

---

# 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  
2       for advancements in fire monitoring techniques for disaster management, environmental  
3       protection and to mitigate negative health outcomes. This research  
4       introduces an innovative, data-driven framework that leverages the semi-supervised  
5       method, pseudo-labeling, to generate smoke plume annotations in geostationary  
6       satellite imagery. Unlike many pseudo-labeling application that want to increase  
7       the 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. While these methods provide valuable local  
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 advanced sensors, such as the Advanced Baseline Imager (ABI) on  
31     the Geostationary Operational Environmental Satellites (GOES), have revolutionized environmental  
32     monitoring. These tools enable the detailed observation of smoke plumes, their particulate density,  
33     and the extent of smoke spread. These satellite-based systems offer the capabilities to provide critical

34 insights into the concentration and movement of smoke particulates, facilitating accurate and timely  
 35 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 [24]. The method uses historical,  
 44 labeled data to extract optimal radiance values for each channel that corresponds with the labeled  
 45 class. These methods are tuned to particular biogeographies and often have issues with generalization  
 46 to new locations with varying fuel types [14].  
 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 [11, 19]. 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 these challenges we can look towards innovative approaches and technological advancements  
 59 in computer vision. Machine learning methods have shown potential in improving the accuracy  
 60 and efficiency of satellite-based wildfire smoke detection and monitoring. For instance, SmokeNet,  
 61 uses a convolutional neural network (CNN) based framework to determine if a scene of MODIS  
 62 imagery contains smoke [1]. Another study also used a CNN to identify smoke on a pixel-wise basis  
 63 using imagery from Himiwi-8 [8]. Additionally, Wen et al. developed a CNN architecture that  
 64 takes GOES-East imagery as input and NOAA-generated annotations for the target labels during  
 65 training [22].  
 66 The success of deep learning methods, such as CNNs, relies heavily on the availability of a large,  
 67 representative dataset [20]. As laid out in table 1, existing methods use relatively small number of  
 68 samples, from 57 [21] to 6825 [22], where one sample represents a satellite image with a singular time  
 69 and geolocation. In contrast, benchmark datasets for image classification contain tens of thousands  
 70 (CIFAR-10 and MNIST) to millions (CIFAR-100 and ImageNet) of data samples. Keeping in mind  
 71 the correlation between both the quality and quantity of data with model performance, we introduce  
 72 the largest known smoke dataset, SmokeViz, containing over 120,000 samples.

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

Reference	Method	# Samples	Satellite	# Channels	Level
[1]	CNN	6255	MODIS	5	image
[22]	CNN	6825	GOES-East	5	pixel
[8]	CNN	975	Himiwi-8	7	pixel
[21]	U-Net	47	Landsat-8	13	pixel
SmokeViz	U-Net	120,000	GOES-East/West	3	pixel

73 An approach to increase the number of labeled samples in a dataset, semi-supervised learning  
 74 leverages a labeled dataset to generate new labels for an often larger, but unlabeled, dataset. Pseudo-  
 75 labeling, a form of semi-supervised learning, uses labeled data to train an initial model, then runs  
 76 that model on unlabeled data to predict pseudo-labels, and finally trains a new model using the  
 77 pseudo-labels [9]. We introduce a variation of pseudo-labeling not to increase the size, but to increase

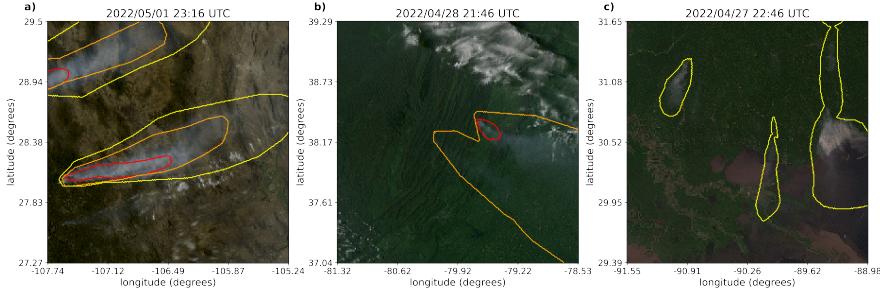


Figure 1: Satellite imagery captured by GOES-East within a few days of each other. The yellow, orange and red contours indicate the extent of Light, Medium and Heavy smoke. a) shows a canonical example of a smoke plume. b) and c) show variations in the density labels.

78 the quality of our dataset by using pseudo-labels to choose the best satellite image out of a given  
79 time-window to represent each smoke plume annotation.

## 80 2 Methods

### 81 Dataset

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

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

### 95 Smoke Labels

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

100 HMS smoke analysis data gives the coordinates of the smoke perimeter as a polygon and classifies  
101 the smoke by density within a given time window. The time windows can range from instantaneous  
102 (same start/end time) to lengths of 5 hours. While the bounds of the smoke annotations can change  
103 within the larger time spans, the analyst is making an approximation that should reflect the smoke  
104 coverage over the duration of the time window. The density information is qualitatively determined  
105 by each analyst based on the apparent smoke opacity in the satellite imagery and categorized as either  
106 light, medium or heavy as seen in figure 1a [12].

### 107 Thermometer Encoding Smoke Densities

108 One of the challenges introduced with using human generated qualitative smoke densities was that, as  
109 seen in figure 1b and 1c, there are variations in what is labeled as heavy or light density smoke. More  
110 generally, reproducing qualitative metrics with quantitative algorithms is a challenging problem, but

111 we apply mathematical approaches that mitigate some of the underlying complications of our specific  
 112 problem. Despite the fact that the smoke densities introduce qualitative complexities, we decided  
 113 that the density approximations were important to use in our dataset because of the differences in  
 114 signatures the densities produce. Within the satellite imagery, the appearance of a light density  
 115 smoke plume will look significantly different than a heavy density smoke plume as seen in figure 1.  
 116 Additionally, a light density smoke plume is expected to be more challenging to detect since it is easier  
 117 for it to be misclassified as not smoke. During the training process, the separate density categories  
 118 allows us to deferentially weight the penalization given to the model for incorrect classifications  
 119 based on category. For example, the model can be given a small penalization for misclassifying light  
 120 smoke as not smoke while given a higher penalization for misclassifying heavy smoke as not smoke.  
 121 In addition to the densities being ordered and categorical, the differences between the density  
 122 categories are not evenly distributed by a given metric, such as particulate matter per square meter.  
 123 The intervals between densities being unknown along with the hierarchical nature of the density labels  
 124 makes the labels ordinal instead of just categorical. This data property allows us to use thermometer  
 125 encoding [3], which leverages the idea that heavy density smoke includes both medium and light  
 126 density smoke, that heavy density smoke is closer to medium than it is to light and automatically  
 127 weights the loss functions and incorporates the ranked ordering of the densities. As seen in Table 2,  
 128 one-hot encoding, commonly used for categorical data, doesn't take ordinal properties of the data  
 129 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]

### 130 Time Windows For Smoke Annotations

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

### 145 Satellite Imagery

146 The Geostationary Operational Environmental Satellites (GOES) are operated by the NOAA in order  
 147 to support meteorology research and forecasting for the United States. We use the latest operational  
 148 satellites, GOES-16 (East), 17 and 18 (West) that carry the ABI, that measure 16 bands between the  
 149 visible and infrared wavelengths. In improvement to the GOES predecessors, imagery is collected  
 150 every 5 minutes for the contiguous United States and every 10 minutes for the full disk. We use bands  
 151 1-3 (Table 3) as input to Satpy's composite algorithm to develop a true color image representation,  
 152 similar to what is used as input by HMS analysts [15] and [2].

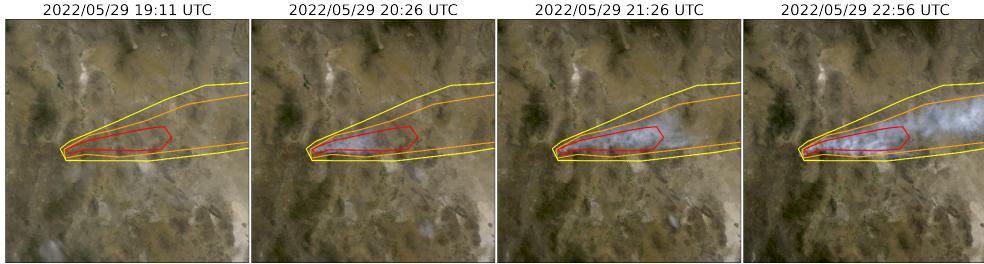


Figure 2: True Color GOES-East imagery from May 2022, Southeast New Mexico ( $31^{\circ}\text{N}$ ,  $100^{\circ}\text{W}$ ) during the start of the Foster Fire. The red, orange and yellow lines represent the heavy, medium and low density HMS smoke annotations that span 19:10–23:00 UTC.

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

### 153 Mie-Derived Dataset

154 We used a physics-informed approach in selecting the initial dataset,  $\mathcal{D}_M$ , we call the Mie-derived  
 155 dataset, for training an initial parent model,  $f_p$ . Prior GOES ABI datasets for machine learning  
 156 applications often include data from only one of the two GOES-series satellites, commonly opting  
 157 for GOES-East [22], [13], [10]. Rather than using one satellite or the cumulative data from both  
 158 GOES-West and GOES-East images, we select between one or the other based on the solar zenith  
 159 angle. For smoke identification, this approach can achieve a much higher signal-to-noise than imaging  
 160 the earth’s surface from an arbitrary angle. The elastic scattering of light is the primary mechanism  
 161 to account for - while the atmosphere is composed of molecules with size  $< 1\text{nm}$ , smoke particles  
 162 can vary from  $100\text{ nm} - 10\text{ }\mu\text{m}$  in diameter,  $d$ . The GOES ABI covers spectral bands from  $0.47\text{ }\mu\text{m} -$   
 163  $13.3\text{ }\mu\text{m}$ , so atmospheric and smoke particle sizes occupy two very different regimes with respect  
 164 to the imaging wavelength  $\lambda$ . In the extreme limit of  $\lambda \gg d$ , the physics of scattering of light off a  
 165 small sphere is captured by Rayleigh scattering. This process has two critical consequences: (1) the  
 166 scattering cross section of light is strongly wavelength dependent (scaling with  $\lambda^{-4}$ ), meaning that  
 167 photons with wavelength closer to the ultraviolet are scattered more strongly than infrared photons. (2)  
 168 the scattering cross section scales with an angular dependent cross section of  $(1 + \cos^2 \theta)$ . Scattered  
 169 photons follow the emission distribution of a radiating dipole, scattering more strongly in the forward  
 170 and backwards directions ( $\theta = 0, \pi$ ) than orthogonal to the direction of propagation ( $\theta = \pi/2, 3\pi/2$ ),  
 171 see figure 3 for Rayleigh scattering schematic.

172 The significance of these scalings is that the observer, or detector, will receive blue photons in most  
 173 directions orthogonal to the source. Equivalently, photons traveling colinearly with line of sight to  
 174 the emission source will mostly have wavelengths in the infrared band. In the converse regime of  
 175  $d > \lambda$ , the elastic scattering of light against matter is modeled through Mie scattering. In comparison  
 176 to Rayleigh scattering, Mie scattering is largely wavelength independent and has a more complicated  
 177 radiation pattern where the cross section has a maximal amplitude in the forward direction. An  
 178 observer downstream of this scatterer will collect more photons than one positioned directly behind it.  
 179 In the context of smoke identification, a sunrise or sunset will lead to a higher Mie scattered signal in  
 180 GOES-West and GOES-East respectively, as shown with a smoke plume producing a stronger signal  
 181 in GOES-East imagery near sunset in figure 2.

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

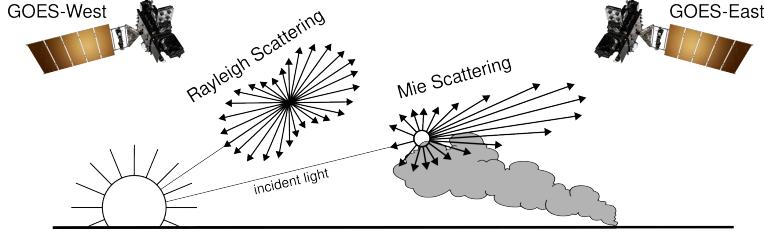


Figure 3: If the particle size is  $< \frac{1}{10}$  the wavelength of the interacting light, then the primary scattering will be Rayleigh. Mie scattering is the predominant scattering mechanism when the particle size is larger than wavelength of light. This schematic demonstrates that when the sun is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

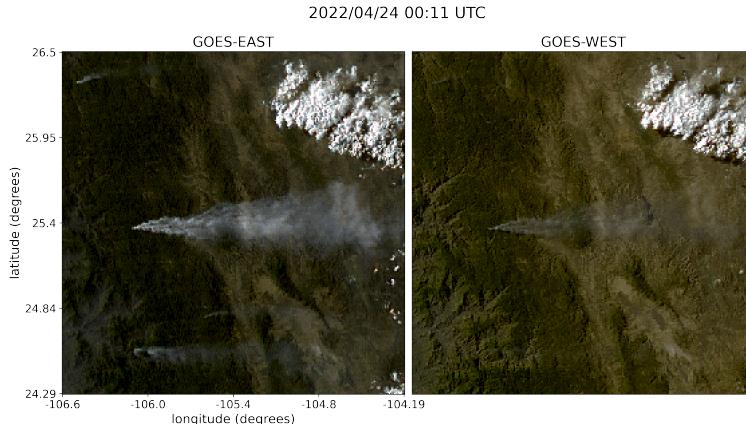


Figure 4: True Color GOES-East (left) and GOES-West (right) imagery from April 24<sup>th</sup>, 2022 in Durango, Mexico. The images were taken  $\sim 0.5$  hours before sunset (01:43 UTC) for this geolocation and time of year.

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

188 Based solely on these criteria, the optimal strategy would be to pull data from GOES-West right after  
 189 sunrise and from GOES-East right before sunset. Another factor to consider is that the time when the  
 190 sun is in optimal alignment with the satellite for smoke detection coincides with when solar zenith  
 191 angle is maximized. Larger angles between the satellite and sun result in an increase in noise due  
 192 to increased atmospheric interactions [18]. This is shown in figure 6, while we optimize for smoke  
 193 signal detection, due to the high solar zenith angle, we introduce atmospheric interaction noise that  
 194 obfuscates the smoke signal. To reduce the noise from large solar zenith angles, if given multiple  
 195 options to choose from, we choose the image with the largest solar zenith angle that is below 80  
 196 degrees.

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

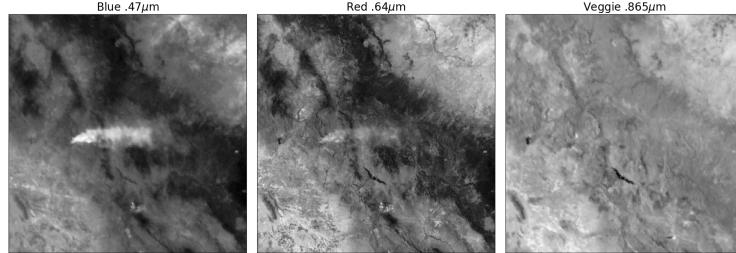


Figure 5: Three bands of GOES-East data are the raw input to generate the True Color image shown in figure 4. These plots show variations in signal-to-noise ratio for smoke detection in relation to the wavelength,  $\lambda$ , of light being measured.

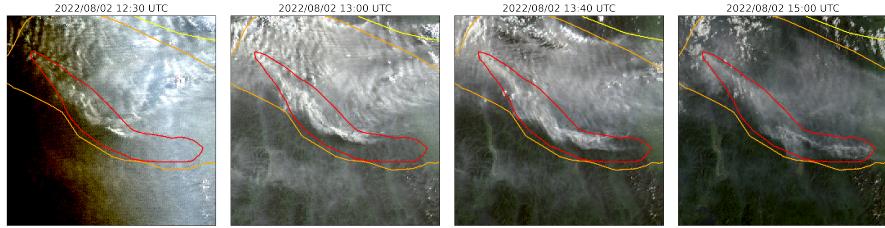


Figure 6: A smoke annotation projected onto GOES-West imagery from August 2022 that spans from 11:00 UTC to 15:00 UTC, sunrise on August 2nd, 2022 at coordinates ( $49^{\circ}24'N$ ,  $115^{\circ}29'W$ ) was 12:15 UTC.

#### 204 Machine Learning Model

205 We implement a deep learning architecture that uses the encoder from the ResNet model [6] and a  
 206 semantic segmentation classifier from the U-Net model [17]. Transfer learning has shown to reduce  
 207 the time and resources needed to train a model by leveraging information from pre-trained models  
 208 [23], [16]. We initialize the values of our model weights using the pre-trained values originally  
 209 trained on the ImageNet dataset [4], containing 1.2 million images and 1000 categories. Our model  
 210 was developed using the Segmentation Models PyTorch package [7] that was written as a high level  
 211 API for implementing models for semantic segmentation problems. We input 256x256x3 snapshots  
 212 of True Color GOES imagery that contains smoke and output a 256x256x3 classification map that  
 213 predicts if a pixel contains smoke and if so, what the density of that smoke is. As mentioned earlier,  
 214 we apply the thermometer encoding shown in table 2 to encode the smoke densities and apply binary  
 215 cross entropy as the loss function per density of smoke.

216 The  $\mathcal{D}_M$  dataset contained over 120,000 samples. To train  $f_p$ , we split  $\mathcal{D}_M$  into training (95,000  
 217 samples), validation (12,000 samples) and testing (12,000) datasets. Training data contains data from  
 218 the years 2018, 2019, 2020, 2021 and 2023 while the data from 2022 is split into validation and  
 219 testing sets by taking data from alternating 10 days of the year. In order to make sure we include  
 220 the monthly variations in wildfire trends over a full year, we split 2022 data up by every 10 days.  
 221 This allowed us to: (1) allocate an additional full year of data for the training set, (2) show yearlong  
 222 trends in both the validation and testing sets and (3) keep the validation and testing datasets relatively  
 223 independent from one another.

224 We trained the parent model,  $f_p$ , for 10 epochs, then ran  $f_p$  on all images,  $x_N$ , within the analyst  
 225 time-window for each annotation to select image that's pseudo-label best matched the HMS smoke  
 226 annotation,  $y^a$ . An image,  $x^*$ , would have the potential be included in  $\mathcal{D}_{PL}$  only if it generated the

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

227 highest Intersection over Union (IoU) value between the image's  $l^*$  and  $y^a$  over all  $x_N$ . The IoU  
 228 metric is given by the ratio of area of overlap to the area of union as shown in equation 1.

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

229 To determine which image,  $x$ , out of the relevant imagery,  $x_N$ , for the given time window best  
 230 represents the analyst annotation,  $y^a$ , we run  $f_p$  on each  $x$  to generate a pseudo-label,  $l^*$ . The output  
 231 of  $f_p$ ,  $l^*$ , give predictions on if smoke is in the image, and if there is smoke, where the smoke is in  
 232 that image and the density of that smoke.  $l^*$  serve as pseudo-labels for each density of smoke and  
 233 are compared to the analyst annotations,  $y^a$ . To compare  $l^*$  and  $y^a$ , we calculate the IoU using the  
 234 total set of pixels for  $l^*$  at that density of smoke and the entire set of pixels for  $y^a$  for a particular  
 235 smoke density in each image. The image with the highest IoU score is chosen as the image,  $x^*$ ,  
 236 that best represents the analyst smoke annotation,  $y^a$ . Often used for pseudo-labeling, a confidence  
 237 threshold value is defined to determine if a pseudo-label should to be included in a dataset [5]. We  
 238 chose a confidence threshold that would include the sample,  $x^*$ , in  $\mathcal{D}_{PL}$  if the maximum overall IoU  
 239 (equation 2) between  $l^*$  and  $x^a$  over all densities was over 0.1.

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

## 242 Results

243 To interpret the performance of  $f_p$ , we report the IoU metrics in table 4 that were computed by  
 244 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  
 245 pixels that  $f_p$  predicts as that density of smoke and the entire set of pixels labeled by the analyst  
 246 as a particular smoke density over all imagery contained in the testing dataset. Additionally, we  
 247 compute the overall IoU for all densities by first computing the number of pixels that intersect their  
 248 corresponding density and divide that by the total number of pixels that make up the union of model  
 249 predicted and analyst labeled smoke in the testing dataset.

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

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

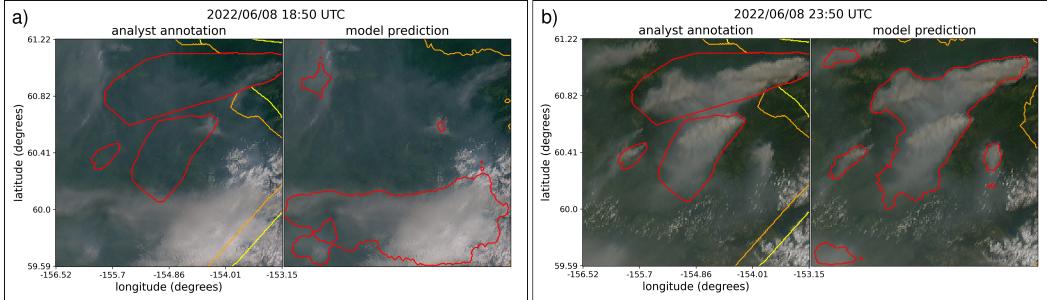


Figure 7: GOES-West imagery showing smoke on June 8th, 2022 in Alaska where, at this geolocation, daylight was between 12:43-7:53 UTC. The HMS smoke annotations displayed span from 18:50 to 23:50 UTC. a) shows the imagery that was selected using the Mie-derived data selection process b) shows the image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

### 256 3 Limitations

257 One of the concerns that comes with using pseudo-labeling methods is that you can perpetuate biases  
 258 from the parent model into subsequent child models. We are not using the pseudo-labels to label  
 259 unlabeled data, but to try to select the image that best represents the label out of a candidate of images.  
 260 Due to the increase in detectable forward scattered light off smoke particular matter, we expect the  
 261 model to have a bias towards producing a higher success rate for smoke detection at larger solar  
 262 zenith angles. This could potentially cause issues with monitoring smoke during middle of the day.

### 263 4 Conclusion

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

280 To improve the dataset’s capacity to accurately represent wildfire smoke plumes, we train a parent  
 281 machine learning model,  $f_p$ , using the Mie-derived dataset,  $\mathcal{D}_M$ , and run it on the relevant satellite  
 282 images within the time-frame. The image with the maximum IoU score between the model’s smoke  
 283 predictions, or pseudo-label, and the analyst smoke annotations are used to create the pseudo-label  
 284 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  
 285 2022 testing sets from  $\mathcal{D}_M$  and  $\mathcal{D}_{PL}$ . The results reported in table 2 suggest that  $\mathcal{D}_{PL}$  was able to  
 286 train a better performing model,  $f_c$  that gave higher IoU metrics on both dataset’s testing sets in  
 287 comparison to the original parent model,  $f_p$ .

288 The result of this study is a representative dataset that can be used to train machine learning models  
 289 for various wildfire smoke applications. The end goal is to produce a robust and reliable machine  
 290 learning based approach for detecting wildfires using satellite imagery. That information can be used  
 291 for wildfire monitoring and as data provided to public health officials for air quality assessments.

292 **5 Acknowledgments and Disclosure of Funding**

293 This work was partially supported by the NOAA Global Systems Laboratory and Cooperative Institute  
294 for Research in Environmental Sciences at the University of Colorado Boulder. We thank Wilfrid  
295 Schroeder and the Hazard Mapping Systems team for giving guidance on how they created their  
296 smoke plume dataset. This work utilized the Alpine high performance computing resource at the  
297 University of Colorado Boulder. Alpine is jointly funded by the University of Colorado Boulder, the  
298 University of Colorado Anschutz, Colorado State University, and the National Science Foundation  
299 (award 2201538).

300 **References**

- 301 [1] R. Ba, C. Chen, J. Yuan, W. Song, and S. Lo. Smokenet: Satellite smoke scene detection using  
302 convolutional neural network with spatial and channel-wise attention. *Remote Sensing*, 11(14):  
303 1702, 2019.
- 304 [2] M. Bah, M. Gunshor, and T. Schmit. Generation of goes-16 true color imagery without a green  
305 band. *Earth and Space Science*, 5(9):549–558, 2018.
- 306 [3] J. Buckman, A. Roy, C. Raffel, and I. Goodfellow. Thermometer encoding: One hot way to  
307 resist adversarial examples. In *International conference on learning representations*, 2018.
- 308 [4] J. Deng, K. Li, M. Do, H. Su, and L. Fei-Fei. Construction and Analysis of a Large Scale Image  
309 Ontology. Vision Sciences Society, 2009.
- 310 [5] R. E. Ferreira, Y. J. Lee, and J. R. Dórea. Using pseudo-labeling to improve performance of  
311 deep neural networks for animal identification. *Scientific Reports*, 13(1):13875, 2023.
- 312 [6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition, 2015.
- 313 [7] P. Iakubovskii. Segmentation models pytorch. [https://github.com/qubvel/segmentation\\_models.pytorch](https://github.com/qubvel/segmentation_models.pytorch), 2019.
- 315 [8] A. Larsen, I. Hanigan, B. J. Reich, Y. Qin, M. Cope, G. Morgan, and A. G. Rappold. A deep  
316 learning approach to identify smoke plumes in satellite imagery in near-real time for health risk  
317 communication. *Journal of exposure science & environmental epidemiology*, 31(1):170–176,  
318 2021.
- 319 [9] D.-H. Lee. Pseudo-label : The simple and efficient semi-supervised learning method for deep  
320 neural networks. *ICML 2013 Workshop : Challenges in Representation Learning (WREPL)*, 07  
321 2013.
- 322 [10] Y. Lee, C. D. Kummerow, and I. Ebert-Uphoff. Applying machine learning methods to detect  
323 convection using geostationary operational environmental satellite-16 (goes-16) advanced  
324 baseline imager (abi) data. *Atmospheric Measurement Techniques*, 14(4):2699–2716, 2021.
- 325 [11] D. McNamara, G. Stephens, M. Ruminski, and T. Kasheta. The hazard mapping system (hms) -  
326 noaa’s multi-sensor fire and smoke detection program using environmental satellites. *Conference  
327 on Satellite Meteorology and Oceanography*, 01 2004.
- 328 [12] NOAA. Hazard mapping system fire and smoke product. URL <https://www.ospo.noaa.gov/Products/land/hms.html#about>.
- 330 [13] T. C. Phan and T. T. Nguyen. Remote sensing meets deep learning: exploiting spatio-temporal-  
331 spectral satellite images for early wildfire detection. 2019.
- 332 [14] T. Randriambelo, S. Baldy, M. Bessafi, M. Petit, and M. Despinoy. An improved detection  
333 and characterization of active fires and smoke plumes in south-eastern africa and madagascar.  
334 *International Journal of Remote Sensing*, 19(14):2623–2638, 1998.
- 335 [15] M. Raspaud, D. Hoese, A. Dybbroe, P. Lahtinen, A. Devasthale, M. Itkin, U. Hamann, L. Ø.  
336 Rasmussen, E. S. Nielsen, T. Leppelt, et al. Pytroll: An open-source, community-driven python  
337 framework to process earth observation satellite data. *Bulletin of the American Meteorological  
338 Society*, 99(7):1329–1336, 2018.

- 339 [16] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: an  
340 astounding baseline for recognition, 2014.
- 341 [17] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image  
342 segmentation, 2015.
- 343 [18] A. Royer, P. Vincent, and F. Bonn. Evaluation and correction of viewing angle effects on  
344 satellite measurements of bidirectional reflectance. *Photogrammetric engineering and remote*  
345 *sensing*, 51(12):1899–1914, 1985.
- 346 [19] W. Schroeder, M. Ruminski, I. Csizar, L. Giglio, E. Prins, C. Schmidt, and J. Morisette.  
347 Validation analyses of an operational fire monitoring product: The hazard mapping system.  
348 *International Journal of Remote Sensing*, 29(20):6059–6066, 2008.
- 349 [20] C. Sun, A. Shrivastava, S. Singh, and A. Gupta. Revisiting unreasonable effectiveness of data in  
350 deep learning era, 2017.
- 351 [21] Z. Wang, P. Yang, H. Liang, C. Zheng, J. Yin, Y. Tian, and W. Cui. Semantic segmentation and  
352 analysis on sensitive parameters of forest fire smoke using smoke-unet and landsat-8 imagery.  
353 *Remote Sensing*, 14(1):45, 2022.
- 354 [22] J. Wen and M. Burke. Wildfire smoke plume segmentation using geostationary satellite imagery.  
355 *ArXiv*, abs/2109.01637, 2021. URL [https://api.semanticscholar.org/CorpusID:  
356 237416777](https://api.semanticscholar.org/CorpusID:237416777).
- 357 [23] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural  
358 networks?, 2014.
- 359 [24] T. X.-P. Zhao, S. Ackerman, and W. Guo. Dust and smoke detection for multi-channel imagers.  
360 *Remote Sensing*, 2(10):2347–2368, 2010. ISSN 2072-4292. doi: 10.3390/rs2102347. URL  
361 <https://www.mdpi.com/2072-4292/2/10/2347>.