1.6. Step 6: Finalize the Codebook and Reliability Documentation

When you have reached your reliability threshold there are a few last details to consider. First, reflect with the labeling team and make sure that all information needed to effectively label is reflected in the codebook, ideally so that someone unfamiliar with the project could use it easily to label. Walk through the document and make sure everything is clear, well-formatted, and ready to be shared. You should also document the reliability trial process, detailing the specific information covered in *Steps* 1–4, such as the number of reliability trials, number of samples per trial, and the sampling approach taken for the trials. You also want to be sure to log the final reliability measures. For the NYSDOT project, the group achieved a KA value of 0.888. All these details should be represented in any documentation or publication associated with the model, which we will do when the project is complete. A key part of the process is to make a plan for how and where you will publish these details and the codebook. An easy way to publish the codebook is to post it in a repository that provides a DOI, such as Zenodo (e.g., Wirz, Cains, et al., 2022), or include it as an appendix with your publications. At this point, all those who achieved reliability are ready to label data that can be used in training the supervised ML model. Since reliability has been assessed, *they do not need to label in teams and can label data independently.*

The case study outlined in this section demonstrates how the concepts and QCA approach can be applied in practice. However, we note this example is relatively simple compared to other potential labeling tasks in the ESS domain, such as labeling for image segmentation. Further work and applications will be needed to continue to refine this approach for a range of contexts. Nonetheless, preliminary work done by the same research team using the same approach demonstrated the human labeling team achieved higher performance using the QCA approach than a model trained using a rain gauge when classifying precipitation in images (Przybylo, 2023). Our codebook also has already demonstrated success in the realm of replicability. We have used the approach to replicate the labeling process on completely new cameras in different regions of NY state that have varying camera quality,

23335084, 2024, 7, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2023EA003364, Wiley Online Library on [02/06/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

lighting, positioning, and backgrounds. Including these additional cameras in the training data set is helpful for building a model that can eventually generalize well to thousands of other cameras that realistically have differing levels of quality. Additionally, the codebook and detailed QCA process proved successful within our labeling group through our ability to achieve ICR again, months apart, after experiencing labeling drift. This shows preliminary evidence for how the QCA builds reproducibility as well.

5. Conclusion

The QCA-based approach we have outlined reduces the subjectivity of hand labeling by creating a codebook that outlines the exact decision-making rules for assigning labels and then empirically evaluates the reliability of labelers in adhering to those rules. This process and the information it provides also increase the reproducibility andreplicability of the ML model. Moreover, the certification of labelers via the reliability trials is more resource efficient because it allows labeling teams to divide and conquer, rather than everyone having to label all of the same samples to check for consistency. The QCA approach allows the research team to identify and address inconsistencies among labelers a priori through reliability assessments. Although QCA does work to eliminate subjectivity from the labeling process, we also note that the end goal is not necessarily perfect agreement among the labelers but rather a baseline for reliability among the labeling team that can be quantified and communicated. This process may take more time in the initial phases of the project, and obtaining sufficient reliability might prove to be challenging. Nonetheless, we argue the additional rigor, transparency, and reproducibility are well worth the effort. To this end, we have provided an actionable path forward for addressing ethical considerations and goals outlined by recent AGU work on ML ethics in ESS (Stall et al., 2023).

5.1. A Foundation for Future Research on Potential Users' Perceptions of ML

The use of QCA also poses potentially fruitful areas for research about users' perceptions of how ML models are developed. More reproducible, replicable, and evaluable hand labeling of training data for supervised ML models may have implications for how potential users and affected communities perceive the overall ML system. First, having clear documentation on the labeling process and statistical measures of how reliable the hand labeling teams were improves the transparency of the overall model. This transparency allows potential users, as well as other researchers, to more meaningfully evaluate the foundation of what the model was trained to do by being able to interrogate the specific decision-making processes that went into the hand labeling (e.g., Haibe-Kains et al., 2020; Lebovitz et al., 2021). Many have argued that transparency and documentation are key for developing "trustworthy AI" (e.g., Arnold et al., 2018; Ashoori & Weisz, 2019). Although ML trustworthiness is a complex and context-dependent concept (Wirz, Demuth, et al., 2022), methodological transparency may be important for how users and other researchers assess ML trustworthiness.

Further, hand-labeled also represent an interesting intersection of human and automation biases, which have implications for how potential users evaluate the subsequent ML systems. Research has shown some individuals tend to view humans as being more biased than automated systems (e.g., Dzindolet et al., 2003; Muir, 1987), while others are skeptical of automation and prefer human involvement or oversight (Ashoori & Weisz, 2019; Kern et al., 2022). Hand-labeled supervised ML models are a potential way to appeal to both groups, if the labeling is done well. Having humans involved in the labeling of training materials may assuage concerns of ML running away "unchecked" in a domain or system that it may not understand. For example, when National Weather Service forecasters explored a non-operational severe convective weather ML product, several forecasters noted that having humans involved in hand-labeling the algorithm training data set increased the trustworthiness of the product for them, "assuming the humans know what they're doing" (Cains et al., 2023). When addressing such concerns related to the human labelers, having clear documentation and statistical measures of reliability for the hand labeling process may assuage the concerns about human subjectivity or ignorance influencing an otherwise more objective approach. However more empirical research is needed to examine these dynamics and potential implications.

5.2. Additional Contributions of QCA to Supervised ML Workflows

From our experiences applying QCA in ESS contexts, we have also noted a few additional benefits of this approach. First, the robust documentation included in the codebook made it much easier to transfer leadership of a project and to spin up new members of the team. Having a resource that clearly articulated the labeling goals,

23335084, 2024, 7, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2023EA003364, Wiley Online Library on [02/06/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

definitions, and procedures that everyone was working from, rather than an abstract set of undocumented shared ideas and experiences, was essential for smooth onboarding. This is especially relevant in many research settings where students and postdocs may move on to new opportunities before work is completed. Second, the articulation process required in making the codebook facilitated co-development with the target end users for the NYSDOT project. We reached out to members of the NYSDOT to ask questions and facilitate discussions to ensure our definitions and labeling scheme met their expectations, practices, and operational needs. Through this process, we better tailored the model to the NYSDOT's needs because we had to commit time and thought into exactly how we were defining each of the labeling classes. In turn, this use-inspired process then helped us put more thought into the design and goals of the model before it was designed and trained. Researchers in this competitive field may find it tempting to respond to the pressure and pace of ML research by emphasizing speed and efficiency over transparency and documentation. However, the deeply interdisciplinary method we provide represents an opportunity for ESS to pave the way by establishing the expectation for reproducibility and replicability of ML models in the context of hand-labeled supervised ML models.

Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

Data Availability Statement

The codebook and intercoder reliability data used for this paper are available here https://zenodo.org/record/8370665 (Sutter et al., 2023). The NYSM website is available here: nysmesonet.org.

Acknowledgments
We would like to acknowledge the support and insights from all of our colleagues from the NSF AI Institute for Research on Trustworthy AI in Weather, Climate, and Coastal Oceanography (AI2ES). We would also like to acknowledge those who have supported our work and provided camera data and valuable feedback at the NYSDOT. This material is based upon work supported by the National Science Foundation under Grant ICER-2019758. This material is based upon work supported by the NSF National Center for Atmospheric Research, which is a major facility sponsored by the National Science Foundation under Cooperative Agreement 1852977. KMJ is supported by the U.S. Department of Energy, Office of Science, Office of Biological & Environmental Research (BER), Regional and Global Model Analysis (RGMA) component of the Earth and Environmental System Modeling Program under Award Number DE-SC0022070 and NSF IA 1947282.

References

Adali, T., Guido, R. C., Ho, T. K., Müller, K.-R., & Strother, S. (2022). Interpretability, reproducibility, and replicability [from the guest editors]. *IEEE Signal Processing Magazine*, *39*(4), 5–7. https://doi.org/10.1109/MSP.2022.3170665

Arnold, M., Bellamy, R. K. E., Hind, M., Houde, S., Mehta, S., Mojsilovic, A., et al. (2018). FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *In* arXiv *[*cs.CY*]. arXiv*. Retrieved from http://arxiv.org/abs/1808.07261

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

Ashoori, M., & Weisz, J. D. (2019). In AI we trust? Factors that influence trustworthiness of AI-infused decision-making processes. In *arXiv* [*cs.CY*]. arXiv. Retrieved from http://arxiv.org/abs/1912.02675

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, *104*(3), 671–732. http://www.jstor.org/stable/24758720

Bechmann, A., & Bowker, G. C. (2019). Unsupervised by any other name: Hidden layers of knowledge production in artificial intelligence on social media. *Big Data & Society*, *6*(1), 2053951718819569. https://doi.org/10.1177/2053951718819569

Bengio, Y., Russell, S., Musk, E., Wozniak, S., Harari, Y. N., Mostaque, E., et al. (2023). Pause giant AI experiments: An open letter. Retrieved from https://futureoflife.org/open-letter/pause-giant-ai-experiments/

Benoit, W. L. (2014). Content analysis in political communication. In *Sourcebook for political communication research* (pp. 290–302). Routledge. Retrieved from https://www.taylorfrancis.com/chapters/edit/10.4324/9781315782713-26/content-analysis-political-communication-william-benoit-ohio-university

Berelson, B. (1952). Content analysis in communication research. 220. Retrieved from https://psycnet.apa.org/fulltext/1953-07730-000.pdf

Beucler, T., Pritchard, M., Rasp, S., Ott, J., Baldi, P., & Gentine, P. (2021). Enforcing analytic constraints in neural networks emulating physical systems. *Physical Review Letters*, *126*(9), 098302. https://doi.org/10.1103/PhysRevLett.126.098302

Biard, J. C., & Kunkel, K. E. (2019). Automated detection of weather fronts using a deep learning neural network. *Advances in Statistical Climatology Meteorology and Oceanography*, *5*(2), 147–160. https://doi.org/10.5194/ascmo-5-147-2019

Bond, C. E. (2015). Uncertainty in structural interpretation: Lessons to be learnt. *Journal of Structural Geology*, *74*, 185–200. https://doi.org/10.1016/j.jsg.2015.03.003

Bond, C. E., Gibbs, A. D., Shipton, Z. K., & Jones, S. (2007). What do you think this is? "Conceptual uncertainty" in geoscience interpretation. *Geological Society of America Today: A Publication of the Geological Society of America*, *17*(11), 4. https://doi.org/10.1130/gsat01711a.1

Boulbes, D. R., Costello, T., Baggerly, K., Fan, F., Wang, R., Bhattacharya, R., et al. (2018). A survey on data reproducibility and the effect of publication process on the ethical reporting of laboratory research. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, *24*(14), 3447–3455. https://doi.org/10.1158/1078-0432.CCR-18-0227

Brotzge, J. A., Wang, J., Thorncroft, C. D., Joseph, E., Bain, N., Bassill, N., et al. (2020). A technical overview of the New York state Mesonet standard network. *Journal of Atmospheric and Oceanic Technology*, *37*(10), 1827–1845. https://doi.org/10.1175/JTECH-D-19-0220.1

Bush, R., Dutton, A., Evans, M., Loft, R., & Schmidt, G. A. (2021). Perspectives on data reproducibility and replicability in paleoclimate and climate science. *Harvard Data Science Review*, *2*(4). Retrieved from https://par.nsf.gov/servlets/purl/10232859

Cains, M. G., Wirz, C. D., Demuth, J. L., Bostrom, A., McGovern, A., Ebert-Uphoff, I., et al. (2023). Exploring what AI/ML guidance features NWS forecasters deem trustworthy. *103rd AMS Annual Meeting*. Retrieved from https://ams.confex.com/ams/103ANNUAL/meetingapp.cgi/Paper/419371

Chapman, W. E., Monache, L. D., Alessandrini, S., Subramanian, A. C., Martin Ralph, F., Xie, S.-P., et al. (2022). Probabilistic predictions from deterministic atmospheric river forecasts with deep learning. *Monthly Weather Review*, *150*(1), 215–234. https://doi.org/10.1175/MWR-D-21-0106.1

Chase, R. J., Harrison, D. R., Burke, A., Lackmann, G. M., & McGovern, A. (2022). A machine learning tutorial for operational meteorology, Part I: Traditional machine learning. *In arXiv [physics.ao-ph]. arXiv*. Retrieved from http://arxiv.org/abs/2204.07492

Coe, K., & Scacco, J. M. (2017). Content analysis, quantitative. In *The international encyclopedia of communication research methods* (pp. 1–11). Wiley. https://doi.org/10.1002/9781118901731.iecrm0045

Daniel, F., Kucherbaev, P., Cappiello, C., Benatallah, B., & Allahbakhsh, M. (2018). Quality control in crowdsourcing: A survey of quality attributes, assessment techniques and assurance actions. *In arXiv [cs.HC]. arXiv*. Retrieved from http://arxiv.org/abs/1801.02546

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies*, *58*(6), 697–718. https://doi.org/10.1016/S1071-5819(03)00038-7

Filipiak, B. C., Bassill, N. P., Corbosiero, K. L., Lang, A. L., & Lazear, R. A. (2023). Probabilistic forecasting methods of winter mixed precipitation events in New York state utilizing a random forest. *Artificial Intelligence for the Earth Systems*, *1*(aop), 1–37. https://doi.org/10.1175/AIES-D-22-0080.1

Franzosi, R. (2008). Content analysis: Objective, systematic, and quantitative description of content. *Content Analysis*, *1*(1), 21–49. Retrieved from https://sociologie.cuso.ch/fileadmin/sociologie/Content-Analysis--Introduction.pdf

Freelon, D. (2013). ReCal OIR: Ordinal, interval, and ratio intercoder reliability as a web service. ijis.net. Retrieved from https://www.ijis.net/ijis8_1/ijis8_1_freelon.pdf

Freelon, D. G. (2010). ReCal: Intercoder reliability calculation as a web service. dfreelon.org. Retrieved from https://dfreelon.org/publications/2010_ReCal_Intercoder_reliability_calculation_as_a_web_service.pdf

Gagne, D. J., McGovern, A., & Brotzge, J. (2009). Classification of convective areas using decision trees. *Journal of Atmospheric and Oceanic Technology*, *26*(7), 1341–1353. https://doi.org/10.1175/2008JTECHA1205.1

Gerbner, G. (1958). On content analysis and critical research in mass communication. *Audio Visual Communication Review*, *6*(2), 85–108. Retrieved from http://www.jstor.org/stable/30216838

Gundersen, O. E., Coakley, K., Kirkpatrick, C., & Gil, Y. (2022). Sources of irreproducibility in machine learning: A review. In arXiv [cs.LG]. arXiv. Retrieved from http://arxiv.org/abs/2204.07610

Gundersen, O. E., & Kjensmo, S. (2018). State of the art: Reproducibility in artificial intelligence. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1). https://doi.org/10.1609/aaai.v32i1.11503

Haibe-Kains, B., Adam, G. A., Hosny, A., & Khodakarami, F., & Massive Analysis Quality Control (MAQC) Society Board of Directors. (2020). Transparency and reproducibility in artificial intelligence [Review of Transparency and reproducibility in artificial intelligence]. *Nature*, *586*(7829), E14–E16. https://doi.org/10.1038/s41586-020-2766-y

Ham, Y.-G., Kim, J.-H., & Luo, J.-J. (2019). Deep learning for multi-year ENSO forecasts. *Nature*, *573*(7775), 568–572. https://doi.org/10.1038/s41586-019-1559-7

Hastie, T., Tibshirani, R., & Friedman, J. (2009). Overview of supervised learning. In T. Hastie, R. Tibshirani, & J. Friedman (Eds.), *The elements of statistical learning: Data mining, inference, and prediction* (pp. 9–41). Springer. https://doi.org/10.1007/978-0-387-84858-7_2

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.

Hutson, M. (2018). Artificial intelligence faces reproducibility crisis. *Science*, *359*(6377), 725–726. https://doi.org/10.1126/science.359.6377.725

Inel, O., & Aroyo, L. (2019). Validation methodology for expert-annotated datasets: Event annotation case study. *Schloss Dagstuhl - Leibniz-Zentrum für Informatik*. https://doi.org/10.4230/OASICS.LDK.2019.12

Kaplan, A. (1943). Content analysis and the theory of signs. *Philosophy of Science*, *10*(4), 230–247. Retrieved from http://www.jstor.org/stable/184268

Kern, C., Gerdon, F., Bach, R. L., Keusch, F., & Kreuter, F. (2022). Humans versus machines: Who is perceived to decide fairer? Experimental evidence on attitudes toward automated decision-making. *Patterns (New York, N.Y.)*, *3*(10), 100591. https://doi.org/10.1016/j.patter.2022.100591

Krippendorff, K. (2004). Reliability in content analysis. *Human Communication Research*, *30*(3), 411–433. https://doi.org/10.1111/j.1468-2958.2004.tb00738.x

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. SAGE Publications. Retrieved from https://play.google.com/store/books/details?id=nE1aDwAAQBAJ

Lacy, S., Watson, B. R., Riffe, D., & Lovejoy, J. (2015). Issues and best practices in content analysis. *Journalism & Mass Communication Quarterly*, *92*(4), 791–811. https://doi.org/10.1177/1077699015607338

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Pritzel, A., et al. (2022). GraphCast: Learning skillful medium-range global weather forecasting. *In arXiv [cs.LG]. arXiv*. Retrieved from http://arxiv.org/abs/2212.12794

Lebovitz, S., Levina, N., & Lifshitz-Assaf, H. (2021). Is AI ground truth really 'true'? The dangers of training and evaluating AI tools based on experts' know-what. In AI tools based on experts'. Retrieved from https://papers.ssrn.com/abstract=3839601

Liu, C., Gao, C., Xia, X., Lo, D., Grundy, J., & Yang, X. (2021). On the reproducibility and replicability of deep learning in software engineering. *ACM Transactions on Software Engineering and Methodology*, *31*(1), 1–46. https://doi.org/10.1145/3477535

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2004). A call for standardization in content analysis reliability. *Human Communication Research*, *30*(3), 434–437. https://doi.org/10.1111/j.1468-2958.2004.tb00739.x

Lovejoy, J., Watson, B. R., Lacy, S., & Riffe, D. (2016). Three decades of reliability in communication content analyses. *Journalism & Mass Communication Quarterly*, *93*(4), 1135–1159. https://doi.org/10.1177/1077699016644558

Mayer, K. J., & Barnes, E. A. (2021). Subseasonal forecasts of opportunity identified by an explainable neural network. *Geophysical Research Letters*, *48*(10). https://doi.org/10.1029/2020gl092092

Mayer, K. J., & Barnes, E. A. (2022). Quantifying the effect of climate change on midlatitude subseasonal prediction skill provided by the tropics. *Geophysical Research Letters*, *49*(14). https://doi.org/10.1029/2022gl098663

McGovern, A., Ebert-Uphoff, I., Gagne, D. J., & Bostrom, A. (2022). Why we need to focus on developing ethical, responsible, and trustworthy artificial intelligence approaches for environmental science. *Environmental Data Science*, *1*. https://doi.org/10.1017/eds.2022.5

Meyer, H., Reudenbach, C., Hengl, T., Katurji, M., & Nauss, T. (2018). Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. *Environmental Modelling & Software*, *101*, 1–9. https://doi.org/10.1016/j.envsoft.2017.12.001

Meyer, H., Reudenbach, C., Wöllauer, S., & Nauss, T. (2019). Importance of spatial predictor variable selection in machine learning applications – Moving from data reproduction to spatial prediction. In arXiv [stat.AP]. arXiv. Retrieved from http://arxiv.org/abs/1908.07805

Mobley, A., Linder, S. K., Braeuer, R., Ellis, L. M., & Zwelling, L. (2013). A survey on data reproducibility in cancer research provides insights into our limited ability to translate findings from the laboratory to the clinic. *PLoS One*, *8*(5), e63221. https://doi.org/10.1371/journal.pone.0063221

Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies*, *27*(5), 527–539. https://doi.org/10.1016/S0020-7373(87)80013-5

23335084, 2024, 7, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2023EA003364, Wiley Online Library on [02/06/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

NASEM. (2019). Reproducibility and replicability in science. *National Academies of Sciences, Engineering, and Medicine*.
Nasteski, V. (2017). An overview of the supervised machine learning methods. *Horizons*, *4*, 51–62. https://doi.org/10.20544/HORIZONS.B.04.1.17.P05
Neuendorf, K. A. (2016). *The content analysis guidebook*. Sage.
New York State Department of Transportation. (2023). *511NY*. Retrieved from https://511ny.org/
Northcutt, C. G., Athalye, A., & Mueller, J. (2021). Pervasive label errors in test sets destabilize machine learning benchmarks. In arXiv [stat.ML]. arXiv. Retrieved from http://arxiv.org/abs/2103.14749
Nüst, D., & Pebesma, E. (2021). Practical reproducibility in geography and geosciences. *Annals of the Association of American Geographers. Association of American Geographers*, *111*(5), 1300–1310. https://doi.org/10.1080/24694452.2020.1806028
O'Brien, T. A., Risser, M. D., Loring, B., Elbashandy, A. A., Krishnan, H., Johnson, J., et al. (2020). Detection of atmospheric rivers with inline uncertainty quantification: TECA-BARD v1.0.1. *Geoscientific Model Development*, *13*(12), 6131–6148. https://doi.org/10.5194/gmd-13-6131-2020
Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., et al. (2022). FourCastNet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *In* arXiv *[*physics.ao-ph*]. arXiv*. Retrieved from http://arxiv.org/abs/2202.11214
Pawlik, M., Hütter, T., Kocher, D., Mann, W., & Augsten, N. (2019). A link is not enough - reproducibility of data. *Datenbank-Spektrum : Zeitschrift Fur Datenbanktechnologie : Organ Der Fachgruppe Datenbanken Der Gesellschaft Fur Informatik e.V*, *19*(2), 107–115. https://doi.org/10.1007/s13222-019-00317-8
Peng, R. D. (2011). Reproducible research in computational science. *Science*, *334*(6060), 1226–1227. https://doi.org/10.1126/science.1213847
Pham, H. V., Qian, S., Wang, J., Lutellier, T., Rosenthal, J., Tan, L., et al. (2021). Problems and opportunities in training deep learning software systems: An analysis of variance. *Proceedings of the 35th IEEE/ACM International Conference on Automated Software Engineering*, 771–783. https://doi.org/10.1145/3324884.3416545
Prabhat, K., Mudigonda, M., Kim, S., Kapp-Schwoerer, L., Graubner, A., Karaismailoglu, E., et al. (2021). ClimateNet: An expert-labeled open dataset and deep learning architecture for enabling high-precision analyses of extreme weather. *Geoscientific Model Development*, *14*(1), 107–124. https://doi.org/10.5194/gmd-14-107-2021
Przybylo, V. (2023). Detecting the presence of precipitation in New York state Mesonet imagery at night using convolutional neural networks. AI2ES sitewide talk. Retrieved from https://www.ai2es.org/publications/ai2es-talks/
Raji, I. D., Bender, E. M., Paullada, A., Denton, E., & Hanna, A. (2021). AI and the everything in the whole wide world benchmark. In arXiv [cs.LG]. arXiv. Retrieved from http://arxiv.org/abs/2111.15366
Riffe, D., Kim, S., & Sobel, M. R. (2018). News borrowing revisited: A 50-year perspective. *Journalism & Mass Communication Quarterly*, *95*(4), 909–929. https://doi.org/10.1177/1077699018754909
Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., & Aroyo, L. M. (2021). "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–15. https://doi.org/10.1145/3411764.3445518
Shamsaliei, S., Gundersen, O. E., Alfredsen, K., & Halleraker, J. H. (2022). Towards historical analysis of riverscape development utilizing semantic segmentation. *Presented at the Workshop on Artificial Intelligence for Social Good, AI4SG-23, at AAAI*, *2023*.
Sobash, R. A., Gagne, D. J., Becker, C. L., Ahijevych, D., Gantos, G. N., & Schwartz, C. S. (2023). Diagnosing storm mode with deep learning in convection-allowing models. *Monthly Weather Review*, *151*(8), 2009–2027. https://doi.org/10.1175/MWR-D-22-0342.1
Stall, S., Cervone, G., Coward, C., Cutcher-Gershenfeld, J., Donaldson, T. J., Erdmann, C., et al. (2023). Ethical and responsible use of AI/ML in the earth, space, and environmental sciences. *Authorea Preprints*. https://doi.org/10.22541/essoar.168132856.66485758/v1
Stodden, V., McNutt, M., Bailey, D. H., Deelman, E., Gil, Y., Hanson, B., et al. (2016). Enhancing reproducibility for computational methods. *Science*, *354*(6317), 1240–1241. https://doi.org/10.1126/science.aah6168
Sutter, C., Sulia, K., Bassill, N. P., Thorncroft, C. D., Wirz, C. D., Przybylo, V., et al. (2023). Quantitative content analysis data for hand labeling road surface conditions in New York state department of transportation camera images. https://doi.org/10.5281/zenodo.8370665
Tatman, R., VanderPlas, J., & Dane, S. (2018). A practical taxonomy of reproducibility for machine learning research. rctatman.com. Retrieved from https://www.rctatman.com/files/Tatman_2018_ICML_poster.pdf
Togelius, J., & Yannakakis, G. N. (2023). Choose your weapon: Survival strategies for depressed AI academics. In arXiv [cs.OH]. arXiv. Retrieved from http://arxiv.org/abs/2304.06035
Valavi, R., Elith, J., Lahoz-Monfort, J. J., & Guillera-Arroita, G. (2019). Block CV: An r package for generating spatially or environmentally separated folds for k -fold cross-validation of species distribution models. *Methods in Ecology and Evolution / British Ecological Society*, *10*(2), 225–232. https://doi.org/10.1111/2041-210x.13107
Welty, C., Paritosh, P. K., & Aroyo, L. (2019). A metrological framework for evaluating crowd-powered instruments. Retrieved from https://www.humancomputation.com/2019/assets/papers/140.pdf
Weyn, J. A., Durran, D. R., Caruana, R., & Cresswell-Clay, N. (2021). Sub-seasonal forecasting with a large ensemble of deep-learning weather prediction models. *Journal of Advances in Modeling Earth Systems*, *13*(7). https://doi.org/10.1029/2021ms002502
Wheeler, M. C., & Hendon, H. H. (2004). An all-season real-time multivariate MJO index: Development of an index for monitoring and prediction. *Monthly Weather Review*, *132*(8), 1917–1932. https://doi.org/10.1175/1520-0493(2004)132<1917:AARMMI>2.0.CO;2
Wirz, C. D., Cains, M. G., Madlambayan, D., Demuth, J. L., & Bostrom, A. (2022). Trust and trustworthiness codebook for content analysis: An example from NWS Forecaster interviews about AI. https://doi.org/10.5281/zenodo.7113671
Wirz, C. D., Demuth, J. L., Cains, M. G., Bostrom, A., Schumacher, A., Madlambayan, D., et al. (2022). (Re)Conceptualizing the trustworthiness of AI as perceptual and context-dependent. Paper presented at the annual convention of the Society for Risk Analysis (SRA).
xCITE. (2023). *ExTreme collaboration, innovation, and technology laboratory*. University at Albany. Retrieved from https://www.albany.edu/asrc/xcite-laboratory
Xu, C., Doi, S. A. R., Zhou, X., Lin, L., Furuya-Kanamori, L., & Tao, F. (2022). Data reproducibility issues and their potential impact on conclusions from evidence syntheses of randomized controlled trials in sleep medicine. *Sleep Medicine Reviews*, *66*, 101708. https://doi.org/10.1016/j.smrv.2022.101708
Zhuang, D., Zhang, X., Song, S. L., & Hooker, S. (2021). Randomness in neural network training: Characterizing the impact of tooling. In arXiv [cs.LG]. arXiv. http://arxiv.org/abs/2106.11872

23335084, 2024, 7, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2023EA003364, Wiley Online Library on [02/06/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License