

---

# SmokeViz: Using Pseudo-Labels to Develop a Deep Learning Dataset of Wildfire Smoke Plumes in Satellite Imagery

---

Anonymous Author(s)

Affiliation

Address

email

## Abstract

1       The increase in the frequency of wildfires on a global scale underscores the need for  
2       advancements in fire monitoring techniques for disaster management, environmental  
3       protection and to mitigate negative health outcomes. This research introduces  
4       an innovative, data-driven framework that leverages the semi-supervised method,  
5       pseudo-labeling, to generate smoke plume annotations in geostationary satellite  
6       imagery. Unlike many pseudo-labeling applications that aim to increase the la-  
7       beled dataset size, the primary objective is use pseudo-labels to refine an existing  
8       National Oceanic and Atmospheric Administration smoke dataset that provides  
9       temporal and geographical information on individual smoke plumes but at variable  
10      and, primarily, low temporal resolution. We use deep learning and pseudo-labels to  
11      pinpoint the singular, most representative, satellite image that optimally illustrates  
12      the smoke annotation within the given time window. By identifying the most  
13      representative imagery of smoke plumes for a given smoke annotation, the study  
14      seeks to create an accurate and relevant machine learning dataset. The resulting  
15      dataset is anticipated to be an instrumental tool in developing further machine  
16      learning models, such as an automated system capable of real-time monitoring and  
17      annotation of smoke plumes directly from streaming satellite imagery.

18     

## 1 Introduction

19     In recent years, the escalation of wildfire incidents worldwide has become a prominent environmental  
20     and public health concern. The combustion process in wildfires releases smoke containing fine  
21     particulate matter (PM2.5) and harmful gases, posing severe hazards to human health and air quality.  
22     These risks underscore the necessity for efficient and effective monitoring methods to mitigate the  
23     adverse health impacts associated with wildfire smoke.

24     Traditionally, wildfire monitoring has relied on ground-based methods, such as forest service patrols,  
25     manned lookout towers, and aviation surveillance [1]. While these methods provide valuable localized  
26     insights, they are constrained by geographical and logistical limitations, often failing to deliver timely  
27     and comprehensive data, especially over large and remote areas. In contrast, satellite imagery offers  
28     a vantage point that overcomes these limitations, providing continuous, wide-area coverage and  
29     real-time data crucial for assessing and responding to the health risks posed by wildfire smoke.

30     Satellite imagery, equipped with state-of-the-art sensors, such as the Advanced Baseline Imager  
31     (ABI) on the Geostationary Operational Environmental Satellites (GOES) [8], have revolutionized  
32     environmental monitoring. These tools enable the detailed observation of smoke plumes, their  
33     particulate density, and the extent of smoke spread. These satellite-based systems offer the capabilities

34 to provide critical insights into the concentration and movement of smoke particulates, facilitating  
 35 real-time assessments of air quality.  
 36 The integration of satellite imagery in wildfire smoke monitoring is not only instrumental in providing  
 37 real-time data but also plays a significant role in public health planning and response. By mapping  
 38 the spread and density of smoke, health authorities can issue timely warnings, implement evacuation  
 39 protocols, and deploy resources effectively to mitigate health risks. Furthermore, long-term data  
 40 gathered from satellite observations can aid in understanding the broader impacts of wildfire smoke  
 41 on public health, influencing policy decisions and preventive measures.  
 42 Currently, multi-channel thresholding is a popular method to distinguish smoke pixels from pixels  
 43 containing dust, clouds or other phenomenon with similar signatures [28]. Thresholds are determined  
 44 by using historical, labeled data to extract optimal radiance values for each channel that corresponds  
 45 with the labeled class. These methods are tuned to particular biogeographies and often have issues  
 46 with generalization to new locations with varying fuel types [18].  
 47 In contrast to the numerical thresholding approach, human visual inspection of satellite imagery  
 48 is another commonly used method for smoke identification. Trained analyst will inspect satellite  
 49 imagery and label the smoke by hand. An example of hand labeled annotations is the National  
 50 Oceanic and Atmospheric Administration (NOAA) Hazard Mapping System (HMS) fire and smoke  
 51 product [15, 23]. For the HMS smoke product, trained satellite analysts use movement characteristics  
 52 to help identify smoke by scanning through a time series of satellite imagery. When visual inspection  
 53 indicates smoke, the analyst will draw a polygon that corresponds to the geolocation and density  
 54 of smoke. By design of the product, the HMS annotations have varying time resolution and are  
 55 released on a rolling but undefined schedule ranging from one to multiple times a day as observation  
 56 conditions permit. This method is potentially not as scalable as an automated approach and is limited  
 57 by the availability of analysts and their time.  
 58 To address the challenges associated with thresholding and manual labels, we can look towards  
 59 innovative approaches and recent technological advancements in computer vision. Machine learning  
 60 methods have shown potential in improving the accuracy and efficiency of satellite-based wildfire  
 61 smoke detection and monitoring. For instance, SmokeNet, uses a convolutional neural network  
 62 (CNN) based framework to determine if a scene of MODIS satellite imagery contains smoke [2].  
 63 Another study, that looked at a singular wildfire event, also used a CNN to identify smoke on a  
 64 pixel-wise basis using imagery from Himiware-8 [12]. Additionally, Wen et al. developed a CNN  
 65 architecture that takes GOES-East imagery as input and the HMS-generated annotations for the target  
 66 labels during training [26].  
 67 The success of deep learning methods, such as CNNs, relies heavily on the availability of a large,  
 68 representative dataset [24]. As laid out in table 1, existing methods use relatively small number of  
 69 samples, from 57 [25] to 6825 [26], where one sample represents a satellite image with a singular  
 70 time and geolocation. In contrast, benchmark datasets for image classification contain tens of  
 71 thousands (CIFAR-10 and MNIST) to millions (CIFAR-100 and ImageNet) of data samples [11],  
 72 [6], [5]. Keeping in mind the correlation between both the quality and quantity of data with model  
 73 performance, we introduce the largest known smoke dataset, SmokeViz, containing over 130,000  
 74 samples.

Table 1: Comparison of different studies including method used, dataset size, satellite source, number of channels used and if classification is performed at a pixel or image level.

Reference	Method	# Samples	Satellite	# Channels	Level
[2]	CNN	6255	MODIS	5	image
[26]	CNN	6825	GOES-East	5	pixel
[12]	CNN	975	Himiware-8	7	pixel
[25]	U-Net	47	Landsat-8	13	pixel
SmokeViz	U-Net	133,871	GOES-East/West	3	pixel

75 An approach to increase the number of labeled samples in a dataset, semi-supervised learning  
 76 leverages a labeled dataset to generate new labels for an often larger, but unlabeled, dataset. Pseudo-  
 77 labeling, a form of semi-supervised learning, uses labeled data to train an initial model, then runs  
 78 that model on unlabeled data to predict pseudo-labels, and finally trains a new model using the

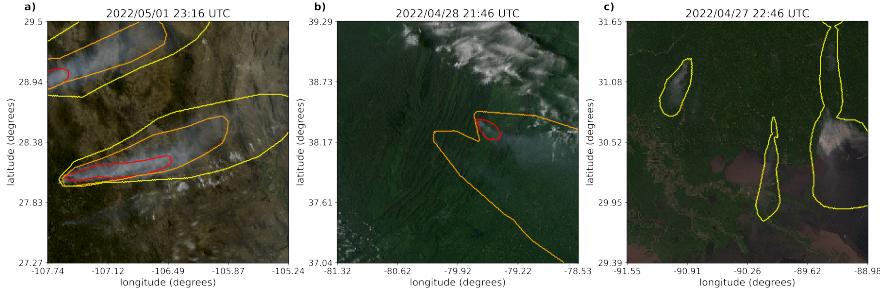


Figure 1: Satellite imagery captured by GOES-East within a few days of each other. The yellow, orange and red contours indicate the extent of Light, Medium and Heavy smoke. a) shows a canonical example of a smoke plume. b) and c) show observable variations in the density labels.

79 pseudo-labels [13]. We introduce a variation of pseudo-labeling not to increase the size, but to  
 80 increase the quality of our dataset by using pseudo-labels to choose the best satellite image out of a  
 81 given time-window to represent each smoke plume annotation.

## 82 2 Methods

### 83 Dataset

84 The initial data source, discussed in further detail in the HMS Smoke Labels section, is uniquely  
 85 characterized by each annotation having corresponding imagery ranging between 1-60 frames, where  
 86 each frame captures 5 minutes of exposure. Additionally, we have two satellites that overlap in  
 87 coverage area, GOES-East and GOES-West, effectively doubling the number of frames for a single  
 88 annotation. We apply pseudo-labeling to develop a dataset that has a one-to-one annotation-to-image  
 89 ratio, where we choose the satellite image that has the maximum overlap between the geolocation of  
 90 smoke in the imagery and the analyst annotation.

91 Dataset development came in three stages. First, we leverage light scattering physics to determine  
 92 which singular satellite image would be in the optimal configuration for smoke detection. Second, we  
 93 used that dataset to train an initial parent model that will identify smoke in satellite imagery. Third,  
 94 we use that parent model to label each satellite image in a given annotation’s time-window and the  
 95 optimal satellite image is chosen based on which image’s pseudo-labels has the greatest overlap with  
 96 the analyst annotation for the given location and densities of smoke.

### 97 HMS Smoke Labels

98 NOAA manages environmental satellite programs such as the HMS program, the HMS program is an  
 99 operational system that uses an aggregation of satellite data to generate active fire and smoke data.  
 100 To train our model, we implement a supervised learning framework that uses the HMS analyst smoke  
 101 product as truth labels during the model training process.

102 HMS smoke analysis data gives the coordinates of the smoke perimeter as a polygon and classifies  
 103 the smoke by density within a given time window. The time windows can range from instantaneous  
 104 (same start and end time) to lengths of 5 hours. While the true bounds of the smoke can change  
 105 within the larger time spans, the analyst is making an approximation that should reflect the smoke  
 106 coverage over the duration of the time window. The density information is qualitatively determined  
 107 by each analyst based on the apparent smoke opacity in the satellite imagery and categorized as either  
 108 light, medium or heavy as seen in figure 1a [16].

### 109 Thermometer Encoding Smoke Densities

110 One of the challenges introduced with using human generated qualitative smoke densities was that, as  
 111 seen in figure 1b and 1c, there are variations in what is labeled as heavy or light density smoke. More

112 generally, reproducing qualitative metrics with quantitative algorithms is a challenging problem, but  
 113 we apply mathematical approaches that mitigate some of the underlying complications of our specific  
 114 problem. Despite the fact that the smoke densities introduce qualitative complexities, we decided  
 115 that the density approximations were important to use in our dataset because of the differences in  
 116 signatures the densities produce. Within the satellite imagery, the appearance of a light density  
 117 smoke plume will look significantly different than a heavy density smoke plume as seen in figure 1.  
 118 Additionally, a light density smoke plume is expected to be more challenging to detect since it is easier  
 119 for it to be misclassified as not smoke. During the training process, the separate density categories  
 120 allows us to deferentially weight the penalization given to the model for incorrect classifications  
 121 based on category. For example, the model can be given a small penalization for misclassifying light  
 122 smoke as not smoke while given a higher penalization for misclassifying heavy smoke as not smoke.  
 123 In addition to the densities being ordered and categorical, the differences between the density  
 124 categories are not evenly distributed by a given metric, such as particulate matter per square meter.  
 125 The intervals between densities being unknown along with the hierarchical nature of the density labels  
 126 makes the labels ordinal instead of just categorical. This data property allows us to use thermometer  
 127 encoding [4], which leverages the idea that heavy density smoke includes both medium and light  
 128 density smoke, that heavy density smoke is closer to medium than it is to light and automatically  
 129 weights the loss functions and incorporates the ranked ordering of the densities. As seen in Table 2,  
 130 one-hot encoding, commonly used for categorical data, doesn't take ordinal properties of the data  
 131 into consideration.

Table 2: A comparison of one-hot encoding used for categorical data to thermometer encoding for ordinal data.

category	one-hot	thermometer
No Smoke	[0 0 0]	[0 0 0]
Light	[0 0 1]	[0 0 1]
Medium	[0 1 0]	[0 1 1]
Heavy	[1 0 0]	[1 1 1]

### 132 Time Windows For Smoke Annotations

133 In order to take into account movement characteristics to help identify smoke, analysts use multi-  
 134 frame animations of the satellite imagery. The resulting annotations often have large time windows  
 135 over multiple hours to represent one smoke plume annotation. Since the goal of these annotations is  
 136 to show the general coverage over that time span, as shown in figure 2, the smoke boundaries don't  
 137 often match up with the satellite imagery over the entire time window. One way to approach this  
 138 problem would be to use all the satellite images the analysts used as input. Since the timespans are  
 139 non-uniform, this would vary the length in imagery inputs into the model, which would be difficult  
 140 with a CNN architecture. Moreover, this would require a large amount of additional memory and  
 141 computational resources. Instead of using the original analysts' many satellite image inputs to one  
 142 annotated output, we develop a one-to-one input-to-output by finding the optimal singular satellite  
 143 image input to represent the annotation. Discussed in further detail in the next section, we do this  
 144 by making physics-driven choices on which satellite and timestamp would give the optimal angle  
 145 between the sun and satellite that would produce the strongest smoke signature for the geolocation  
 146 and timestamp of the smoke plume.

### 147 Satellite Imagery

148 The GOES satellites are operated by NOAA in order to support meteorology research and forecasting  
 149 for the United States. We use the latest operational satellites, GOES-16 (East), 17 and 18 (West)  
 150 that each carry the ABI, that measure 16 bands between the visible and infrared wavelengths. In  
 151 improvement to the GOES predecessors, imagery is collected every 5 minutes for the contiguous  
 152 United States and every 10 minutes for the full disk. Using PyTroll, a python framework for  
 153 processing satellite data [19], we input bands 1-3 (Table 3) to a GOES true color composite algorithm  
 154 to develop a true color image representation [3], similar to what is seen by HMS analysts.

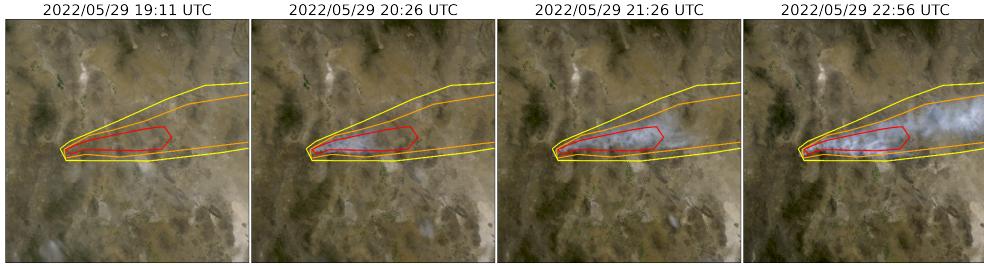


Figure 2: True color GOES-East imagery from May 2022, Southeast New Mexico ( $31^{\circ}\text{N}$ ,  $100^{\circ}\text{W}$ ) during the start of the Foster Fire. The red, orange and yellow lines represent the heavy, medium and low density HMS smoke annotations that span 19:10–23:00 UTC.

Table 3: To create a true color image, we use the following bands from the ABI Level 1b CONUS (ABI-L1b-RadC) product.

band	description	center wavelength	spatial resolution (km)
C01	blue visible	0.47	1
C02	red visible	0.64	0.5
C03	veggie near infrared	0.865	1

### 155 Mie-Derived Dataset

156 We used a physics-informed approach in selecting the initial dataset,  $\mathcal{D}_M$ , we call the Mie-derived  
 157 dataset, for training an initial parent model,  $f_p$ . Prior GOES ABI datasets for machine learning  
 158 applications often include data from only one of the two GOES-series satellites, commonly opting  
 159 for GOES-East [26], [17], [14]. Rather than using one satellite or the cumulative data from both  
 160 GOES-West and GOES-East images, we select between one or the other based on the solar zenith  
 161 angle. For smoke identification, this approach can achieve a much higher signal-to-noise than imaging  
 162 the earth’s surface from an arbitrary angle. The elastic scattering of light is the primary mechanism  
 163 to account for - while the atmosphere is composed of molecules with size  $< 1\text{nm}$ , smoke particles  
 164 can vary from  $100\text{ nm} - 10\text{ }\mu\text{m}$  in diameter,  $d$ . The GOES ABI covers spectral bands from  $0.47\text{ }\mu\text{m} -$   
 165  $13.3\text{ }\mu\text{m}$ , so atmospheric and smoke particle sizes occupy two very different regimes with respect  
 166 to the imaging wavelength  $\lambda$ . In the extreme limit of  $\lambda \gg d$ , the physics of scattering of light off a  
 167 small sphere is captured by Rayleigh scattering. This process has two critical consequences: (1) the  
 168 scattering cross section of light is strongly wavelength dependent (scaling with  $\lambda^{-4}$ ), meaning that  
 169 photons with wavelength closer to the ultraviolet are scattered more strongly than infrared photons. (2)  
 170 the scattering cross section scales with an angular dependent cross section of  $(1 + \cos^2 \theta)$ . Scattered  
 171 photons follow the emission distribution of a radiating dipole, scattering more strongly in the forward  
 172 and backwards directions ( $\theta = 0, \pi$ ) than orthogonal to the direction of propagation ( $\theta = \pi/2, 3\pi/2$ ),  
 173 see figure 3 for Rayleigh scattering schematic.

174 The significance of these scalings is that the observer, or detector, will receive blue photons in most  
 175 directions orthogonal to the source. Equivalently, photons traveling colinearly with line of sight to  
 176 the emission source will mostly have wavelengths in the infrared band. In the converse regime of  
 177  $d > \lambda$ , the elastic scattering of light against matter is modeled through Mie scattering. In comparison  
 178 to Rayleigh scattering, Mie scattering is largely wavelength independent and has a more complicated  
 179 radiation pattern where the cross section has a maximal amplitude in the forward direction. An  
 180 observer downstream of this scatterer will collect more photons than one positioned directly behind it.  
 181 In the context of smoke identification, a sunrise or sunset will lead to a higher Mie scattered signal in  
 182 GOES-West and GOES-East respectively, as shown with a smoke plume producing a stronger signal  
 183 in GOES-East imagery near sunset in figure 4.

184 Smoke identification therefore amounts to extracting a signal of  $d > \lambda$  photons from the  $\lambda \gg d$   
 185 background. Positioning a detector along line of sight to the scatterer will result in a higher signal

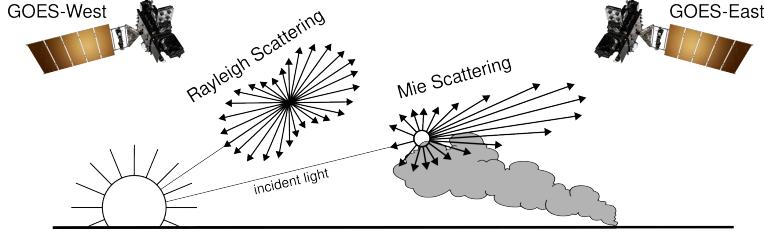


Figure 3: If the particle size is  $< \frac{1}{10}$  the wavelength of the interacting light, then the primary scattering will be Rayleigh. Mie scattering is the predominant scattering mechanism when the particle size is larger than wavelength of light. This schematic demonstrates that when the sun is setting in the West, the Mie scattering will predominately forward scatter towards GOES-East.

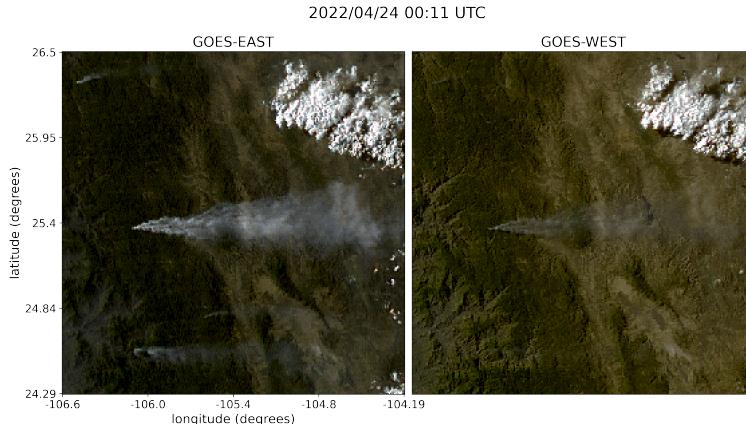


Figure 4: True color GOES-East (left) and GOES-West (right) imagery from April 24<sup>th</sup>, 2022 in Durango, Mexico. The images were taken  $\sim 0.5$  hours before sunset (01:43 UTC) for this geolocation and time of year.

186 from smoke particles (figure 3). Filtering the imaged wavelength can enhance this signal; photons  
 187 collected in the blue spectrum will have a naturally lower background along the line of sight to the  
 188 illumination source do their high level of Rayleigh scattering as. Therefore, as demonstrated in figure  
 189 5, this configuration results in the highest signal to noise imaging for smoke particles.

190 Based solely on these criteria, the optimal strategy would be to pull data from GOES-West right after  
 191 sunrise and from GOES-East right before sunset. Another factor to consider is that the time when the  
 192 sun is in optimal alignment with the satellite for smoke detection coincides with when solar zenith  
 193 angle is maximized. Larger angles between the satellite and sun result in an increase in noise due  
 194 to increased atmospheric interactions [22]. This is shown in figure 6, while we optimize for smoke  
 195 signal detection, due to the high solar zenith angle, we introduce atmospheric interaction noise that  
 196 obfuscate the smoke signal. To reduce the noise from large solar zenith angles, if given multiple  
 197 options to choose from, we choose the image with the largest solar zenith angle that is below 80  
 198 degrees.

199 The resulting image selection process takes into account atmospheric properties and light scattering  
 200 physics to generate an estimate of which singular satellite image within the analyst time-window  
 201 could give the highest smoke signal-to-noise ratio. The resulting Mie-derived dataset,  $\mathcal{D}_M$ , was then  
 202 used to train a model,  $f_p$ , that would generate  $N$  pseudo-labels,  $l^*$ , for every sample, where  $N$  is  
 203 determined by how many images, taken at a 10 minute interval, fit within the analyst time-window  
 204 for that sample. Chosen from the  $N$  images,  $x^*$  is the image with the highest alignment between the  
 205  $f_p$  prediction of smoke,  $l^*$ , in the image and the HMS analysts' annotation  $y^a$ .

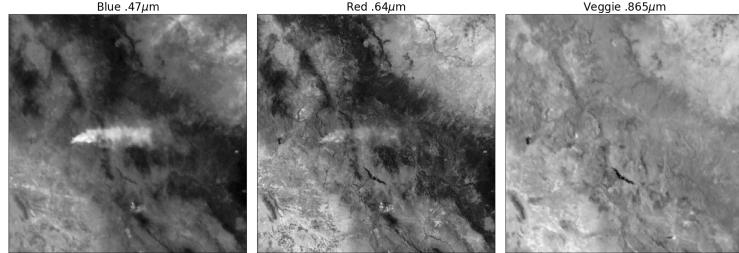


Figure 5: Three bands of GOES-East data are the raw input to generate a true color image. These plots show variations in the signal-to-noise ratio for smoke detection in relation to the wavelength,  $\lambda$ , of light being measured.

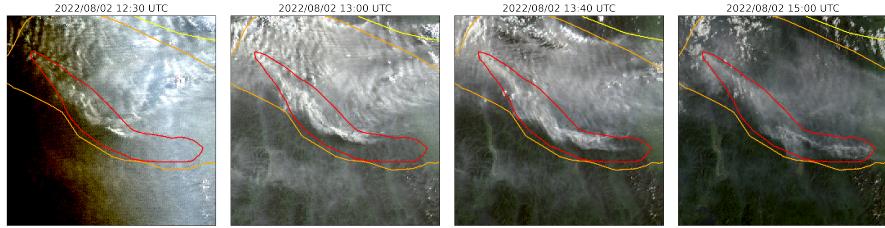


Figure 6: A smoke annotation projected onto GOES-West imagery from August 2022 that spans from 11:00 UTC to 15:00 UTC, sunrise on August 2nd, 2022 at coordinates (49°24'N, 115°29'W) was 12:15 UTC.

## 206 Machine Learning Model

207 We implement a deep learning architecture that uses the encoder from the ResNet model [9] and a  
 208 semantic segmentation classifier from the U-Net model [21]. Transfer learning has shown to reduce  
 209 the time and resources needed to train a model by leveraging information from pre-trained models  
 210 [27], [20]. We initialize the values of our model weights using the pre-trained values originally  
 211 trained on the ImageNet dataset [5], containing 1.2 million images and 1000 categories. Our model  
 212 was developed using the Segmentation Models PyTorch package [10] that was written as a high level  
 213 API for implementing models for semantic segmentation problems. We input 256x256x3 snapshots  
 214 of true color GOES imagery that contains smoke and output a 256x256x3 classification map that  
 215 predicts if a pixel contains smoke and if so, what the density of that smoke is. As mentioned earlier,  
 216 we apply the thermometer encoding shown in table 2 to encode the smoke densities and apply binary  
 217 cross entropy as the loss function per density of smoke.

218 The  $\mathcal{D}_M$  dataset contained over 130,000 samples. To train  $f_p$ , we split  $\mathcal{D}_M$  into training (118,691  
 219 samples), validation (8,100 samples) and testing (7,080) datasets. Training data contains data from  
 220 the years 2018, 2019, 2020, 2021 and 2023 while the data from 2022 is split into validation and  
 221 testing sets by taking data from alternating 10 days of the year. In order to make sure we include  
 222 the monthly variations in wildfire trends over a full year, we split 2022 data up by every 10 days.  
 223 This allowed us to: (1) allocate an additional full year of data for the training set, (2) show yearlong  
 224 trends in both the validation and testing sets and (3) keep the validation and testing datasets relatively  
 225 independent from one another since only two out of every ten days of data will have adjacent days in  
 226 validation and testing.

227 We trained the parent model,  $f_p$ , for 10 epochs, then ran  $f_p$  on all images,  $x_N$ , within the analyst  
 228 time-window for each annotation to select image that's pseudo-label best matched the HMS smoke

Table 4: IoU results per density of smoke and over all densities using the parent and child models and M.

	$f_p$		$f_c$	
	$\mathcal{D}_M$	$\mathcal{D}_{PL}$	$\mathcal{D}_M$	$\mathcal{D}_{PL}$
Light	0.394	0.551	0.418	0.538
Medium	0.283	0.392	0.340	0.411
Heavy	0.233	0.290	0.270	0.325
Overall	0.365	0.510	0.396	0.503

229 annotation,  $y^a$ . The candidate image,  $x^*$ , would have the potential be included in  $\mathcal{D}_{PL}$  only if it  
230 generated the highest Intersection over Union (IoU) value between the image's  $l^*$  and  $y^a$  over all  $x_N$ .  
231 The IoU metric is given by the ratio of area of overlap to the area of union as shown in equation 1.

$$IoU = \frac{|y^a \cap l^*|}{|y^a| \cup |l^*|} \quad (1)$$

232 To determine which image,  $x$ , out of the relevant imagery,  $x_N$ , for the given time window best  
233 represents the analyst annotation,  $y^a$ , we run  $f_p$  on each  $x$  to generate a pseudo-label,  $l^*$ . The output  
234 of  $f_p$ ,  $l^*$ , give predictions on if smoke is in the image, and if there is smoke, where the smoke is in  
235 that image and the density of that smoke.  $l^*$  serve as pseudo-labels for each density of smoke and  
236 are compared to the analyst annotations,  $y^a$ . To compare  $l^*$  and  $y^a$ , we calculate the IoU using the  
237 total set of pixels for  $l^*$  at that density of smoke and the entire set of pixels for  $y^a$  for a particular  
238 smoke density in each image. The image with the highest IoU score is chosen as the image,  $x^*$ ,  
239 that best represents the analyst smoke annotation,  $y^a$ . Often used for pseudo-labeling, a confidence  
240 threshold value is defined to determine if a pseudo-label should to be included in a dataset [7]. We  
241 chose a confidence threshold that would include the sample,  $x^*$ , in  $\mathcal{D}_{PL}$  if the maximum overall IoU  
242 (equation 2) between  $l^*$  and  $x^a$  over all densities was over 0.1.

243 Finally, we use  $\mathcal{D}_{PL}$  to train an additional child model,  $f_c$ . We use the same dataset split method and  
244 model setup but change  $\mathcal{D}_M$  to  $\mathcal{D}_{PL}$  to train the model over 10 epochs.

## 245 Results

246 To interpret the performance of  $f_p$ , we report the IoU metrics in table 4 that were computed by  
247 running  $f_p$  and  $f_c$  on  $\mathcal{D}_M$  and  $\mathcal{D}_{PL}$ . For each density, we calculate the IoU using the total set of  
248 pixels that  $f_p$  predicts as that density of smoke and the entire set of pixels labeled by the analyst  
249 as a particular smoke density over all imagery contained in the testing dataset. Additionally, we  
250 compute the overall IoU for all densities by first computing the number of pixels that intersect their  
251 corresponding density and divide that by the total number of pixels that make up the union of model  
252 predicted and analyst labeled smoke in the testing dataset.

$$IoU_{overall} = \frac{\sum_{\substack{i=light \\ heavy}}^{heavy} |y_i^a \cap l_i^*|}{\sum_{\substack{i=light \\ heavy}}^{heavy} |y_i^a| \cup |l_i^*|} \quad (2)$$

253 An illustration of a pseudo-label picked image better representing the analyst annotation when  
254 compared to the Mie-derived image selection is evident in Figure 7, where the heavy density smoke  
255 IoU increases from 0.01 to 0.59. The analyst annotation for these densities cover 5 hours of imagery,  
256 the Mie-derived selection optimizes for the image closest to sunrise while the pseudo-label image  
257 selection chooses the image with the highest overlap between the pseudo-label and the analyst  
258 annotation.

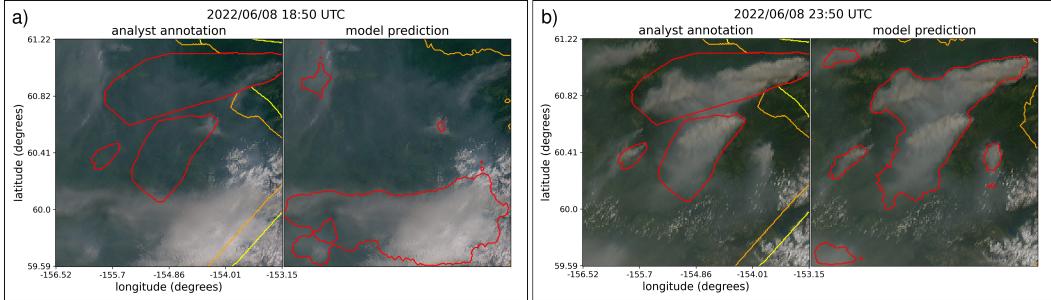


Figure 7: GOES-West imagery showing smoke on June 8th, 2022 in Alaska where, at this geolocation, daylight was between 12:43-7:53 UTC. The HMS smoke annotations displayed span from 18:50 to 23:50 UTC. a) shows the imagery that was selected using the Mie-derived data selection process b) shows the image that had the highest IoU score between the  $f_p$  generated pseudo-label and the analyst annotation.

### 259 3 Limitations

260 One of the concerns that comes with using pseudo-labeling methods is that you can perpetuate biases  
 261 from the parent model into subsequent child models. Due to the increase in detectable forward  
 262 scattered light off smoke particular matter, we expect the model to have a bias towards producing a  
 263 higher success rate for smoke detection at larger solar zenith angles. This could potentially cause  
 264 poor model performance when monitoring smoke during the middle of the day.

### 265 4 Conclusion

266 In this study, we have refined an existing dataset originally curated by NOAA’s HMS team, trans-  
 267 forming it from a many-to-one imagery-to-annotation format to a, more succinct, one-to-one satellite  
 268 image-to-annotation dataset. The initial HMS dataset primarily provided a general approximation  
 269 of where smoke had been present for a given time window, though it did not guarantee the actual  
 270 existence of smoke in the labeled pixels during the given times. Our goal was to create a dataset  
 271 that could be used, along with additional applications, to train a model to detect wildfire smoke in  
 272 real-time on an image-by-image level. The Mie-derived dataset selection process determines that if  
 273 smoke is present, what timestamp within the analyst time window would give the highest smoke  
 274 signal-to-noise ratio. While optimizing for being able to detect smoke, if it is present, the Mie-dataset  
 275 selection had no metric to determine if the smoke was effectually present in the selected image. Since  
 276 many of the images within the HMS time-window either contained no smoke at all or the smoke was  
 277 not contained within the geospatial bounds of the annotations, the Mie-derived dataset contained  
 278 a large number of mislabeled samples. Discrepancies between data and labels can be detrimental  
 279 towards the model’s capacity to improve on feature representations in the target domain. During  
 280 model training, the penalization of accurate predictions can inadvertently introduce biases towards  
 281 misclassifying noise as meaningful signal.

282 To improve the dataset’s capacity to accurately represent wildfire smoke plumes, we train a parent  
 283 machine learning model,  $f_p$ , using the Mie-derived dataset,  $\mathcal{D}_M$ , and run it on the relevant satellite  
 284 images within the time-frame. The image with the maximum IoU score between the model’s smoke  
 285 predictions, or pseudo-label, and the analyst smoke annotations are used to create the pseudo-label  
 286 generated dataset,  $\mathcal{D}_{PL}$ . We then train a child model,  $f_c$ , using  $\mathcal{D}_{PL}$  and test  $f_p$  and  $f_c$  on both the  
 287 2022 testing sets from  $\mathcal{D}_M$  and  $\mathcal{D}_{PL}$ . The results reported in table 4 suggest that  $\mathcal{D}_{PL}$  was able to  
 288 train a better performing model,  $f_c$  that gave higher IoU metrics on both dataset’s testing sets in  
 289 comparison to the original parent model,  $f_p$ .

290 The result of this study is a representative dataset that can be used to train machine learning models  
 291 for various wildfire smoke applications. A future goal is to produce a robust and reliable machine  
 292 learning based approach for detecting wildfires using satellite imagery. That information can be used  
 293 for wildfire monitoring and as data provided to public health officials for air quality assessments. On  
 294 a broader scale, we show how pseudo-labeling can be used to optimize a dataset when the resolution

295 for the data and corresponding labels do not match. This could be useful in similar applications  
296 involving time-series/video data with a singular label where the data can be compressed while still  
297 remaining representative of the label.

## 298 **5 Acknowledgments and Disclosure of Funding**

299 This research was supported in part by NOAA cooperative agreement NA22OAR4320151, for the  
300 Cooperative Institute for Earth System Research and Data Science (CIESRDS). We thank Wilfrid  
301 Schroeder and the Hazard Mapping Systems team for giving guidance on how they created their  
302 smoke plume dataset. This work utilized the Alpine high performance computing resource at the  
303 University of Colorado Boulder. Alpine is jointly funded by the University of Colorado Boulder, the  
304 University of Colorado Anschutz, Colorado State University, and the National Science Foundation  
305 (award 2201538). The statements, findings, conclusions, and recommendations are those of the  
306 author(s) and do not necessarily reflect the views of NOAA or the U.S. Department of Commerce.

## 307 **References**

- 308 [1] R. S. Allison, J. M. Johnston, G. Craig, and S. Jennings. Airborne optical and thermal remote  
309 sensing for wildfire detection and monitoring. *Sensors*, 16(8):1310, 2016.
- 310 [2] R. Ba, C. Chen, J. Yuan, W. Song, and S. Lo. Smokenet: Satellite smoke scene detection using  
311 convolutional neural network with spatial and channel-wise attention. *Remote Sensing*, 11(14):  
312 1702, 2019.
- 313 [3] M. Bah, M. Gunshor, and T. Schmit. Generation of goes-16 true color imagery without a green  
314 band. *Earth and Space Science*, 5(9):549–558, 2018.
- 315 [4] J. Buckman, A. Roy, C. Raffel, and I. Goodfellow. Thermometer encoding: One hot way to  
316 resist adversarial examples. In *International conference on learning representations*, 2018.
- 317 [5] J. Deng, K. Li, M. Do, H. Su, and L. Fei-Fei. Construction and Analysis of a Large Scale Image  
318 Ontology. Vision Sciences Society, 2009.
- 319 [6] L. Deng. The mnist database of handwritten digit images for machine learning research [best of  
320 the web]. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012. doi: 10.1109/MSP.2012.  
321 2211477.
- 322 [7] R. E. Ferreira, Y. J. Lee, and J. R. Dórea. Using pseudo-labeling to improve performance of  
323 deep neural networks for animal identification. *Scientific Reports*, 13(1):13875, 2023.
- 324 [8] S. J. Goodman, T. J. Schmit, J. Daniels, and R. J. Redmon. *The GOES-R series: a new  
325 generation of geostationary environmental satellites*. Elsevier, 2019.
- 326 [9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition, 2015.
- 327 [10] P. Iakubovskii. Segmentation models pytorch. [https://github.com/qubvel/segmentation\\_models.pytorch](https://github.com/qubvel/segmentation_models.pytorch), 2019.
- 328 [11] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. 2009.
- 329 [12] A. Larsen, I. Hanigan, B. J. Reich, Y. Qin, M. Cope, G. Morgan, and A. G. Rappold. A deep  
330 learning approach to identify smoke plumes in satellite imagery in near-real time for health risk  
331 communication. *Journal of exposure science & environmental epidemiology*, 31(1):170–176,  
332 2021.
- 333 [13] D.-H. Lee. Pseudo-label : The simple and efficient semi-supervised learning method for deep  
334 neural networks. *ICML 2013 Workshop : Challenges in Representation Learning (WREPL)*, 07  
335 2013.
- 336 [14] Y. Lee, C. D. Kummerow, and I. Ebert-Uphoff. Applying machine learning methods to detect  
337 convection using geostationary operational environmental satellite-16 (goes-16) advanced  
338 baseline imager (abi) data. *Atmospheric Measurement Techniques*, 14(4):2699–2716, 2021.

- 340 [15] D. McNamara, G. Stephens, M. Ruminski, and T. Kasheta. The hazard mapping system (hms) -  
341 noaa's multi-sensor fire and smoke detection program using environmental satellites. *Conference*  
342 *on Satellite Meteorology and Oceanography*, 01 2004.
- 343 [16] NOAA. Hazard mapping system fire and smoke product. URL <https://www.ospo.noaa.gov/Products/land/hms.html#about>.
- 345 [17] T. C. Phan and T. T. Nguyen. Remote sensing meets deep learning: exploiting spatio-temporal-  
346 spectral satellite images for early wildfire detection. 2019.
- 347 [18] T. Randriambelo, S. Baldy, M. Bessafi, M. Petit, and M. Despinoy. An improved detection  
348 and characterization of active fires and smoke plumes in south-eastern africa and madagascar.  
349 *International Journal of Remote Sensing*, 19(14):2623–2638, 1998.
- 350 [19] M. Raspaud, D. Hoese, A. Dybbroe, P. Lahtinen, A. Devasthale, M. Itkin, U. Hamann, L. Ø.  
351 Rasmussen, E. S. Nielsen, T. Leppelt, et al. Pytroll: An open-source, community-driven python  
352 framework to process earth observation satellite data. *Bulletin of the American Meteorological Society*,  
353 99(7):1329–1336, 2018.
- 354 [20] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: an  
355 astounding baseline for recognition, 2014.
- 356 [21] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image  
357 segmentation, 2015.
- 358 [22] A. Royer, P. Vincent, and F. Bonn. Evaluation and correction of viewing angle effects on  
359 satellite measurements of bidirectional reflectance. *Photogrammetric engineering and remote*  
360 *sensing*, 51(12):1899–1914, 1985.
- 361 [23] W. Schroeder, M. Ruminski, I. Csizar, L. Giglio, E. Prins, C. Schmidt, and J. Morisette.  
362 Validation analyses of an operational fire monitoring product: The hazard mapping system.  
363 *International Journal of Remote Sensing*, 29(20):6059–6066, 2008.
- 364 [24] C. Sun, A. Shrivastava, S. Singh, and A. Gupta. Revisiting unreasonable effectiveness of data in  
365 deep learning era, 2017.
- 366 [25] Z. Wang, P. Yang, H. Liang, C. Zheng, J. Yin, Y. Tian, and W. Cui. Semantic segmentation and  
367 analysis on sensitive parameters of forest fire smoke using smoke-unet and landsat-8 imagery.  
368 *Remote Sensing*, 14(1):45, 2022.
- 369 [26] J. Wen and M. Burke. Wildfire smoke plume segmentation using geostationary satellite imagery.  
370 *ArXiv*, abs/2109.01637, 2021. URL <https://api.semanticscholar.org/CorpusID:237416777>.
- 372 [27] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural  
373 networks?, 2014.
- 374 [28] T. X.-P. Zhao, S. Ackerman, and W. Guo. Dust and smoke detection for multi-channel imagers.  
375 *Remote Sensing*, 2(10):2347–2368, 2010. ISSN 2072-4292. doi: 10.3390/rs2102347. URL  
376 <https://www.mdpi.com/2072-4292/2/10/2347>.