Location generalizability of image-based air quality models

Supplementary Material

1. National Parks AQI Dataset

The National Parks AQI Dataset is composed of images from the National Park Service's (NPS) air quality web cameras [34] paired with AQI values obtained via the NPS Gaseous Pollutant Monitoring Program (GPMP) [24], which maintains monitoring stations at each national park to monitor and protect air quality. This multi-location AQI dataset is publicly available for download on the PNNL DataHub at the following URL: https://data.pnnl.gov/group/nodes/dataset/33967.

Each monitoring station is equipped with a suite of co-located air quality sensors and a web camera. We note that the camera and air quality sensors are identical across all sites, which is particularly useful for assessing location generalizability, since the fixed hardware minimizes any variation between sites that isn't driven by environmental factors (e.g., color correction differences between cameras, sensor calibration).

1.1. Dataset Assembly

To assemble the dataset, we collect GPMP data from 17 monitoring stations ("sites") between 2016 to 2023. For AQI measurements, the GPMP collects gaseous pollutant data at 5–30 minute intervals: ozone (O_3) in ppm for most sites and SO_2 in ppm for some sites. For imagery, the NPS web cameras capture high-resolution images (1200 x 1600 pixels) of the horizon at 15-minute intervals.

We convert the 8-hour running O_3 concentrations to AQI values following the approach specified by the Environmental Protection Agency (EPA) [1]. To match the 15-minute frequency of the webcam imagery, the AQI values are either down-sampled or up-sampled depending on the sampling frequency for a given park.

Time displacements between AQI measurements and images is never greater than 15 minutes; over such short time periods we expect AQI values to vary smoothly. Several parks have both O_3-based AQI and SO_2-based AQI values; for consistency, we use O_3-based AQI values where possible. HAVO is the only park that only has SO_2-based AQI. While AQI is designed to be a common scale across multiple pollutants, the optical effects of SO_2 pollution will be slightly different from O_3 pollution.

The final National Parks AQI Dataset includes data from 17 sites in 15 national parks: Acadia, Big Bend, Denali, Dinosaur National Monument, Grand Canyon, Great Smoky Mountains (3 sites), Grand Tetons, Hawai'i Volcanoes, Joshua Tree, Mammoth Cave, Mount Rainier, Sequoia and Kings, Shenandoah, Theodore Roosevelt, and Yosemite. Example images from each site are shown in Figures 5 and 6.

The final National Parks AQI dataset includes 146,882 images and AQI values. We include both AQI values for regression tasks and the EPA-established air quality category for classification: "good" (0–50), "moderate" (50–100), "sensitive" (100-150), "unhealthy" (150-200), "very unhealthy" (200-300), and "hazardous" (above 300). Table 4 shows the image count broken down by site and EPA category.

1.2. Dataset Characterization

In this work, AQI is calculated from O_3 where possible, except for HAVO, which only has SO_2 data. The original O_3 and SO_2 sensor measurements are provided in the released dataset as 8-hour running averages in ppb. Most sites (15/17) have additional meteorological measurements, which include temperature and relative humidity. Table 5 shows the sensor measurements available for each site.

1.2.1 Diurnal distribution

Figure 7 shows the temporal distribution of samples broken down by AQI category, with time of day measured on a 24 hour clock. The distributions of the moderate, sensitive, and unhealthy categories are fairly similar, with the most samples taken mid-day (11am to 1pm). The good category distribution is slightly different, with a more uniform time distribution. As described in Sec. 3, for days when the AQI never goes above 50 ("good"; 0–50), only a single image is downloaded, at a randomly sampled daylight hour.

1.2.2 Site-specific AQI distributions

Figures 8 and 9 show the AQI distribution for each site. The sites have very different distributions in AQI, both in terms of overall range and in median values. HAVO is the the only park with imagery at AQI values over 200 ("very unhealthy" or "hazardous"), driven by SO_2 in volcanic emissions; volcanic plumes are clearly visible in the webcam imagery in Fig. 6. HAVO is the only site without ozone-based AQI values, in part because SO_2 is always the dominant air pollutant.

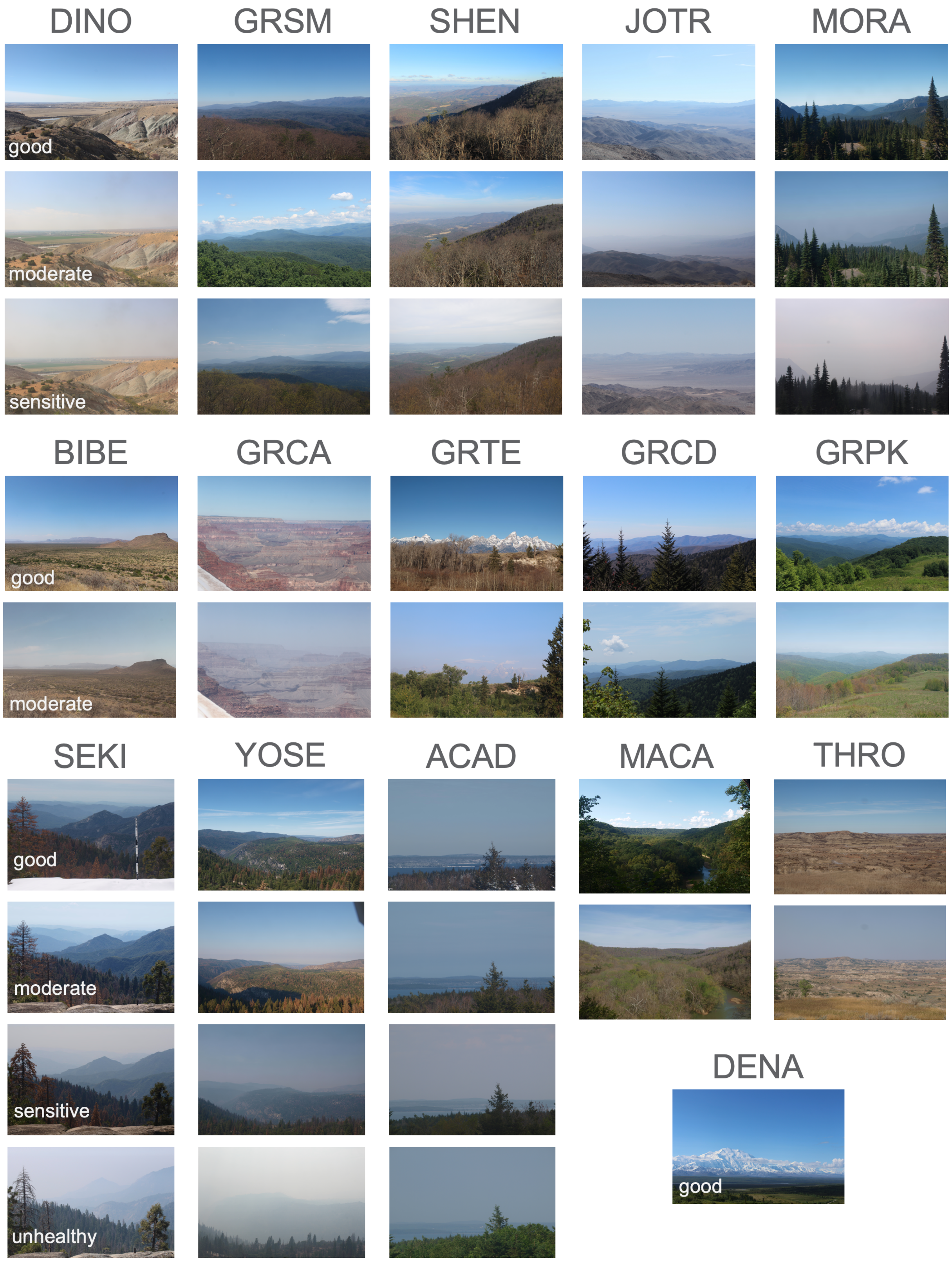

Figure 5. Example imagery from each of the sites included in the National Parks AQI dataset.

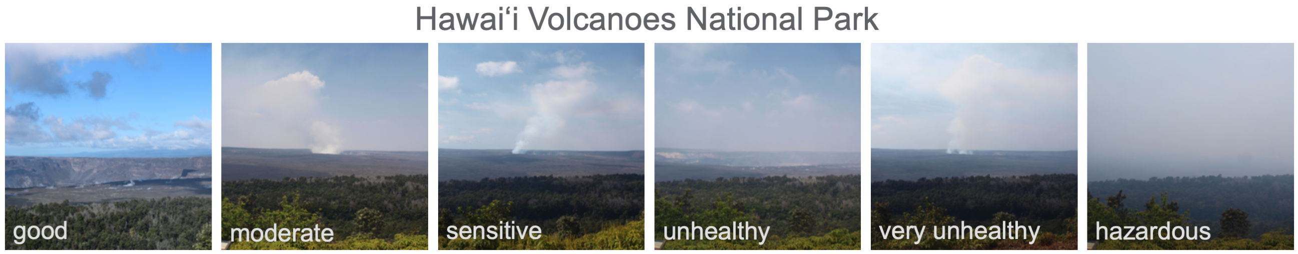

Figure 6. Example imagery from Hawai‘i Volcanoes National Park. This is the only park with imagery categorized as “very unhealthy” and “hazardous”, driven by the SO_2 emission in volcanic plumes.

Park Name	Site	Total	Good	Moderate	Sensitive	Unhealthy	V. Unhealthy	Hazard
Acadia	ACAD	3,898	2,293	1,483	119	3		
Big Bend	BIBE	6,498	2,037	4,461				
Denali	DENA	452	452					
Dinosaur Monument	DINO	6,485	1,305	5,093	87			
Grand Canyon	GRCA	13,176	2,094	11,082				
Grand Tetons	GRTE	5,209	2,335	2,874				
Great Smokey	GRSM	8,285	2,433	5,765	87			
Mountains	GRCD	3,473	1,115	2,358				
	GRPK	4,371	1,542	2,829				
Joshua Tree	JOTR	13,318	1,428	11,431	459			
Mammoth Cave	MACA	5,645	2,367	3,278				
Mt. Rainier	MORA	3,104	1,558	1,446	100			
Sequoia & Kings	SEKI	28,640	1,504	19,554	7,218	364		
Shenandoah	SHEN	5,647	2,200	3,429	18			
Theodore Roosevelt	THRO	2,938	1,525	1,413				
Yosemite	YOSE	19,465	1,843	16,110	1,374	138		
Hawai‘i Volcanoes	HAVO	16,218	1,0081	2,417	1,881	689	685	465
		Total:	38,112	95,023	11,343	1,194	685	465

Table 4. National Parks AQI Dataset imagery statistics.

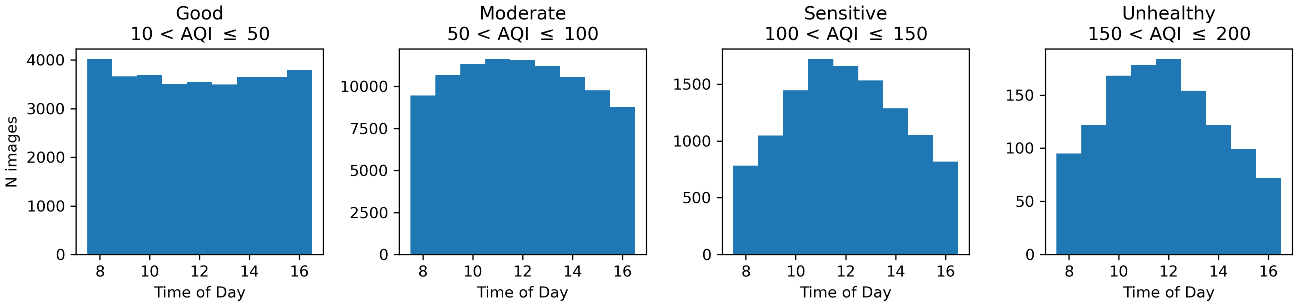

Figure 7. Time distribution of samples for each EPA air quality category.

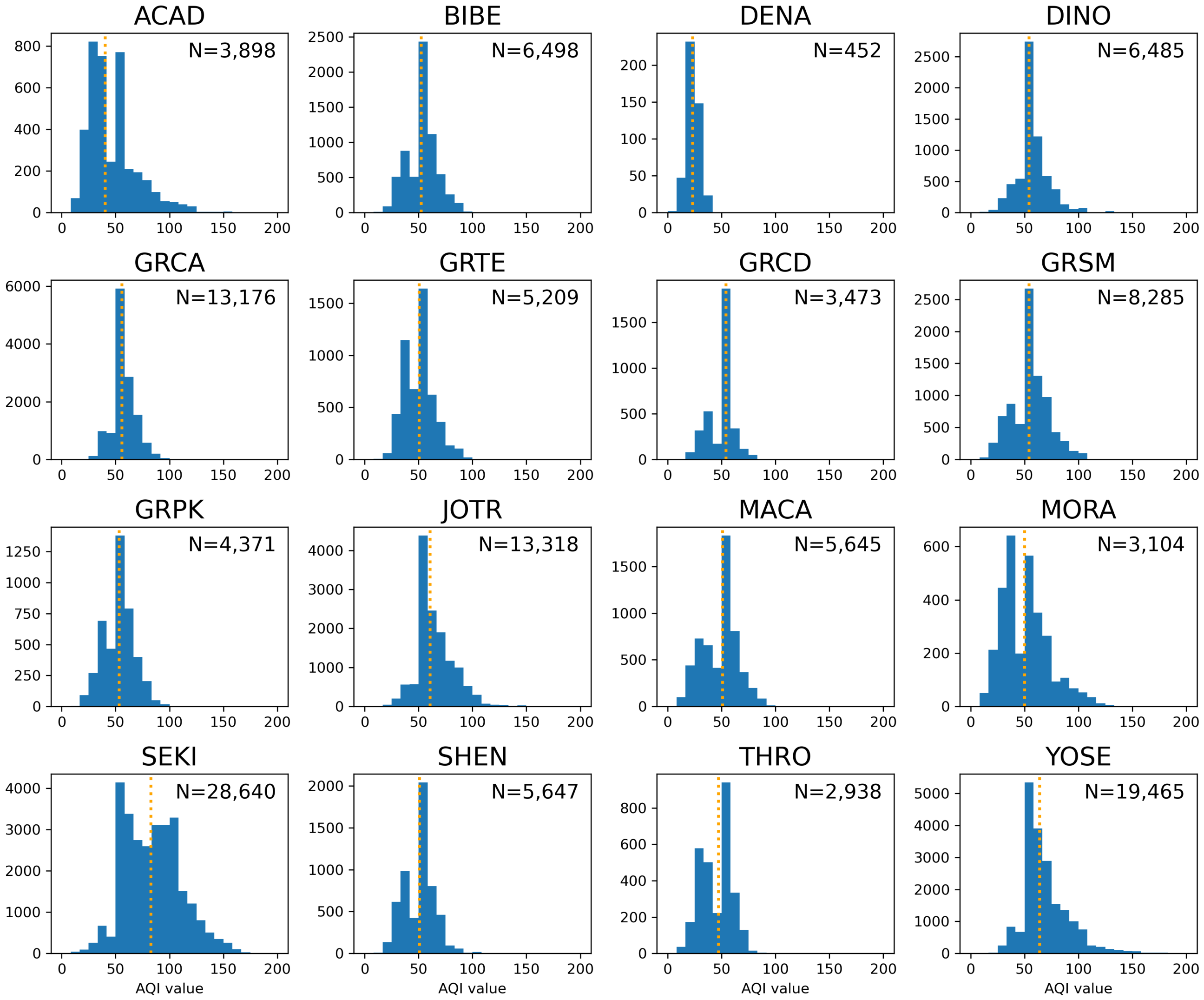

Figure 8. AQI distribution for each site in the National Parks AQI Dataset. For each site, the yellow dashed line shows the median AQI value.

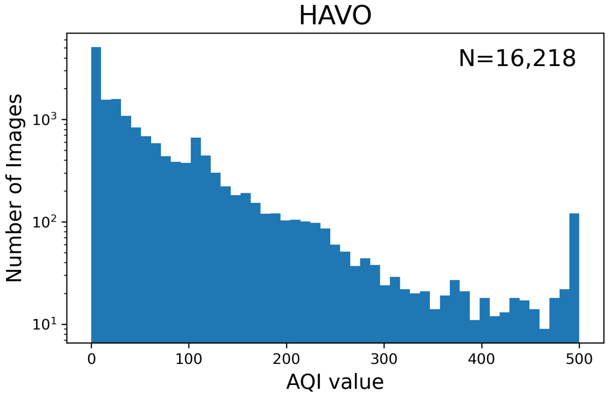

Figure 9. AQI distribution for Hawai‘i Volcanoes National Park.

Site	O_3	SO_2	AQI	Meteorological Data
ACAD	x	x	x	x
BIBE	x		x	x
DENA	x		x	x
DINO	x		x	*
GRCA	x		x	x
GRTE	x		x	x
GRSM	x		x	x
GRCD	x		x	x
GRPK	x		x	
JOTR	x		x	x
MACA	x	x	x	x
MORA	x		x	
SEKI	x		x	x
SHEN	x		x	x
THRO	x	x	x	x
YOSE	x		x	x
HAVO		x	x	x

Table 5. Available data by site in the National Parks AQI dataset. O_3 and SO_2 measurements are 8-hour running averages in ppb. Meteorological data includes temperature and humidity measurements. The asterisk means that DINO only has temperature measurements, and no relative humidity measurements.