# SmokeViz: A Pseudo-Labeled Smoke Plume Dataset For Deep Learning Applications

from satellite images

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

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

## 1 Introduction

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

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

geostationary satelites provider continous (if you define that as having a temporal resolution of 30-min or better.)

Satellite imagery, equipped with advanced sensors, such as the Advanced Baseline Imager (ABI) on the Geostationary Operational Environmental Satellites (GOES), have revolutionized environmental monitoring. These tools enable the detailed observation of smoke plumes, their particulate density, and the extent of smoke spread. These satellite-based systems offer the capabilities to provide critical insights into the concentration and movement of smoke particulates, facilitating accurate and timely assessments of air quality.

again, here it would be useful to put a time sampling number in the discussion. For evolution, (moovement), I think 30-min updates or faster are needed

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

The integration of satellite imagery in wildfire smoke monitoring is not only instrumental in providing real-time data but also plays a significant role in public health planning and response. By mapping the spread and density of smoke, health authorities can issue timely warnings, implement evacuation protocols, and deploy resources effectively to mitigate health risks. Furthermore, long-term data gathered from satellite observations can aid in understanding the broader impacts of wildfire smoke on public health, influencing policy decisions and preventive measures.

Currently, multi-channel thresholding is a popular method to distinguish smoke pixels from pixels containing dust, clouds or other phenomenon with similar signatures. The method uses historical, labeled data to extract optimal radiance values for each channel that corresponds with the labeled class. These methods are tuned to particular biogeographies and often have issues with generalization to new locations with varying fuel types [11].

This calls for a reference or two. I would like to encourage you to refernece papers by Pavolonis et al from NESDIS — if you are talking non-AI methods. It wasn't clear from the context

In contrast to the numerical thresholding approach, human visual inspection of satellite imagery is another commonly used method for smoke identification. Trained analyst will inspect imagery and label the smoke by hand. This method is not as scalable as an automated approach and is limited by the availability of analysts and their time.

To address these challenges we can look towards innovative approaches and technological advancements in computer vision. Machine learning methods have shown potential in improving the accuracy and efficiency of satellite-based wildfire smoke detection and monitoring. For instance, SmokeNet, uses a convolutional neural network (CNN) based framework to determine if a scene of MODIS imagery contains smoke [1]. Another study also used a CNN to identify smoke on a pixel-wise basis using imagery from Himiwari-8 [7]. Additionally, Wen et al. developed a CNN architecture that takes GOES-EAST imagery as input and National Oceanic and Atmospheric Administration (NOAA) generated annotations for the target labels during training [19].

Tangent: I don't konw this journal at all, but you are assuming all readers know what a CNN, U-Net, and deep learning networks mean. Perhaps that is fine for this audience.

The success of deep learning methods, such as CNNs, relies heavily on the availability of a large, representative dataset [17]. Existing methods use relatively small amounts of data, from 57 [18] to 6825 [19]. In contrast, benchmark datasets for image classification contain tens of thousands (CIFAR-10 and MNIST) to millions (CIFAR-100 and ImageNet) of data samples. Keeping in mind the correlation between both the quality and quantity of data with model performance, we introduce the largest known smoke dataset, SmokeViz, containing over 120,000 samples.

When you use "amount of data" and then state "from 57 to 6825", I think you need to be more clear on what constitudets 1 piece of data. Are thesee independent samples, or will they have some overlaps (coindicent regions) or be close in time (and thus a strong temporal correlation)? This correlation question is critical, I think, and is seldom discussed (but I encourage you strongly to do so and talk about its importance)

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 |
| [19] | CNN | 6825 | GOES-EAST | 5 | pixel |
| [7] | CNN | 975 | Himiwari-8 | 7 | pixel |
| [18] | U-Net | 47 | Landsat-8 | 13 | pixel |
| SmokeViz | U-Net | 120,000 | GOES-EAST/WEST | 3 | pixel |

Connecting to my previous Q: in an image which has thousands of pixes, do the pixel-level algos treat each pixel as a sample? I doubt it, but you need to be more clear.

An approach to increase the number of labeled samples in a dataset, semi-supervised learning leverages a labeled dataset to generate new labels for an often larger, but unlabeled, dataset. Pseudo-labeling, a form of semi-supervised learning, uses labeled data to train an initial model, then runs that model on unlabeled data to predict pseudo-labels, and finally trains a new model using the pseudo-labels [8]. We introduce a variation of pseudo-labeling not to increase the size, but to increase the quality of our dataset by using the pseudo-labels to choose the best satellite image out of a given time-window to represent each smoke plume annotation.

Which is why I stress you need to talk about the importance of having independent data for training above, as you can reinfoce why this is a good approach here

## 2 Methods

### Dataset

The initial data source, discussed in further detail in the next section, is uniquely characterized by each annotation having corresponding imagery ranging between 1-60 frames, where each frame captures 5 minutes of exposure. Additionally, we have two satellite sensors, GOES-EAST and GOES-WEST, doubling the number of frames for a single annotation. We apply pseudo-labeling to

Nitpicky comment: the sensor is the ABI on these two platforms

That means you are foocusing only on a particular longitude band where both sensors have overlapping views? If so, you should state that more explicitly

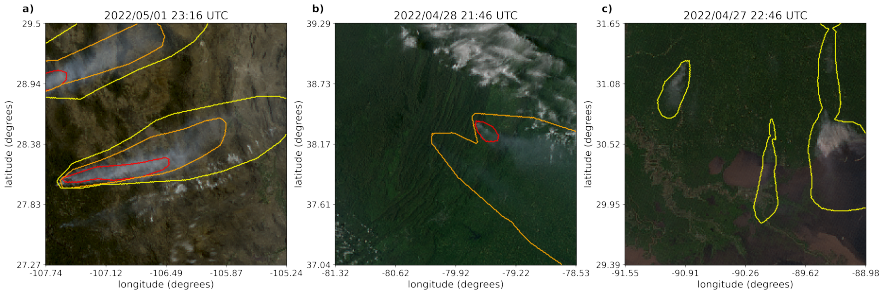



Figure 1: Satellite imagery captured by GOES-EAST within a few days of each other. The yellow, orange and red contours indicate the extent of Light, Medium and Heavy smoke. a) shows a canonical example of a smoke plume. b) and c) show variations in the density labels. b) we show Medium and Heavy densities of smoke that, upon visual inspection, could be interpreted as less dense than portions of the Light density smoke labeled in c).

as determined by what methocd? what do you mean by "variations in the density labels"?

This yellow text could be due to the density of the smoke actually being different, or a difference in the viewing angle of the satellite, the angle of the sun into the scene, the brightness of the underlying surface, etc. As long as you know this is a very qualitatitive statement that really has no physical meaning from a smoke property perspective

develop a dataset that has a one-to-one annotation-to-image ratio, where we choose the best satellite image that represents where the smoke is located in the analyst annotation.

super vague. What is "best". A single analyst did the labeling, or was is a consensus? Or was the training data labeled by another automated algorithnm (e.g., from NESDIS or NASA)?

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

## Smoke Labels

This is redundant, as NESDIS is within NOAA.

The National Environmental Satellite, Data and Information Service (NESDIS) and NOAA manage environmental satellite programs such as the Hazard Mapping System (HMS) [9, 16]. The HMS program is an operational system that uses an aggregation of satellite data to generate active fire and smoke data that is used in applications such as air quality assessments and serves as verification and validation for NOAA's smoke forecasting model, HYSPLiT, [14]. To train our model, we implement a supervised learning framework that uses the HMS analyst smoke product as truth labels during the model training process.

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

So with this info here, perhaps in my Qs above just parantheticavlly say (described later) or similar

## Thermometer Encoding Smoke Densities

One of the challenges introduced with using human generated qualitative smoke densities was that, as seen in figure 1b and 1c, there are variations in what is labeled as heavy or light density smoke. More generally, reproducing qualitative metrics with quantitative algorithms is a challenging problem, but we apply mathematical approaches that mitigate some of the underlying complications of our specific problem. Despite the fact that the smoke densities introduce qualitative complexities, we decided that the density approximations were important to use in our dataset because of the differences in signatures the densities produce. Within the satellite imagery, the appearance of a light density smoke plume will look significantly different than a heavy density smoke plume as seen in figure 1. Additionally, a light density smoke plume is expected to be more challenging to detect since it is easier for it to be misclassified as not smoke. During the training process, the separate density categories allows us to deferentially weight the penalization given to the model for incorrect classifications

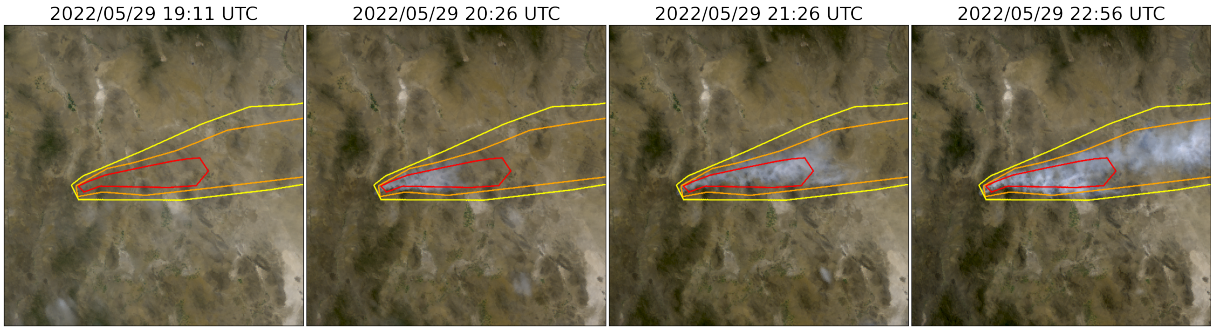



Figure 2: True Color GOES imagery from May 2022, Southeast New Mexico (31°N, 100°W) during the start of the Foster Fire. The HMS annotations for the smoke outlines shown here spanned from 19:10–23:00 UTC but visually match the smoke location for the last part of the timewindow.

based on category. For example, the model can be given a small penalization for misclassifying light smoke as not smoke while given a higher penalization for misclassifying heavy smoke as not smoke.

In addition to the densities being ordered and categorical, the differences between the density categories are not evenly distributed by a metric, such as particulate matter per square meter. The intervals between densities being unknown along with the hierarchical nature of the density labels makes the labels ordinal instead of just categorical. This data property allows us to use thermometer encoding, which leverages the idea that heavy density smoke includes both medium and light density smoke, that heavy density smoke is closer to medium than it is to light and automatically weights the loss functions and incorporates the ranked ordering of the densities. As seen in Table 2, one-hot encoding, commonly used for categorical data, doesn't take ordinal properties of the data 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] |

## Time Windows For Smoke Annotations

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

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 |

## Satellite Imagery

The Geostationary Operational Environmental Satellites (GOES) are operated by the NOAA and NESDIS support meteorology research and forecasting for the United States. We use the latest operational satellites, GOES-16 (EAST), 17 and 18 (WEST) that carry the ABI, that measure 16 bands between the visible and infrared wavelengths. In improvement to the GOES predecessors, imagery is collected every 5 minutes for the contiguous United States and every 10 minutes for the full disk. We use bands 1-3 (Table 3) as input to Satpy's composite algorithm to develop a true color image representation, similar to what is used as input by HMS analysts [12] and [2].

redundant

An

We used a physics-informed approach in selecting the initial dataset for training our model. Rather than use the cumulative data from GOES-WEST and GOES-EAST images, we select one or the other based on the solar zenith angle. For smoke identification, this approach can achieve a much higher signal-to-noise than imaging the earth's surface from an arbitrary angle. The elastic scattering of light is the primary mechanism to account for - while the atmosphere is composed of molecules with size <1nm, smoke particles can vary from 100 nm - 10 $\mu$m in diameter, $d$. The GOES ABI covers spectral bands from 0.47 $\mu$m - 13.3 $\mu$m, so atmospheric and smoke particle sizes occupy two very different regimes with respect to the imaging wavelength $\lambda$, as shown in figure 3. In the extreme limit of $\lambda \gg d$, the physics of scattering of light off a small sphere is captured by Rayleigh scattering. This process has two critical consequences: (1) the scattering cross section of light is strongly wavelength dependent (scaling with $\lambda^{-4}$), meaning that photons with wavelength closer to the ultraviolet are scattered more strongly than infrared photons. (2) the scattering cross section scales with an angular dependent cross section of $(1 + \cos^2\theta)$. Scattered photons follow the emission distribution of a radiating dipole, scattering more strongly in the forward and backwards directions $(\theta = 0, \pi)$than orthogonal to the direction of propagation $(\theta = \pi/2, 3\pi/2)$, see figure 4 for Rayleigh scattering schematic.

The significance of these scalings is that the observer, or detector, will receive blue photons in most directions orthogonal to the source. Equivalently, photons traveling colinearly with line of sight to the emission source will mostly have wavelengths in the infrared band. In the converse regime of $d > \lambda$, the elastic scattering of light against matter is modeled through Mie scattering. Unlike Rayleigh scattering, Mie scattering is largely wavelength independent and has a more complicated radiation pattern where the cross section has a maximal amplitude in the forward direction. An observer downstream of this scatterer will collect more photons than one positioned directly behind it. In the context of smoke identification, a sunrise or sunset will lead to a higher Mie scattered signal in GOES-WEST and GOES-EAST respectively, as shown with a smoke plume producing a stronger signal in GOES-EAST imagery near sunset in figure 2. I think you mean Fig 4

Yes and no. If thee size parameter (ratio of Pi * d / lambda) is greater than 10, then in the geopmetric optics regime

No, Mie scattering is wavelength dependent. But it depends on the square, not on the 4th power. I would just say is has a different wavelength dependence than Rayleigh scattering

Smoke identification therefore amounts to extracting a signal of $d > \lambda$ photons from the $\lambda \gg d$ background. Positioning a detector along line of sight to the scatterer will result in a higher signal from smoke particles (figure 4). Filtering the imaged wavelength can enhance this signal; photons collected in the blue spectrum will have a naturally lower background along the line of sight to the illumination source do their high level of Rayleigh scattering as. Therefore, as demonstrated in figure 6, this configuration results in the highest signal to noise imaging for smoke particles.

If the goal is to create a dataset that can be used to develop more robust smoke-determination algorithms, this approach is going to result in a dataset that suffers from a bad selection bias. For example, we will want to be able to identify smoke in GOES-W images oveer CA in thee afternoons, but that won't exist in this datasset at all (as I understand it)

Based on these criteria, the optimal strategy is to pull data from GOES-WEST right after sunrise and from GOES-EAST right before sunset. Another consideration to account for was that when the sun is in optimal alignment with the satellite for detecting smoke also coincides with the maximal amount of atmosphere the light travels through. This is shown in figure 7, where the noise introduced by higher amounts of atmospheric interactions can obfuscate the signal from the smoke, despite the smoke signal being at its highest. This phenomenon is even more prominent the further the smoke is longitudinally from the satellite since the light must travel through more atmosphere between

I think you are misinterpreting Fig 7, or at least this explanation isn't correct. The visual changee in the signal is indeed assocaited with the relative difference in the solar angle and the viewing angle, and hence thes scattering phase function, but less assocaited with the "amoiunt of atmosphere" comment. If you do the math, the change in the atmospheric path langth traveled by the photons is (much) smaller than 2, even though the signal looks more than a factor of 2 stronger in the example shown in the figure

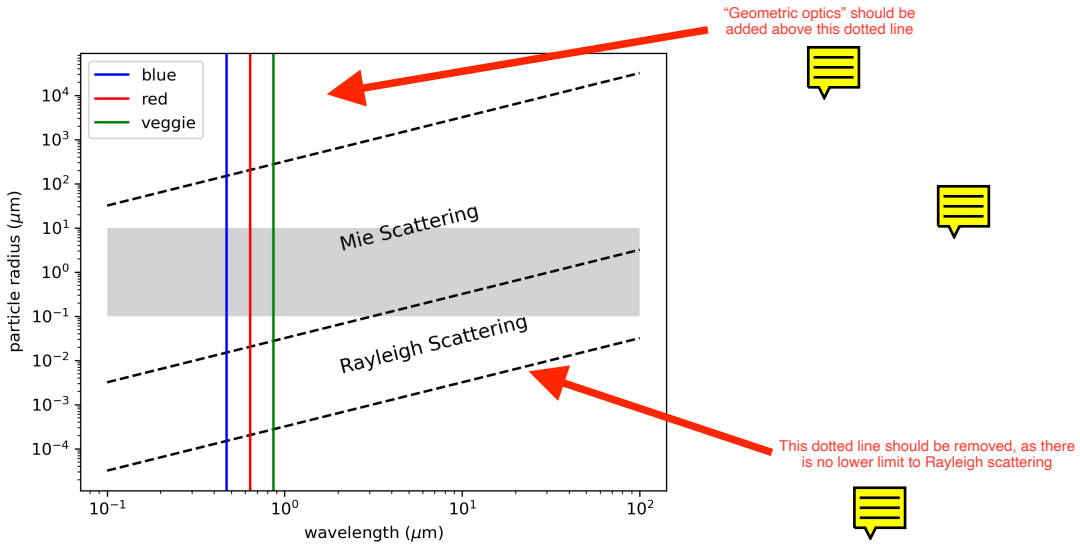



Figure 3: Relationship between the size of a particle, the wavelength of light interacting with the particle and the type of scattering behavior induced by that interaction. The dotted lines represent rough estimates of the boundaries between the scattering regimes [10]. The gray area represents the range of particle radius relevant to smoke particulate matter.

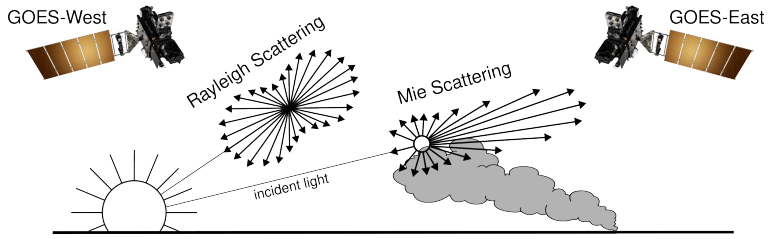



Figure 4: If the particle size is $< \frac{1}{10}$ the wavelength of the interacting light, then the primary scattering will be Rayleigh. Mie scattering is the predominant scattering mechanism when the particle size is larger than wavelength of light.

This caption is misleading. According to the ttext, this is an example at Sunset showing the importance of viewing geometry. This needs to be stated exolicitly here

scattering off the smoke and reaching the detector. Additionally shown in figure 7, and is especially evident for data close to sunrise, when the time window is large, the smoke has often not dispersed to the extent of the analysts' annotation boundaries. We consider the atmospheric interaction noise in our algorithms to develop the dataset by choosing a lag time between sunrise and optimal image timestamp as a function of longitude.

Why is the time window usuaally large near sunrise (as opposed to midday or sunset)? Why is ana analyzsts boundaries so different at sunrise?

The resulting algorithm used atmospheric properties and light scattering physics to make an estimate of which singular satellite image within the analyst time-window would give the best representation of the smoke plume label. That dataset was then used to train a model that would generate pseudo-labels for every image within the time-window and choose the image with the highest alignment between smoke in the image and annotation.

### Machine Learning Model

We implement a deep learning architecture that uses the encoder from the ResNet model [5] and a semantic segmentation classifier from the U-Net model [15]. Transfer learning has shown to reduce the time and resources needed to train a model by leveraging information from pre-trained models [20], [13]. We initialize the values of our model weights using the pre-trained values originally trained on the ImageNet dataset [3], containing 1.2 million images and 1000 categories. Our model was developed using the Segmentation Models PyTorch package [6] that was written as a high level API for implementing models for semantic segmentation problems. We input 256x256x3 snapshots

Is this 256 pixels by 256 pixels by 3 wavelengths? Or has the pixel data already been preprocessed (e.g., using a convolution)?

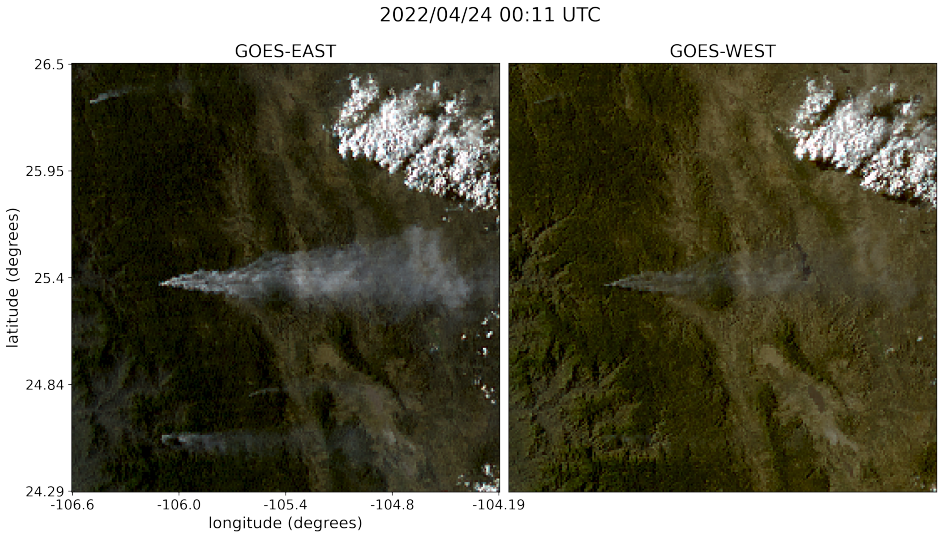
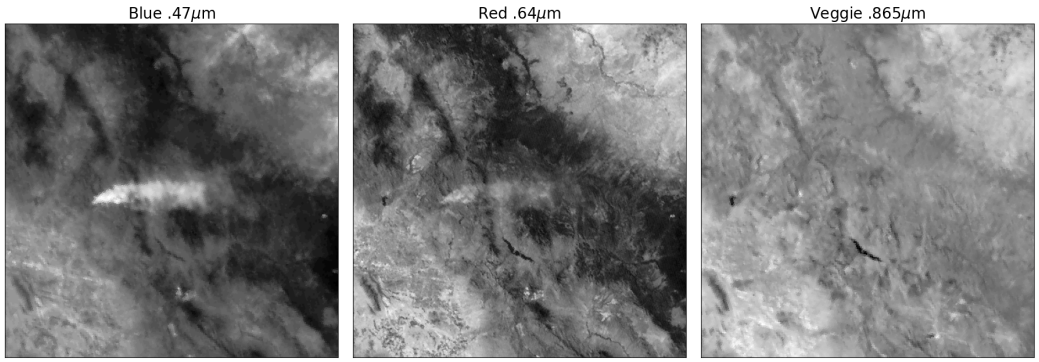



Figure 5: True Color GOES-EAST (left) and GOES-WEST (right) imagery from April $24^{th}$, 2022. The images were taken about 1.5 hours before sunset for this geolocation and time of year (01:43 UTC).

Where is this image (ii.e., what region of the country)? what is the size of the image (e.g., is this a 100. km by 100 km image?

Very nitpicky — I would put the GOES-E satellite image on the right side

Very nitpicky: please don't write a number in a paper that starrts with a decimal point. In those cases, use a leading zero. So 0.47 $\mu$m, not .47 $\mu$m
(For the headers of these images)




Always good to list the actual wavelengths of the channels (or the channel numbers, connecting to the above table, in the caption. Don't make the reader go looking

Figure 6: The three bands of GOES-EAST data are the raw input to generate the True Color image shown in figure 5. There appears to be a higher signal-to-noise ratio for smoke detection as the wavelength, $\lambda$, of light being measured decreases.

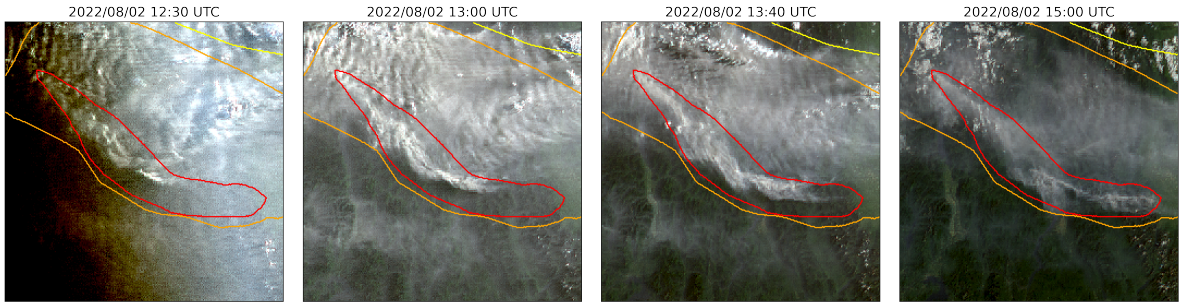



Figure 7: A smoke annotation projected onto GOES-WEST imagery from August 2022 that spans from 11:00 UTC to 15:00 UTC, sunrise on August 2nd, 2022 at coordinates (49°24'N, 115°29'W) was 12:15 UTC. 15 minutes after the calculated sunrise, the imagery contains a noticeably higher levels of noise than the subsequent imagery that's light travels through (thus interacts with) less atmosphere as the sun rises and the angle between GOES-WEST and the sun decreases.

You've sort of answer4ed my questio here, but you could be more clear earlier in the sentence

of True Color GOES imagery that contains smoke and output a 256x256x3 classification map that predicts if a pixel contains smoke and if so, what the density of that smoke is. As mentioned earlier, we apply the thermometer encoding shown in table 2 to encode the smoke densities and apply binary cross entropy as the loss function per density of smoke.

Not clear what you mean here

The original dataset developed using the Mie algorithm contained over 120,000 samples. To train our model, we split the dataset into training (95,000 samples), validation (12,000 samples) and testing (12,000) datasets. Training data contains data from the years 2018, 2019, 2020, 2021 and 2023 while the data from 2022 is split into validation and testing data by taking data from alternating 10 days of the year. Splitting 2022 data by 10 days allowed us to leave more full years of data for the training set and allowed each dataset to show yearlong trends while trying to keep the datasets independent from one another.

We trained a model over 20 epochs and then used that model to develop the dataset by determining which satellite image provided the best Intersection over Union (IoU) value. The IoU metric is given by the ratio of area of overlap to the area of union as defined in equation 1, where A and B are the truth labels and the model's predictions.

$$IoU = \frac{|A \cap B|}{|A| \cup |B|} \quad (1)$$

Amazingly, I hadn 't seen this metric before. I like it

To determine which image best represents the analyst annotation, we gather all the satellite imagery for the given time window and run them through the machine learning model. The output of the model gives a prediction on if there is smoke in the image, and if there is smoke, where the smoke is in that image and what the density of that smoke is. The model output generates pseudo-labels for each density of smoke that are compared to the analyst annotations. To compare the pseudo labels and analyst labels, we calculate the IoU using the total set of pixels for the pseudo-labels at that density of smoke and the entire set of pixels for the analyst labels for a particular smoke density in each image. The image with the highest IoU score is chosen as the image that best represents the analyst smoke annotation. Generally, a confidence threshold value is defined to decide if a pseudo-label should to be included in a dataset [4]. We chose a confidence threshold that would include the sample in the dataset if the maximum IoU value was over 0.1.

What do you do if the IoU > 0.1 for heavy smolke, but IoU < 0.1 for light smokle? In other words, do you assign different weights to the smokle density, and if so, what are those weights?

## Results

To interpret the performance of our trained model, we report the IoU metrics in table 4 that were computed by running the model on the Mie algorithm derived dataset and the pseudo-labeled dataset.

Same question as above (what is this)? And I guess there is a related question: how are clouds treated in this algorithm / analyisis? That is a 2-pronged question: (a) clouds being mis-identified as smoke, and (b) scenes that have both smoke and aerosol intermingled.

Table 4: IoU results per density of smoke and over all densities.

| category | IoU Mie Dataset | IoU Pseudo-Labeled Dataset |
|---|---|---|
| Light | 0.394 | 0.551 |
| Medium | 0.283 | 0.392 |
| Heavy | 0.233 | 0.290 |
| Overall | 0.365 | 0.510 |

I would really like to see another paragraph or two iat the end of section 2 that interprets this table. What is the message here?

And is there any important degree of "cross-talk" here (i.e., mis-labeling the density) that is important to point out? For example, perhaps a lot of the events that have medium density are being labeled as light?

For each density, we calculate the IoU using the total set of pixels that the model predicts as that density of smoke and the entire set of pixels labeled by the analyst as a particular smoke density over all imagery contained in the testing dataset. Additionally, we compute the overall IoU for all densities by first computing the number of pixels that intersect their correct density and divide that by the total number of pixels that make up the union of model predicted and analyst labeled smoke as shown in equation 2.

$$IoU_{overall} = \frac{\sum_{i=light}^{heavy} |A_i \cap B_i|}{\sum_{i=light}^{heavy} |A_i| \cup |B_i|} \tag{2}$$

## 3 Conclusion

In this study, we have refined an existing dataset originally curated by the HMS team, transforming it from a many-to-one imagery-to-annotation format to a, more concise, one-to-one satellite image-to-annotation dataset. The initial HMS dataset primarily gave an approximation of where smoke had been present for a given time window, though it did not confirm the actual existence of smoke in the pixels of the selected images. Due to that nature of the HMS dataset, our Mie derived dataset gave an approximation of when we'd best be able to measure the smoke signal but did not factor in if the smoke was actually present in the selected image. This discrepancy can be detrimental when training a machine learning model, as it may penalize accurate predictions and inadvertently introduce biases towards misclassifying noise as meaningful signal.

To make improvements on the dataset's reliability, we apply a machine learning model trained on the Mie-derived dataset to select the satellite image within the time-frame that best overlaps with the analyst's annotation. A notable illustration of the improvements introduced by the machine learning method is evident in Figure 8. The annotation associated with this example encompasses five hours of imagery and we show the images that is chosen by each of our two methods. While the Mie algorithm tries to optimize for the highest possible signal-to-noise, which is the image closest to sunrise, our machine learning algorithm chooses the image that maximizes the overlap of smoke predicted by the model with the analyst's annotation.

Totaly unfair comparison — see comments around the figure

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

## 4 Acknowledgments and Disclosure of Funding

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

See GSL web page for guidance here. Depends on who the coauthors are

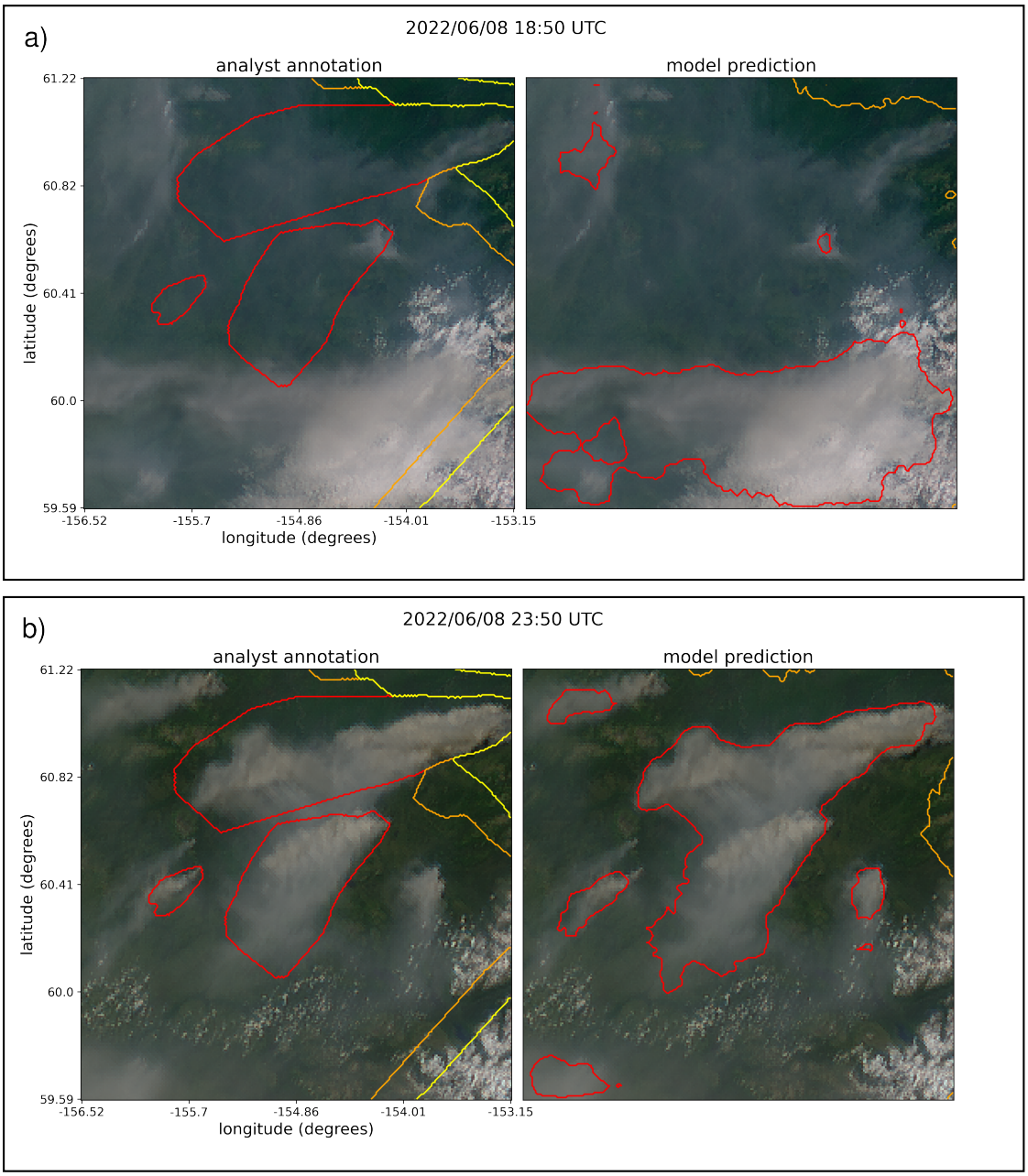



Figure 8: GOES-WEST imagery showing smoke on June 8th, 2022 in Alaska where at the coordinates (61°03'N, 156°07'W), daylight was between 12:43-7:53 UTC. The smoke annotations displayed span from 18:50 to 23:50 UTC. a) shows the imagery that was selected using the Mie algorithm, which optimizes for the image closest to sunrise. b) shows the imagery chosen by the pseudo-label that had the highest IoU score. The IoU scores are similar for the low and medium density smoke, but the high density smoke IoU for a) is .01 while b) is significantly higher at .59.

Please adjust the headers on the right-hand panels from "model prediction" to "Mie algorithm prediction"and "Pseudo-label prediction" or similar.

Multiple things here. First, this figure needs to be part of section 2 and used to help describe the performance between the two approaches. It is not proper to put it in the conclusions. Second, wow — by using different times for the left hand panels (and the left hand panel in (a) clearly shows a cloud in my mind), this is super misleading on what the Mie algprithm can really do. What does the Mie algorithm show at 23:50? If you leave this figure as-is, it really leaves me with the impression (for rright or wrong) that (a) the Mie algorithm is pretty stinky so of course it will be easy to beat (table 4) and (b) you used the Mie algorithm as input to start the pseudo-labeling — so why is it seeming to work? You will want to revisit this and incorporate this into (proabably new) paragraphs within section 2.

# References


[1] R. Ba, C. Chen, J. Yuan, W. Song, and S. Lo. Smokenet: Satellite smoke scene detection using convolutional neural network with spatial and channel-wise attention. *Remote Sensing*, 11(14): 1702, 2019.

[2] M. Bah, M. Gunshor, and T. Schmit. Generation of goes-16 true color imagery without a green band. *Earth and Space Science*, 5(9):549–558, 2018.

[3] J. Deng, K. Li, M. Do, H. Su, and L. Fei-Fei. Construction and Analysis of a Large Scale Image Ontology. Vision Sciences Society, 2009.

[4] R. E. Ferreira, Y. J. Lee, and J. R. Dórea. Using pseudo-labeling to improve performance of deep neural networks for animal identification. *Scientific Reports*, 13(1):13875, 2023.

[5] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition, 2015.

[6] P. Iakubovskii. Segmentation models pytorch. `https://github.com/qubvel/segmentation_models.pytorch`, 2019.

[7] A. Larsen, I. Hanigan, B. J. Reich, Y. Qin, M. Cope, G. Morgan, and A. G. Rappold. A deep learning approach to identify smoke plumes in satellite imagery in near-real time for health risk communication. *Journal of exposure science & environmental epidemiology*, 31(1):170–176, 2021.

[8] D.-H. Lee. Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks. *ICML 2013 Workshop : Challenges in Representation Learning (WREPL)*, 07 2013.

[9] D. McNamara, G. Stephens, M. Ruminski, and T. Kasheta. The hazard mapping system (hms) - noaa's multi-sensor fire and smoke detection program using environmental satellites. *Conference on Satellite Meteorology and Oceanography*, 01 2004.

[10] G. Petty. *A First Course in Atmospheric Radiation*. Sundog Pub., 2006. ISBN 9780972903318.

[11] T. Randriambelo, S. Baldy, M. Bessafi, M. Petit, and M. Despinoy. An improved detection and characterization of active fires and smoke plumes in south-eastern africa and madagascar. *International Journal of Remote Sensing*, 19(14):2623–2638, 1998.

[12] M. Raspaud, D. Hoese, A. Dybbroe, P. Lahtinen, A. Devasthale, M. Itkin, U. Hamann, L. Ø. Rasmussen, E. S. Nielsen, T. Leppelt, et al. Pytroll: An open-source, community-driven python framework to process earth observation satellite data. *Bulletin of the American Meteorological Society*, 99(7):1329–1336, 2018.

[13] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition, 2014.

[14] G. D. Rolph, R. R. Draxler, A. F. Stein, A. Taylor, M. G. Ruminski, S. Kondragunta, J. Zeng, H.-C. Huang, G. Manikin, J. T. McQueen, et al. Description and verification of the noaa smoke forecasting system: the 2007 fire season. *Weather and Forecasting*, 24(2):361–378, 2009.

[15] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[16] W. Schroeder, M. Ruminski, I. Csiszar, L. Giglio, E. Prins, C. Schmidt, and J. Morisette. Validation analyses of an operational fire monitoring product: The hazard mapping system. *International Journal of Remote Sensing*, 29(20):6059–6066, 2008.

[17] C. Sun, A. Shrivastava, S. Singh, and A. Gupta. Revisiting unreasonable effectiveness of data in deep learning era, 2017.

[18] Z. Wang, P. Yang, H. Liang, C. Zheng, J. Yin, Y. Tian, and W. Cui. Semantic segmentation and analysis on sensitive parameters of forest fire smoke using smoke-unet and landsat-8 imagery. *Remote Sensing*, 14(1):45, 2022.

[19] J. Wen and M. Burke. Wildfire smoke plume segmentation using geostationary satellite imagery. *ArXiv*, abs/2109.01637, 2021. URL https://api.semanticscholar.org/CorpusID:237416777.

[20] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks?, 2014.