DEEP LEARNING DETECTION AND QUANTIFICATION OF VOLCANIC THERMAL SIGNALS IN INFRARED SATELLITE DATA

By

Pablo Saunders-Shultz

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of

Master of Science
in
Geology

University of Alaska Fairbanks
May 2024

APPROVED:

Dr. Taryn Lopez, Committee Chair
Dr. Hannah Dietterich, Committee Member
Dr. Társilo Girona, Committee Member
Dr. Ronni Grapenthin, Committee Member
Dr. Bernard Coakley, Chair
Department of Geosciences
Karsten Hueffer, Dean
College of Natural Science and Mathematics
Richard Collins, Director
Graduate School

Abstract

Volcanic eruptions pose hazards to human lives and livelihoods (Loughlin et al., 2015). To mitigate these hazards, volcano monitoring groups aim to detect signs of unrest and eruption as early as possible. Prior to eruption volcanoes may show various signals of unrest, including: increased surface temperatures, surface deformation, increased seismicity, increased degassing, and more. Here we focus on one approach to monitor volcanic unrest: detecting high-temperature localized volcanic heat emissions, otherwise known as hotspots. The presence of hotspots can indicate subsurface and surface volcanic processes that precede, or coincide with, eruptions. Space-borne infrared sensors can identify hotspots in near-real-time; however, automatic hotspot detection systems are needed to efficiently analyze the large quantities of data produced. While hotspots have been automatically detected for over 20 years with simple thresholding algorithms, new computer vision technologies, such as convolutional neural networks (CNNs), enable improved detection capabilities. Here we introduce HotLINK: the Hotspot Learning and Identification Network, a CNN-based model to detect volcanic hotspots in VIIRS (Visible Infrared Imaging Radiometer Suite) imagery. We find that HotLINK achieves an accuracy of 96% when evaluated on a validation dataset of ~1,700 unseen images from Mount Veniaminof and Mount Cleveland volcanoes, Alaska, and 95% when evaluated on a test dataset of ~3,000 images from six additional Alaska volcanoes (Augustine Volcano, Bogoslof Island, Okmok Caldera, Pavlof Volcano, Redoubt Volcano, Shishaldin Volcano). Additional testing on ~700 labeled MODIS images demonstrates that our model is applicable to this sensor's data as well, achieving an accuracy of 98%. We apply HotLINK to 10 years of VIIRS data and 22 years of MODIS data for the eight aforementioned Alaska volcanoes. From these time series we find that HotLINK accurately characterizes background and eruptive periods, similar to a threshold-based method, MIROVA, but also detects more subtle warming signals, potentially related to volcanic unrest. In particular, analysis of the Mount Veniaminof record demonstrates that HotLINK is able to detect subtle hotspot signals that are coincident with elevated seismicity, potentially indicative of surface heating due to shallow magma intrusion and/or degassing. We identify three advantages to our model over its predecessors: (1) the ability to detect more subtle volcanic hotspots and produce fewer false positives, especially in daytime imagery; (2) the incorporation of probabilistic predictions for each detection that provide a measure of detection confidence;

and (3) its transferability to multiple sensors and multiple volcanoes without the need for threshold tuning, suggesting the potential for global application. HotLINK is able to identify eruptions and potentially precursory warming signals in infrared satellite data, making it a valuable tool for monitoring volcanoes and tracking their heat released over time.

Plain Word Summary

Volcanoes are dangerous forces of nature, producing lava, explosions, and other hazards. However, prior to erupting, volcanoes may produce various warning signals such as small earthquakes, slight deformation of the surface, increased gas emissions, high surface temperatures, and more, which allow for some degree of eruption forecasting. Here we focus on one approach to volcano monitoring, detecting unusually high surface temperatures, or hotspots. Monitoring the presence of volcanic hotspots can help us determine if a volcano is erupting or might erupt soon. Hotspots can be detected by satellite sensors which measure infrared radiation. Traditionally, volcanologists or simple computer programs would identify the hotspots in infrared images. Now, advanced computer algorithms based on artificial intelligence can accurately identify complex features in images. We applied these algorithms to improve the way we detect volcanic hotspots. Our approach detects more subtle heat signals than other algorithms, which is useful for detecting different types of volcanic activity and may contribute to better forecasting of eruptions. By creating an automated method, we can also analyze more data than would be possible manually. We use our new automated system, called HotLINK – the Hotspot Learning and Identification Network, to detect hotspots at eight volcanoes in Alaska for the years 2000-2022. The data produced by HotLINK records multiple eruptions, and may be useful to detect future eruptions if implemented on real time data.

Dedication

For my cat, Orla. Who has never worried about volcanoes even once.

Acknowledgements

Thanks to my advisor Taryn Lopez and other members of my committee Hannah Dietterich, Társilo Girona, and Ronni Grapenthin. Another thanks again to Taryn, Hannah, and Társilo who are coauthors on chapter 2 of this thesis.

Thanks to Alaska Volcano Observatory, the Geophysical Institute, the Department of Geosciences, the College of Natural Sciences and Mathematics, and the graduate school at the University of Alaska Fairbanks.

Thanks to the National Science Foundation for funding this project through a Prediction of and Resilience against Extreme Events (PREEVENTS, award number 1855126) award.

Thanks to the Department of Geosciences for additional funding and Ronni, who let me TA for him one semester.

Thanks to the Commission on the Chemistry of Volcanic Gases and Patrick Allard, former president of the International Association of Volcanology and Chemistry of the Earth's Interior, for providing travel funding to attend a field workshop in Peru.

Thanks to the Geophysical Institute Graduate Student Association and the Associated Students of the University of Alaska Fairbanks, especially Cole Funke, for providing additional travel funds.

This work was made possible by readers like you, thank you.

Table of Contents

Table of Figures

Page

Table of Tables

Page

Chapter 1: Introduction

Eons before there was any life on Earth, volcanoes existed, bringing heat to the surface in the form of molten rock from deep within the planet's interior (Rogers, 1996). Three billion years later the planet is entering the Anthropocene – so called because humanity has become the dominant force changing the climate and life on our planet (Crutzen, 2006). However, despite our capability to modify conditions on our planet, volcanoes remain a force of nature far outside of human command. Volcanic eruptions are responsible for hundreds of deaths annually on average over the past 400 years, and have large impacts on society, human and environmental health, and economies (Baxter, 2005; Brown et al., 2017). In the Holocene, large eruptions have been responsible for rapid climate forcing, famines, and destruction of entire cities (Robock, 2000; Cashman and Giordano, 2008). Incredibly rare volcanic events such as super-eruptions and formation of large igneous provinces have played a role in mass extinction events (Ernst, 2014; Racki, 2020). Although we have no control over the timing and magnitude of volcanic eruptions, it is possible to forecast and mitigate some volcanic hazards (Cassidy et al., 2023). Volcano monitoring agencies aim to understand the underlying processes that drive volcanism and detect signs of eruption and unrest with the observations available. Their goal is to monitor volcanic activity, forecast eruptions, identify potential hazards, and convey that information to stakeholders to minimize the threat that volcanoes pose.

Volcano observatories look for indications of unrest that volcanoes may exhibit prior to eruption. Potential signals of unrest include: increased surface temperatures, surface deformation, increased seismicity, increased degassing, and more. The type, occurrence, and frequency of unrest signals can vary substantially among different volcanoes, and as a function of deep and shallow processes occurring at any given volcano and time. While eruptions sometimes occur without any detectable precursory unrest signals, studies have shown that with robust monitoring it is possible to anticipate volcanic eruptions (Sparks, 2003; Tilling, 2008; Segall, 2013; Poland et al., 2020). A plethora of techniques are used to monitor volcanoes for the potential precursors listed above. Early volcano monitoring tools included manual temperature measurements of fumaroles and surface lava, seismograph stations, and simple tilt-measuring devices (Wood, 1913). Today, many volcanoes are monitored with real-time data from networked stations containing seismometers, global navigation satellite system receivers

(GNSS), webcams, gas monitoring instruments, and more (Guffanti et al., 2009). For many volcanoes, especially those without local monitoring stations, remote satellite observations can provide regular opportunities to monitor for deformation, degassing, and thermal activity (Poland et al., 2020). In this work we focus on monitoring volcanic unrest through satellite observations of surface temperatures. Specifically, we are looking for signals of localized zones of relatively high-temperature volcanic thermal emissions, or, more briefly – hotspots.

The presence of a hotspot in satellite data indicates that some area of Earth's surface is at a higher temperature than its surroundings. On volcanoes, temperatures above background can be produced by a variety of sources, including lava flows (Harris et al., 1997; Dehn et al., 2000; Hirn et al., 2009; Blacket, 2013), pyroclastic flows, dome growth (Carter et al., 2007), degassing of a hot vent and/or fumarole field (Blacket, 2013), or increased surface meltwater in the case of glaciated volcanoes (Pieri and Abrams 2005; Blackett 2013; Bleick et al., 2013; Reath et al., 2016). Examples of various eruptive products and signals of unrest are captured in high-resolution (1-30 meter pixel size) satellite imagery, and shown in Figure 1.1.

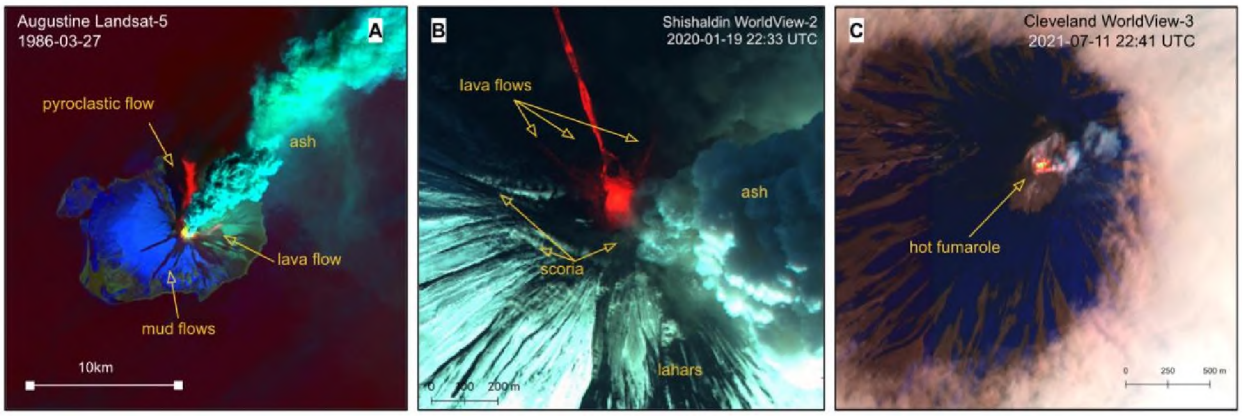

Figure 1.1: Examples of heat-producing phenomena on volcanoes. A) False color image of Augustine Volcano 1986, from Landsat-5, with Red = band 6 (TIR), Green = band 7 (SWIR), and Blue = band 3 (blue). With this band combination the hottest material appears yellow. B) False color image of Shishaldin Volcano 2020, from WorldView-2, with red showing the near-Infrared (NIR) band, and green and blue showing their true color. C) False color image of Mount Cleveland 2021, from WorldView-3, with red showing the short-wave Infrared (SWIR) band, and green showing the near-infrared, and blue showing true color. Images courtesy of Alaska Volcano Observatory, the US Geological Survey (USGS), the Alaska Department of Geological and Geophysical Surveys (DGGS), and K. R. Papp (A), H. Dietterich (B), and M. W. Loewen (C).

Many of the volcanic processes which release heat are eruptive, including lava flows, pyroclastic flows, and lava dome growth (Fig. 1A,B) – all of which bring significant amounts of magma to the surface at temperatures of 700-1300ºC (Philpotts and Ague, 2022). This results in large increases in heat emissions over the area covered by eruptive material. Heat can also be brought to the surface through non-eruptive processes. For example, fumaroles bring exsolved volcanic gases to the surface which can range in temperature from 100-1000ºC (Allaby, 2013). Similarly, hydrothermal systems can bring heat to the surface through magmatic or meteoric fluids which are heated at depth and can be up to 100ºC at the surface (Pipolo et al., 2017). By detecting hotspots on volcanoes and analyzing their temperatures it is possible to identify the type of activity occurring on the surface. Further, eruptions are often preceded by changes to the flux and temperature of degassing and hydrothermal systems (Edmonds and Woods, 2018), for which hotspot detections are able to provide insight. Detection of heat emissions at volcanoes – whether high temperature and eruptive, or lower temperature and non-eruptive – can provide many insights into the evolution and state of unrest of magmatic, volcanic, and hydrothermal systems. Hotspot detection algorithms have been used in the past to observe thermal precursors to eruption, track and characterize eruptions through time, quantify lava volumes, and more (Dehn et al., 2002; Harris et al., 2009; Wright 2016; Girona et al., 2021; Chevrel et al., 2023; Coppola et al., 2023). Due to the utility of these observations, thermal satellite data are used by volcano observatories as a part of daily monitoring operations (Dehn et al., 2000; Dehn et al., 2002; Harris et al., 2016; Coombs et al., 2018; Cameron et al., 2018; Coppola et al., 2020; Pritchard et al., 2022).

1.1 Thermal remote sensing

Hotspots can be detected in infrared satellite data due to the distinct signature of their radiation. The electromagnetic radiation produced by hotspots is characterized by Planck's Law (Planck, 1914), defined as

$$L_\lambda = \frac{2hc^2}{\lambda^5 e^{hc/K_B T\lambda} - 1} \qquad (1.1)$$

where L_λ is the spectral radiance ($W \cdot m^{-2} \cdot sr^{-1} \cdot \mu m^{-1}$) at a given wavelength (λ) and temperature (T, in kelvin), c is the speed of light (2.99×10^8 $m \cdot s^{-1}$), h is Planck's constant (6.626×10^{-34} $J \cdot s$), and K_B is the Boltzmann constant (1.38×10^{-23} $J \cdot K^{-1}$). Planck's law describes the radiation emitted by an ideal blackbody, that is, an object which absorbs radiation at all frequencies, reflects no radiation and allows no radiation to pass through. It states that as the temperature of a blackbody increases, the spectrum it emits will increase in radiance, and the peak radiance will shift to shorter wavelengths. Therefore, a high temperature volcanic hotspot can be identified by an elevated Thermal Infrared (TIR, 8 – 15 μm) radiance above background and an even higher Mid-Infrared (MIR, 3 – 8 μm) radiance above background (Blacket, 2013). For especially hot surfaces (>1000 K, Figure 1.2), the peak radiance emission is in the shortwave infrared (SWIR) part of the spectrum. Based on Planck's law, it is possible to calculate the temperature of a surface using the measured radiance at any wavelength. This is referred to as brightness temperature (BT), and can be derived by rearranging equation 1.1 (substituting Temperature, T, for brightness temperature, BT, to indicate that BT is derived and not known a priori):

$$\mathrm{BT} = \frac{hc}{K_B \lambda} * ln^{-1}\left(1 + \frac{2hc^2}{\lambda^5 L_\lambda}\right) \qquad (1.2)$$

However, the amount of radiation observed by a satellite sensor is not only a function of the surface temperature, but also the radiative properties of the surface as well as the transmittance of radiation through the atmosphere (Dehn et al., 2000; Harris et al., 2009; Rogic et al., 2019). Unfortunately, the ground does not behave like an ideal blackbody – it has variations in absorption and reflection of radiation at different wavelengths and for different surface properties (Aggarwal, 2004). Similarly, the atmosphere does not behave as a perfect transmitter of radiation – it has clouds and atmospheric effects which complicate how much radiation is able to pass through and at which wavelengths (Aggarwal, 2004). MIR and TIR bands are used in this study because they are within atmospheric windows, a range of wavelengths in which there is little absorption in the atmosphere. Still, transmittance through the atmosphere is not perfect and the brightness temperature calculated using equation 1.2 will not equal the true temperature at the surface.

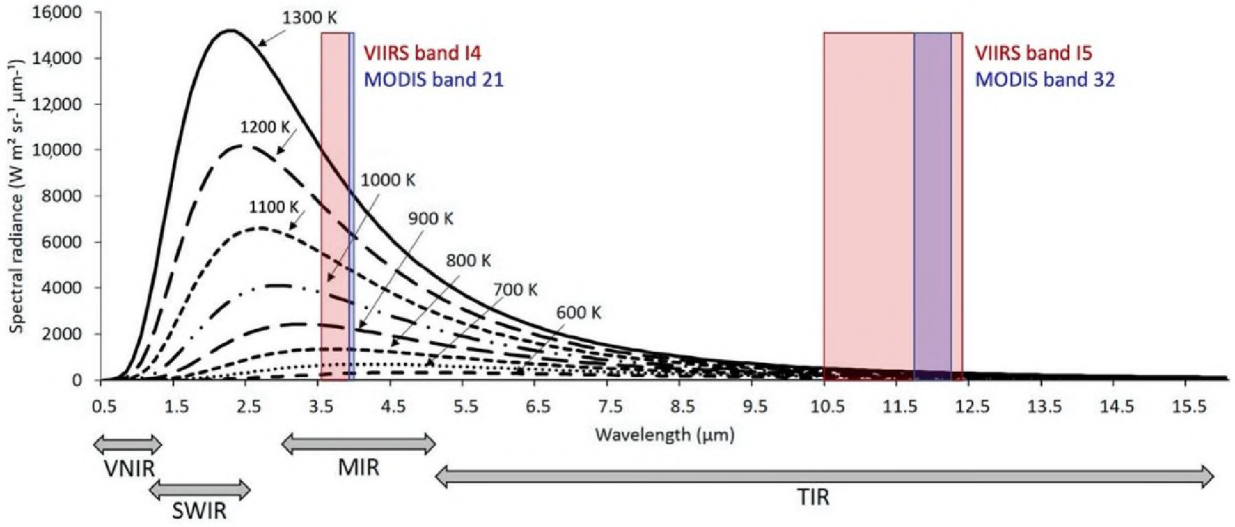

Figure 1.2: Emission spectra of a blackbody at various temperatures. Curves demonstrate the relationship between temperature, wavelength, and spectral radiance as described by Planck's Law. The spectral window of VIIRS imagery bands 4 and 5 are shown with red squares. The spectral window of MODIS bands 21 and 32 are shown with blue squares. Regions of the infrared spectrum are shown by their prefix (VNIR = very near infrared, SWIR = short-wave infrared, MIR = middle infrared, TIR = thermal infrarea). Modified from Blackett (2017).

One final consideration is that satellite sensors measure radiance values integrated over the area covered by each pixel. This is especially relevant for volcanic signals, which are often much smaller than the resolution of space-borne sensors, leading to underestimation of the maximum surface temperatures (Poland et al., 2020). For this project we are using data from two infrared sensors designed for meteorology, the Visible Infrared Imaging Radiometer Suite (VIIRS) aboard the Suomi-National Polar-orbiting Partnership (SNPP) and NOAA-20 satellites, and the Moderate Resolution Imaging Spectroradiometer (MODIS) aboard Aqua & Terra satellite platforms (Figure 1.2). The VIIRS bands we use have a resolution of 375 meters and the MODIS bands have a resolution of 1 kilometer, both of which are larger in size than many potential volcanic signals. Further details on these sensors and our motivation for using them can be found in section 2.3.

1.2 The need for automation

Infrared satellite imagery is used to monitor and forecast global weather, wildfires, and other natural hazards. The utility of this data in a variety of fields has resulted in more satellites,

and the development of new sensors with higher spatial and temporal resolutions. The availability of greater quantity and quality of thermal satellite data is a boon to the volcano monitoring community; however, it has also exceeded the capacity of human analysts. This motivates research to automate the detection and quantification of volcanic hotspots, so that imagery can be analyzed in near-real-time, and timely interpretations can be efficiently provided to volcano observatory scientists to inform decision making. Automated tools can make it easier to track real time thermal activity on volcanoes which may pose a threat, and also provide a mechanism to generate historical time series of thermal activity for volcanoes around the world. Observations over extended time periods can be used to determine baseline activity, identify periods of volcanic unrest, and characterize the thermal evolution of eruptions (Dehn et al., 2002; Wright 2016; Girona et al., 2021; Chevrel et al., 2023; Coppola et al., 2023).

In order to automate detection of hotspots, previous studies have used thresholds to automatically identify anomalous pixels (e.g., Wright et al., 2004). For example: if the brightness temperature of a pixel exceeds a certain value, that pixel is flagged as a hotspot. This can work to identify volcanic hotspots in some cases, but in other instances hotspot and background pixels can occur with similar radiance and brightness temperature values, even within the same image. Figure 1.3 illustrates the overlap of hotspot and background pixels in MIR and TIR brightness temperature data from Mount Veniaminof (labeling of these data is described in more detail in section 2.4). As this figure demonstrates, a simple thresholding approach would work to identify many hotspots with MIR brightness temperatures >330 K. However, there are many hotspot pixels that would go undetected with this thresholding approach (Figure 1.3). Thresholding will not work to identify the weakest thermal signals – caused by smaller hot areas or lower temperatures, which are the type of signals we expect to accompany precursory volcanic unrest. To address this issue many automated algorithms have been developed, each using some combination of band indices, spatial filters, and corrections in order to accentuate the differences between hotspot and background pixels (Higgins and Harris 1997; Wright et al., 2004; Pergola et al., 2004; Ganci et al., 2011; Coppola et al., 2016; Wright, 2016; Pergola et al., 2016; Murphy et al., 2016; Gouhier et al., 2016; Lombardo 2016; Valade et al., 2019; Genzana et al., 2020; Castaño et al., 2020; Mazzeo et al., 2021; Corradino et al., 2023). In sections 2.3 and 2.4, we take a closer look at two existing automated hotspot detection algorithms to elucidate their function and draw comparison with the model we developed.

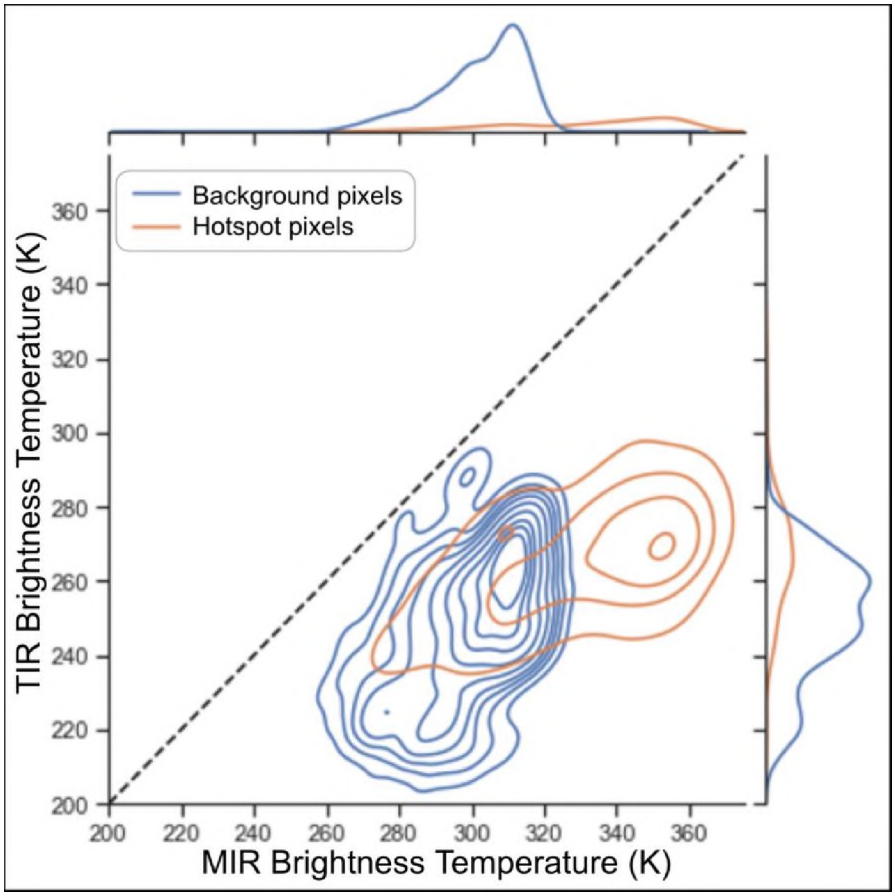

Figure 1.3: Brightness Temperature distributions of hotspot and background pixels. Pixels are selected from the VIIRS training dataset (labeled VIIRS data from Mount Veniaminof and Mount Cleveleland, see Table 2.2). While some hotspot pixels can be identified by thresholding (for example, if I4bt > 330'), still other hotspot pixels overlap with background pixels. This demonstrates that the two classes are not linearly separable in these dimensions, such that a simple thresholding approach will not work well.

Most existing automated techniques use various thresholding approaches to automate the flagging of hotspot pixels in pre-defined indices. The ability of each algorithm to distinguish hotspots from background pixels depends on how successfully the indices of the algorithm are able to separate the two classes (hotspot and background) and the accuracy and precision of the thresholds. Here, we take a different approach. Rather than defining our own indices, we use a convolutional neural network (CNN) which automatically learns the spectral and spatial patterns present in infrared satellite data to identify hotspot pixels. To realize this, we provide the model with thousands of images where hotspot and background pixels were manually labeled, and from that dataset the model derives spectral and spatial patterns to solve the classification problem.

Variants of CNNs have been applied to numerous problems in the field of computer vision, including to identify cancer cells in MRIs (Adoui et al., 2019), facial unlock in cellphones (Apple Support, 2018), and reverse image search algorithms (Wan et al., 2014). At the start of this project, we hypothesize that this data-driven approach has the capability to enhance hotspot detection, and detect subtle signals which might have been missed by other approaches. Indeed, a CNN has already successfully been applied to volcanic hotspot detection in imagery from the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) (Corradino et al., 2023). In section 2.4 we provide more detailed information about the theory of CNNs, and also explain the specific architecture and training process used here.

1.3 Introducing HotLINK

Our final trained model is called HotLINK: the Hotspot Learning and Identification Network. After training and testing (described in detail in section 2.4), HotLINK is applied to VIIRS data from 2012-2022 and MODIS data from 2000-2022 for eight target volcanoes: Augustine Volcano, Mount Cleveland, Bogoslof Island, Okmok Caldera, Pavlof Volcano, Redoubt Volcano, Shishaldin Volcano, and Mount Veniaminof. The final result of these analyses provides 22 years of hotspot detections for eight Alaska volcanoes, ten years of which have both VIIRS and MODIS observations. The three main questions we aim to address with these analyses are:

1. Is a CNN approach able to detect hotspots in infrared data better than a thresholding approach?
2. What volcanic processes can be identified in the time series generated?
3. Are any retrospective hotspot detections precursors to an eruption?

The second chapter of this thesis is a manuscript that is currently in review in the "Applications of Machine Learning in Volcanology" special issue of the Frontiers in Earth Sciences journal (https://www.frontiersin.org/research-topics/49896/applications-of-machine-learning-in-volcanology). The final chapter presents our overall conclusions and addresses the key questions raised above. We find that while our model showed good performance in the testing we conducted, and does an excellent job of tracking eruptions, the capability of the model to distinguish specific volcanic signals and identify precursory warming signals requires further analysis.

Chapter 2: Automatic identification and quantification of volcanic hotspots in Alaska using HotLINK: the Hotspot Learning and Identification Network

2.1 Abstract

An increase in volcanic thermal emissions can indicate subsurface and surface processes that precede, or coincide with, volcanic eruptions. Space-borne infrared sensors can detect hotspots – defined here as localized volcanic thermal emissions – in near-real-time. However, automatic hotspot detection systems are needed to efficiently analyze the large quantities of data produced. While hotspots have been automatically detected for over 20 years with simple thresholding algorithms, new computer vision technologies, such as convolutional neural networks (CNNs), can enable improved detection capabilities. Here we introduce HotLINK: the Hotspot Learning and Identification Network, a CNN trained to detect hotspots with a dataset of ~3,800 satellite-based, Visible Infrared Imaging Radiometer Suite (VIIRS) images from Mount Veniaminof and Mount Cleveland volcanoes, Alaska. We find that our model achieves an accuracy of 96% when evaluated on ~1,700 unseen images from the same volcanoes, and 95% when evaluated on ~3,000 images from six additional Alaska volcanoes (Augustine Volcano, Bogoslof Island, Okmok Caldera, Pavlof Volcano, Redoubt Volcano, Shishaldin Volcano). In comparison with an existing threshold-based hotspot detection algorithm, MIROVA (Coppola et al., 2016), our model detects 22% more hotspots and produces 12% fewer false positives. Additional testing on ~700 labeled Moderate Resolution Imaging Spectroradiometer (MODIS) images from Mount Veniaminof demonstrates that our model is applicable to this sensor's data as well, achieving an accuracy of 98%. We apply HotLINK to 10 years of VIIRS data and 22 years of MODIS data for the eight aforementioned Alaska volcanoes and calculate the radiative power of detected hotspots. From these time series we find that HotLINK accurately characterizes background and eruptive periods, similar to MIROVA, but also detects more subtle warming signals, potentially related to volcanic unrest. We identify three advantages to our model over its predecessors: (1) the ability to detect more subtle volcanic hotspots and produce fewer false positives, especially in daytime images; (2) probabilistic predictions provide a measure of detection confidence; and (3) its transferability, i.e., the successful application to multiple sensors and multiple volcanoes without the need for threshold tuning, suggesting the

potential for global application.

2.2 Plain word summary

Volcanoes release heat on their surface, and by monitoring this heat, we can determine if a volcano is erupting or might erupt soon. Heated areas, called hotspots, can be detected by satellite sensors, which generate images from space in infrared wavelengths. Traditionally, volcanologists or simple computer programs would identify the hotspots in infrared images. Now, advanced computer algorithms based on artificial intelligence can accurately identify complex features in images. We used these algorithms to improve the way we detect volcanic hotspots. Our approach detects more subtle heat signals than other algorithms, which is useful for detecting different types of volcanic activity, and may contribute to better forecasting of volcanic eruptions.

2.3 Introduction

Volcanic eruptions pose hazards to human life and society (Loughlin et al., 2015). To mitigate these hazards, volcano monitoring agencies aim to detect signs of unrest and eruption as early as possible. Local monitoring stations and remote satellite observations are commonly used to monitor volcanic unrest (e.g., Dehn et al., 2000; Cameron et al., 2018; Girona et al., 2021). Here we will focus on one satellite-based approach to monitor thermal unrest: detecting localized volcanic heat emissions, also referred to as volcanic hotspots. In a single satellite image, hotspots may be identified as a few pixels of elevated infrared radiance caused by high temperature volcanic features. Hotspots may be produced by various types of volcanic activity, including lava flows (Harris et al., 1997; Dehn et al., 2000; Hirn et al., 2009; Blackett, 2013), explosive and strombolian activity (Harris et al., 1997; Coppola et al., 2012; Coppola et al., 2014), dome growth (Carter et al., 2007; Ramsey et al., 2012; Coppola et al., 2022), degassing of a hot vent or fumarole field (Oppenhemier et al 1993; Harris and Stevenson, 1997; Blackett, 2013; Laiolo et al., 2017), or increased surface meltwater in the case of glaciated volcanoes (Pieri and Abrams 2005; Blackett 2013; Bleick et al., 2013; Reath et al., 2016). Therefore, monitoring changes in hotspot activity can provide key insights into a volcano's behavior by indicating the presence of

thermal volcanic features and characterizing them over time. Due to the utility of these observations, thermal satellite data are used by volcano observatories as part of their daily monitoring operations (Dehn et al., 2000; Dehn et al., 2002; Harris et al., 2016; Harris et al., 2017; Coombs et al., 2018; Cameron et al., 2018; Coppola et al., 2020; Pritchard et al., 2022; Chevrel et al., 2023). Automating the detection and quantification of volcanic hotspots can provide near-real time information to volcano observatory scientists to inform decision-making and provide a mechanism to generate long time series of thermal activity for volcanoes around the world. Time series observations are useful for determining baseline activity, identifying periods of volcanic unrest, characterizing the thermal evolution of ongoing eruptions, and retrospectively studying eruptive histories and processes (Dehn et al., 2002; Wright 2016; Girona et al., 2021; Chevrel et al., 2023; Coppola et al., 2023).

Surface hotspots will result in increased spectral radiance (Wm^{-2} sr^{-1} μm^{-1}) in both Mid-Infrared (MIR, 3 – 5 μm) and Thermal-Infrared (TIR, 5 – 20 μm) wavelengths (Harris, 2013). This behavior is characterized by Planck's Law, which states that as the temperature of a blackbody increases, the spectrum of energy it emits will increase in radiance, and the peak radiance will shift to shorter wavelengths. Therefore, a volcanic hotspot can be identified by an elevated TIR radiance above background and an even greater signal above background in MIR radiance (e.g., Blackett, 2013; Blackett 2017). For especially hot surfaces (>950 K), the peak radiance emission is in the shortwave infrared (SWIR, 1.4 – 3 μm) part of the spectrum. The distinct features produced by hotspots in MIR and TIR bands have been exploited to automate their detection by different algorithms (Higgins and Harris 1997; Wright et al., 2004; Pergola et al., 2004; Ganci et al., 2011; Coppola et al., 2016; Gouhier et al., 2016; Lombardo 2016; Valade et al., 2019; Genzano et al., 2020; Castaño et al., 2020; Massimetti et al., 2020; Layana et al., 2020; Ramsey et al., 2023; Corradino et al., 2023).

One of the first algorithms to automate volcanic hotspot detection, MODVOLC (Wright et al., 2004), applies a threshold to the Normalized Thermal Index (NTI), constructed from radiance values of MIR and TIR bands:

$$NTI = \frac{MIR - TIR}{MIR + TIR} \qquad (2.1)$$

MODVOLC flags nighttime pixels with NTI greater than -0.8, and daytime pixels with NTI greater than -0.55 as hotspots, because of the large impact of solar reflections and heating on daytime images (Wright et al., 2004; Wright, 2016). These thresholds were found by manual analysis of histograms of NTI at 100 locations to minimize false positive detections (Wright et al., 2004). Another popular approach, the MIROVA algorithm, incorporates a new spectral index in addition to NTI, and spatially filters both spectral indices to improve hotspot detections (Coppola et al., 2016, further details on the MIROVA algorithm and its application in this study can be found in section 2.4). While these and other algorithms define their own band indices, ratios, spatial filters, and corrections in order to accentuate the differences between hotspot and background pixels, each of these approaches use thresholding to automate the flagging of hotspot pixels. The ability of each algorithm to distinguish hotspots from background pixels depends on how successfully their index is able to separate the two classes, and the accuracy and precision of the threshold set for that index. MODVOLC and MIROVA have successfully generated decades long time series of hotspots at volcanoes across the globe, which has allowed for detection and monitoring of eruptions in near-real time and the study of thermal output from different eruptions and volcanic systems (Wright, 2016; Coppola et al., 2023). Still, both datasets contain false detections and missed hotspots, due to the fact that there will inevitably be non-volcanic thermal signals exceeding the set thresholds, and real volcanic signals lower than the detection thresholds.

In this paper, we aim to enhance the automatic detection of volcanic hotspots in infrared satellite data by applying a convolutional neural network (CNN). CNNs are a machine learning technique commonly employed for image analysis (LeCun et al., 2010). They have been applied to numerous problems in the field of computer vision, including to identify cancer cells in MRIs (El Adoui et al., 2019), facial unlock in cellphones (Apple, 2023), and reverse image search algorithms (Wan et al., 2014). In our approach the use of CNNs can be conceptualized as identifying hotspots based on what they look like, rather than by thresholding a particular thermal index. While previous methods employ human-created indices to highlight hotspot pixels, our approach is data-driven – deriving the spectral and spatial characteristics that define hotspots from a large labeled dataset of the hotspots themselves. Rather than defining our own indices, we label a large dataset of hotspots and then allow the model to learn patterns which

distinguish volcanic hotspots from background pixels. In this way, the CNN mimics the pattern recognition of a human analyst.

The type of CNN used here is a U-net (Ronneberger et al., 2015). U-nets are a popular architecture for image segmentation, or tasks in which a prediction is made for each pixel in order to both detect and locate features of interest. A U-net was successfully applied to volcanic hotspot detection in data from the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER), achieving a high accuracy (Corradino et al., 2023). In this study, we apply a similar method to data from the Visible Infrared Imaging Radiometer Suite (VIIRS) and Moderate Resolution Imaging Spectroradiometer (MODIS) satellite sensors. Although ASTER has a finer spatial resolution (90 m in TIR bands, used in Corradino et al., 2023) than VIIRS (375 m) and MODIS (1000 m), we chose to apply this methodology to VIIRS and MODIS data due to their high acquisition rates and MIR and TIR bands. High acquisition rates result in more frequent opportunities to detect and track changes in volcanic unrest. At the time of this writing, VIIRS sensors provide coverage of each Alaska volcano 8 – 15 times per day, while MODIS sensors provide coverage 1 – 6 times per day. Volcanoes at higher latitudes are imaged more frequently than those at lower latitudes by the polar-orbiting satellites used here. Detection frequency will increase in the future with the planned launch of additional VIIRS instruments. Although MODIS has a coarser spatial resolution than VIIRS, it has a longer operational history (satellites Terra and Aqua launched in 1999 and 2002, respectively), so it is useful for studying eruptions prior to the launch of VIIRS (Suomi-National Polar-Orbiting Partnership, SNPP, launched in 2011, and National Oceanic and Atmospheric Administration 20, NOAA-20, launched in 2017).

We incorporate data from eight Alaska volcanoes with a wide range of volcanic thermal signals to develop our model for broad applicability to many volcanic settings (Table 2.1). Alaska volcanoes have frequent eruptions, but are very remote, necessitating remote sensing as a primary method for eruption monitoring, forecasting, and response. We use images of Mount Veniaminof (Alaska) acquired between 2018 – 2019 covering an effusive-explosive eruption, and images of Mount Cleveland (Alaska) between 2017 – 2018 with coverage of lava dome growth in order to train our model. The Mount Veniaminof eruption captures high temperature basaltic lava flows into a large, ice-filled caldera (Loewen et al., 2021). Mount Cleveland activity consists of explosions, dome growth, and degassing within the summit crater of a stratovolcano

(Werner et al., 2017). These volcanoes are quite different in terms of morphology, eruption style, and governing subsurface processes. They also differ in the source of hotspot detections, namely lava surrounded by ice at Mount Veniaminof, versus hot rock surrounded by cold rock at Mount Cleveland. These source differences result in hotspots that may differ slightly in intensity and appearance, leading to a more robust model than it would be if trained on just one of these volcanoes alone.

Volcano	Eruptive styles	Eruptions within study period (2000-2022)
Mount Cleveland	Explosive, dome-building	2001, 2005, 2006, 2007, 2009, 2010, 2011, 2013, 2014, 2016, 2017, 2019, 2020
Okmok Caldera	Explosive, phreato-magmatic	2008
Bogoslof Island	Phreato-magmatic, explosive, dome-building	2016-2017
Shishaldin Volcano	Effusive, explosive	2004, 2014-2015, 2019-2020
Pavlof Volcano	Explosive, effusive	2007, 2013, 2014, 2016, 2021
Mount Veniaminof	Effusive, explosive	2002, 2004, 2005, 2006, 2008, 2009, 2013, 2018, 2021
Augustine Volcano	Explosive, dome-building	2006
Redoubt Volcano	Explosive, dome-building	2009

Table 2.1: Volcanoes used in this study, in order from west to east. Eruption dates and eruption styles are composited from information available on the Alaska Volcano Observatory website (www.avo.alaska.edu/).

The other six volcanoes in this study are used for model testing, and were chosen to comprise a wide range of edifice morphologies, magma compositions, eruption frequencies, and eruption styles. These include the frequently erupting and typically mafic volcanoes Okmok Caldera, Shishaldin Volcano and Pavlof Volcano, and the less frequently erupting and typically more silicic volcanoes Augustine Volcano, Bogoslof Island, and Redoubt Volcano. Importantly, all have erupted since the launch of the MODIS sensors. Although our development is focused in Alaska, the volcanoes compiled here range widely in terms of the thermal signatures we expect

to identify and the meaning of those signatures in terms of eruptive potential. This dataset can help to evaluate the effectiveness of the model across volcanic systems, and inform future application of the model.

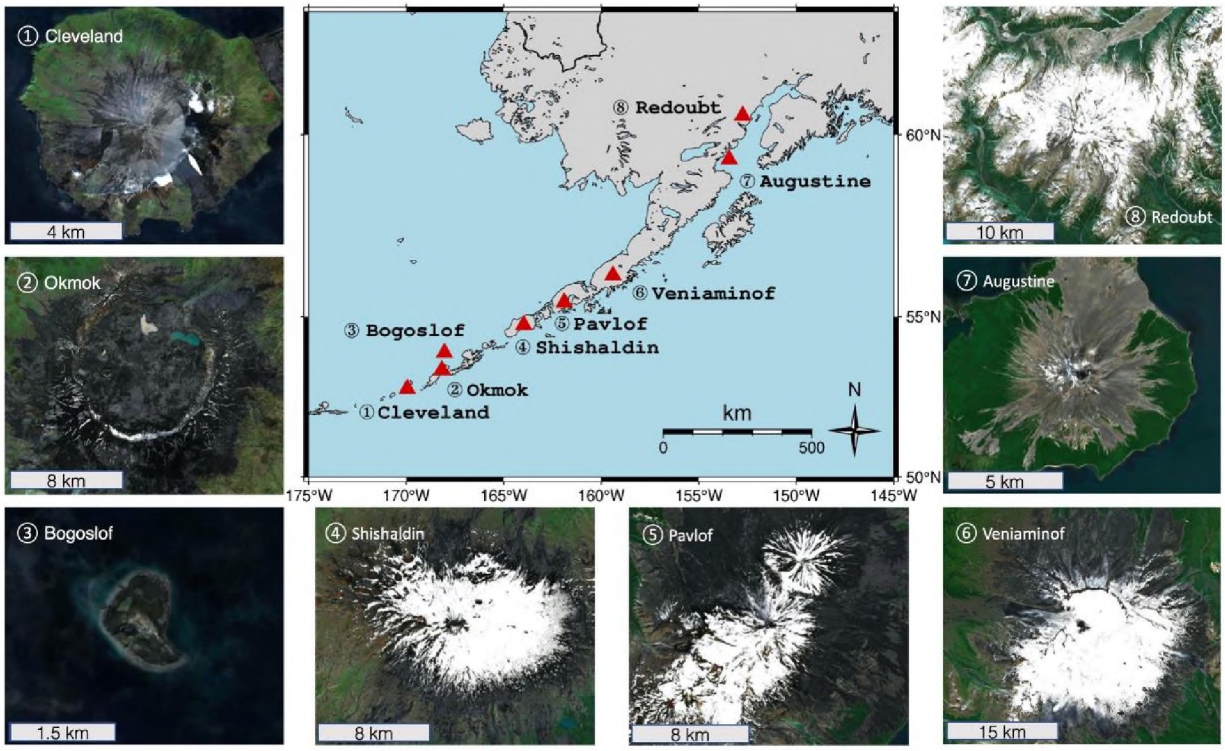

Figure 2.1: Volcanoes used in this study. The map in the center shows all volcano locations in Alaska. Numbered images shows high-resolution satellite data of the volcanoes at various zoom levels, from west to east (1) Mount Cleveland, (2) Okmok Caldera. (3) Bogoslof Island, (4) Shishaldin Volcano, (5) Pavlof Volcano, (6) Mount Veniaminof, (7) Augustine Volcano, and (8) Redoubt Volcano. Satellite data are from Sentinel-2 and composited by CalTopo to provide cloud-free viewing.

We call the final version of our trained U-net model HotLINK: the Hotspot Learning and Identification Network. After testing and training, HotLINK is applied to VIIRS data from 2012-2022 and MODIS data from 2000-2022 for the eight target volcanoes. The result of these analyses are 22 years of hotspot detections for these volcanoes, ten years of which have both VIIRS and MODIS observations. We also implement an optimized version of the MIROVA algorithm for our target volcanoes to compare the performance of the machine learning and thresholding approaches. We choose to compare our results with MIROVA because it is one of the most widely used algorithms for global volcanic hotspot monitoring, and was already

familiar to the authors. Through this work we hope to improve the accuracy of hotspot detections in infrared satellite data and share our methodology so that it can be applied elsewhere. We aim to address the questions: (1) is a CNN approach able to detect volcanic hotspots in infrared data better than a thresholding approach? (2) Can a computer vision model trained on VIIRS data be reasonably applied to MODIS data with a different resolution? (3) What are the limitations of HotLINK in terms of generalizability to other volcanoes, and detection limits for VIIRS and MODIS, night and daytime images? For each detection we calculate radiative power to quantify the heat emissions over the 22-year study period for the target volcanoes. We then discuss the capabilities and limitations of this approach for volcano monitoring.

2.4 Methodology

Our model takes as input a VIIRS or MODIS image with MIR and TIR bands, and outputs the probability that each pixel in a central region of the scene contains a volcanic hotspot. Once a hotspot is detected we calculate the total volcanic radiative power (RP in Watts) and area (m^2) of the hotspot. The methodology applied here involves the use of four separate VIIRS datasets to: (1) *train* the network, (2) *validate* hyperparameter selection (i.e., tuning parameters that configure the model and training, as opposed to parameters that are used within the model to make predictions), (3) *test* the model's accuracy when applied to new volcanoes, and (4) *analyze* detections and calculate RP for each volcano over an extended time period. Each of these four datasets (with names italicized above) is assembled for the VIIRS sensor, and additional test and analysis datasets are assembled for the MODIS sensor to produce six datasets in total (Table 2.2).

HotLINK is trained to detect hotspots in VIIRS infrared images on a manually labeled dataset (VIIRS training) of 3,783 images of Mount Veniaminof and Mount Cleveland volcanoes. We opt for a manual labeling approach because our goal is to create an automated system that simulates the manual hotspot identification which is done on a daily basis by duty satellite scientists at the Alaska Volcano Observatory (AVO). The same training dataset is used to optimize the thresholds of the MIROVA algorithm (Coppola et al., 2016), and results from both the optimized implementation of the MIROVA algorithm and HotLINK are compared using the same validation dataset, which consists of 1,275 images from the same volcanoes. After training

and validation, the accuracy of the model is estimated by applying it to the VIIRS test dataset, which is also manually labeled and consists of images from the six other Alaska volcanoes (Figure 2.1): Okmok Caldera, Shishaldin Volcano, Augustine Volcano, Redoubt Volcano, Pavlof Volcano, and Bogoslof Island.

Dataset	Labeled	Volcanoes (dates)	Number of images
VIIRS Training	By pixel	Veniaminof (2018), Cleveland (2018-2019)	3,783
VIIRS Validation	By pixel	Veniaminof (2018), Cleveland (2018-2019)	1,275
VIIRS Test	By image	Okmok, Shishaldin, Augustine, Redoubt, Pavlof, Bogoslof (Mar, Jun, Sep, and Dec 2017)	3,280 (includes 66 ambiguous images moved from the VIIRS validation dataset)
VIIRS Analysis	None	Veniaminof, Cleveland, Okmok, Shishaldin, Augustine, Redoubt, Pavlof, Bogoslof (2012-2022)	160,497
MODIS Test (Aqua)	By image	Veniaminof (2018)	634
MODIS Analysis (Aqua and Terra)	None	Veniaminof, Cleveland, Okmok, Shishaldin, Augustine, Redoubt, Pavlof, Bogoslof (2000-2022)	385,426

Table 2.2: Datasets used in this study.

Although HotLINK is only trained on VIIRS data, we test its applicability to MODIS data simply by inputting the MODIS test dataset into the VIIRS-trained HotLINK model. Data pre-processing for MODIS follows all of the same steps as for VIIRS data (see section 2.4.1). Finally, HotLINK is used to detect volcanic hotspots in 10 years of VIIRS data (VIIRS analysis dataset) and 22 years of MODIS data (MODIS analysis dataset) from all eight of the previously mentioned Alaska volcanoes. A subset of the MODIS analysis dataset (MODIS test data,

manually labeled for Mount Veniaminof) is reviewed and used to estimate the accuracy of the model when applied to MODIS.

2.4.1 Dataset pre-processing

The pre-processing for all VIIRS and MODIS datasets is the same. First, files containing any of the 8 target volcanoes are downloaded using the Atmosphere Science Investigator-led Processing System API (sips.ssec.wisc.edu) or NASA Earthdata portal (search.earthdata.nasa.gov). Next, terrain and atmospherically corrected radiance data (level 1b) are resampled onto a uniform grid of 64 x 64 pixels centered on the volcano using the nearest neighbor resampling method and the nadir pixel resolution. For VIIRS this corresponds to an area of roughly 24 x 24 km^2 and for MODIS this is an area of 64 x 64 km^2. We use VIIRS image bands I4 (3.55 – 3.93 μm, MIR) and I5 (10.5 – 12.4 μm, TIR), and MODIS bands 21 (3.929 – 3.989 μm, MIR) and 32 (11.77 – 12.27 μm, TIR). Spectral radiance values have the pixel area (m^2), spectral bandwidth (m), and angular aperture (steradians) factored out of the raw radiative power measurement (W), which allows for direct comparison between data from the two sensors, and normalization using the same factors.

Spectral radiance values (L) are normalized to the minimum (L_{min}) and maximum (L_{max}) possible radiance values for the VIIRS sensor, as determined by scale and offset factors (available in the VIIRS level-1b product user guide; NASA, 2018). Physically, L_{min} and L_{max} represent the limits of the sensor, and possible retrieval values are always within this range. Although the true radiance may be outside this range, the sensor will always return at least L_{min} and will saturate at values greater than L_{max} (NASA, 2018). The equation used to normalize the spectral radiance data is as follows:

$$L_{norm} = \frac{L - L_{min}}{L_{max} - L_{min}} \tag{2.2}$$

Normalization is important to prevent issues with vanishing or exploding gradients which would make it difficult for the CNN model to converge on a solution (Sola and Sevilla, 1997). We use the same L_{min} and L_{max} for both VIIRS and MODIS data despite the sensors having different minimum and maximum possible spectral radiance values. This is because once the

model has been trained on spectral radiance data normalized to a certain range, it must be applied to data normalized in the same way. Lastly, since VIIRS data saturates at a lower spectral radiance than MODIS data, some exceedingly rare MODIS pixels have values higher than one after normalization (<0.002% of pixels in the MODIS test dataset). To remedy this, values are capped at a maximum value of one.

The VIIRS training and validation datasets are assembled by collecting all (day and night) VIIRS data from the SNPP and NOAA-20 satellites with coverage of Mount Veniaminof for the year of 2018 and NOAA-20 VIIRS data (only) with coverage of Mount Cleveland for both 2017 and 2018. These volcanoes and time frames were selected to encompass background non-eruptive behavior, increasing unrest, and eruption. From this dataset, 75% of images are grouped into the VIIRS training dataset, and the remaining 25% are put into the VIIRS validation dataset. The validation dataset is smaller because it is only used to ensure the model is not overfitting, and a representative population is sufficient. Whereas the training dataset is larger because data in this group is used to actually train the model, and more data results in better model performance. The grouping between these two datasets is done randomly, with the exception that each image is grouped together with its closest temporal neighbor, since overpasses of SNPP and NOAA-20 satellites can be within ~45 minutes of each other. This prevents having one image in the training dataset and a nearly identical image in the validation dataset.

Images are manually classified into three groups: 'active' defined as images containing a volcanic hotspot, 'inactive' or images with no volcanic hotspot, and 'ambiguous,' where we cannot conclusively identify whether or not the image contains a volcanic hotspot (Figure 2.2). Next, all hotspot pixels within the active-labeled images are identified to construct pixel-wise masks. The ambiguous images are not used for training or validation since we only want images we can characterize with confidence in those datasets. All ambiguous images from the VIIRS validation and training datasets are moved into the VIIRS test dataset, which can have images of any class (66 ambiguous images in total are moved). The final training dataset contains 3,783 images and the final validation dataset contains 1,275 images. In both the VIIRS training and validation datasets, 45% of images are of Mount Veniaminof, 55% are of Mount Cleveland, and 32% of the total are classified as active.

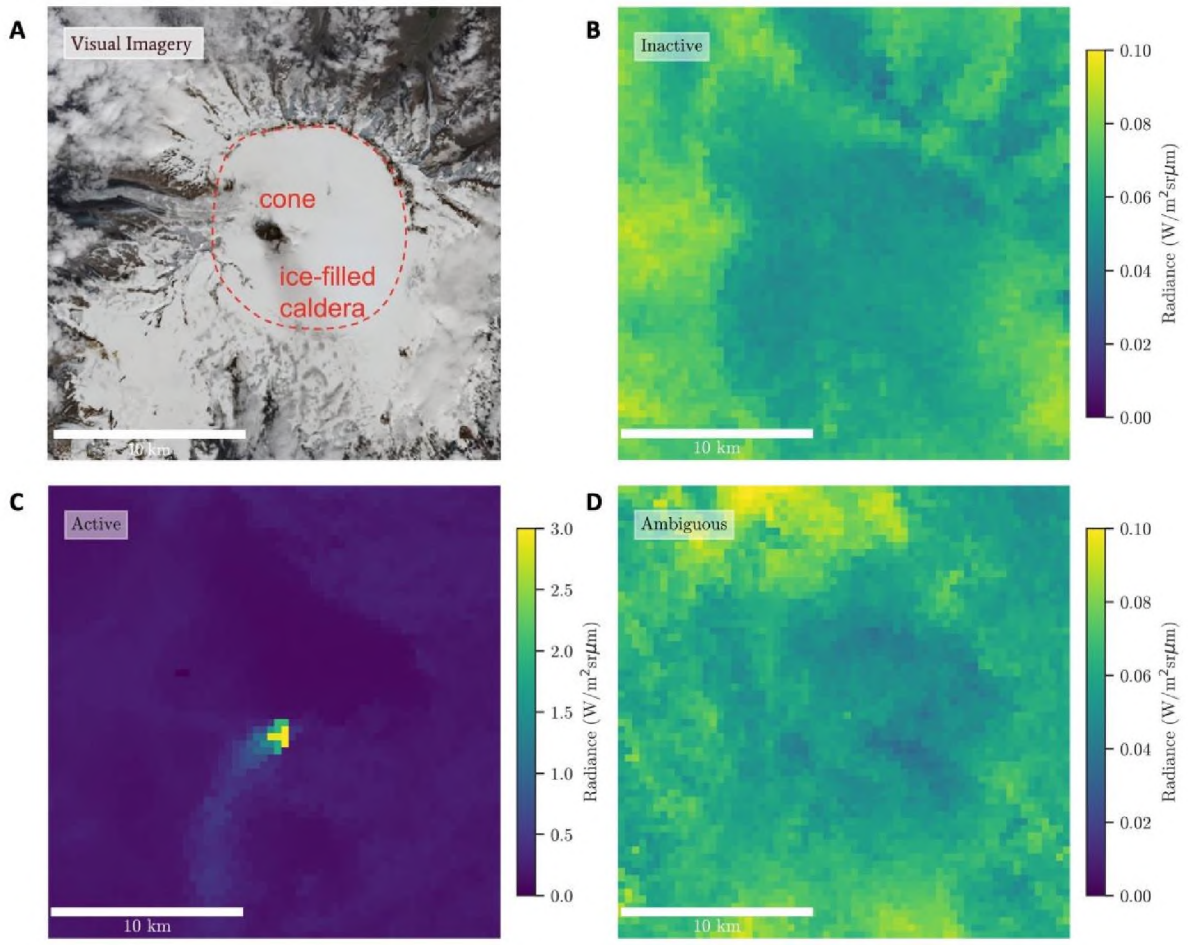

Figure 2.2: Classified example images. (A) Sentinel-2 visible RGB image (enhanced natural color visualization) of the ice-filled summit caldera and central cone. The classes of MIR VIIRS images used while training our model shown are (B) inactive – not containing a volcanic hotspot as identified by a human analyst, (C) active – containing a volcanic hotspot, and (D) ambiguous, which could not be confidently categorized into either class. All examples are nighttime images, showing the same cropped region of Mount Veniaminof (24 by 24 km). Note that images C and D have the same color mapping, but image B is scaled differently. Color bars show the range of radiance values in each image.

To evaluate how well the model generalizes to other volcanoes not used in training, a test dataset is assembled consisting of four months (March, June, September, and December 2017) of VIIRS data for the six additional Alaska volcanoes (Augustine Volcano, Bogoslof Island, Okmok Caldera, Pavlof Volcano, Redoubt Volcano, and Shishaldin Volcano). These months are chosen from throughout the year to capture the full extent of Alaska's seasonal variations. Of our target volcanoes, only Bogoslof Island had an eruption during 2017, so few volcanic hotspots are

expected in the VIIRS test dataset. Although choosing data from eruptive periods would have resulted in more hotspot detections, we elected to standardize the time period we were using for all volcanoes. The resulting dataset is a good indicator of the model's performance when applied to new volcanoes during typical conditions. Images in the test dataset are also manually classified as active, inactive, or ambiguous, but not further classified on a pixel-wise basis. Therefore, the VIIRS test dataset is only used to test the ability of the model to detect images containing hotspots, not whether it accurately retrieves all of the pixels associated with the hotspot.

The VIIRS analysis dataset consists of the remaining (unlabeled) data, which are analyzed by the trained model and used to generate a hotspot detection time series from 2012–2022 for each of the eight volcanoes in this study. It is the largest VIIRS dataset of our study, consisting of 160,497 individual images of the volcanoes. Note that the VIIRS analysis dataset encompasses data that is already a part of the VIIRS training, validation, and test datasets.

We generate additional MODIS test and analysis datasets in order to test the applicability of our model to MODIS data, compare time series results for VIIRS and MODIS, and extend the time series of detections back to the year 2000. The MODIS test dataset consists of all 2018 MODIS data from the Aqua satellite of Mount Veniaminof classified by image. This volcano and time period were chosen for the MODIS test dataset to encompass a known eruption at Mount Veniaminof that was included in the VIIRS training data. The MODIS analysis dataset consists of all MODIS data from both Aqua and Terra satellites from 2000 – 2022 with coverage of the eight target volcanoes.

2.4.2 U-net architecture and training

CNNs utilize 3 x 3 (or other sized) matrices, known as convolution kernels, to search for specific patterns within an image (LeCun et al., 2010). The kernel is moved across the image and multiplied with each 3 x 3 subsection to create a new filtered image that shows the degree of correlation between the features of the kernel and the image. This allows the network to identify and locate specific spatial patterns within the image. By stacking multiple layers of convolutions, the network is able to detect increasingly larger and more complex features. At first the network's kernels are populated randomly, but through an iterative training process the kernels are adapted to identify spatial patterns optimized for the task at hand.

Training a CNN involves inputting batches of labeled images into the model. As each image is passed into the model the probabilistic prediction (initially computed by the randomly initialized kernels) is compared to the truth value (the class of each pixel), which is known by prior manual analysis. Then a value, the "loss," is calculated to quantify how well the model prediction compares to the truth value. This is calculated by the "loss function," which, in simple terms, is a quantitative measure of how poorly the model performs – so, a lower loss score indicates better performance. Importantly, the loss function is differentiable with respect to the model – meaning that the gradient of the loss function can be calculated for the entire model. The gradient is very high dimensional, with a value for each trainable parameter of the entire model. By taking a small step in the direction of the gradient, each parameter of the model is adjusted slightly in the optimal direction to decrease the loss, which thereby increases the performance. With each pass over the training dataset, or epoch, each parameter is adjusted slightly, the loss decreases, and the performance of the model improves. This iterative training process is called gradient descent, since the model is descending step-by-step down the gradient of the loss function with the goal of reaching a local minimum. For a more comprehensive explanation of the training, underlying mathematics, and applications of CNNs, see LeCun et al. (2010).

We chose a U-net CNN architecture, because it allows for predictions to be made in the same resolution as the input (Figure 2.3; Ronneberger et al., 2015). This allows individual pixels to be flagged as hotspots or not. The input for our model is normalized radiance data from the MIR and TIR bands of the VIIRS or MODIS sensor, resampled to uniform resolution and cropped to 64 x 64 pixels centered on the main vent of the volcano of interest (64 x 64 pixels and 2 channels). The output is the probability that each pixel in a central area of the input belongs to one of three classes: background, hotspot, or hotspot-adjacent (24 x 24 pixels and 3 classes). The third class of pixels, hotspot-adjacent, helps the model to train faster; these pixels are considered background pixels during validation and testing. The output region is smaller than the input, due to the fact that convolutions of border pixels are undefined, resulting in a smaller image after each convolution. We consider that a 24 x 24 area of pixels is sufficient for detecting most hotspots (9 x 9 km^2 for VIIRS, and 24 x 24 km^2 for MODIS), but acknowledge that it may miss distal regions of large lava flows, or eruptions which occur far from the main vent.

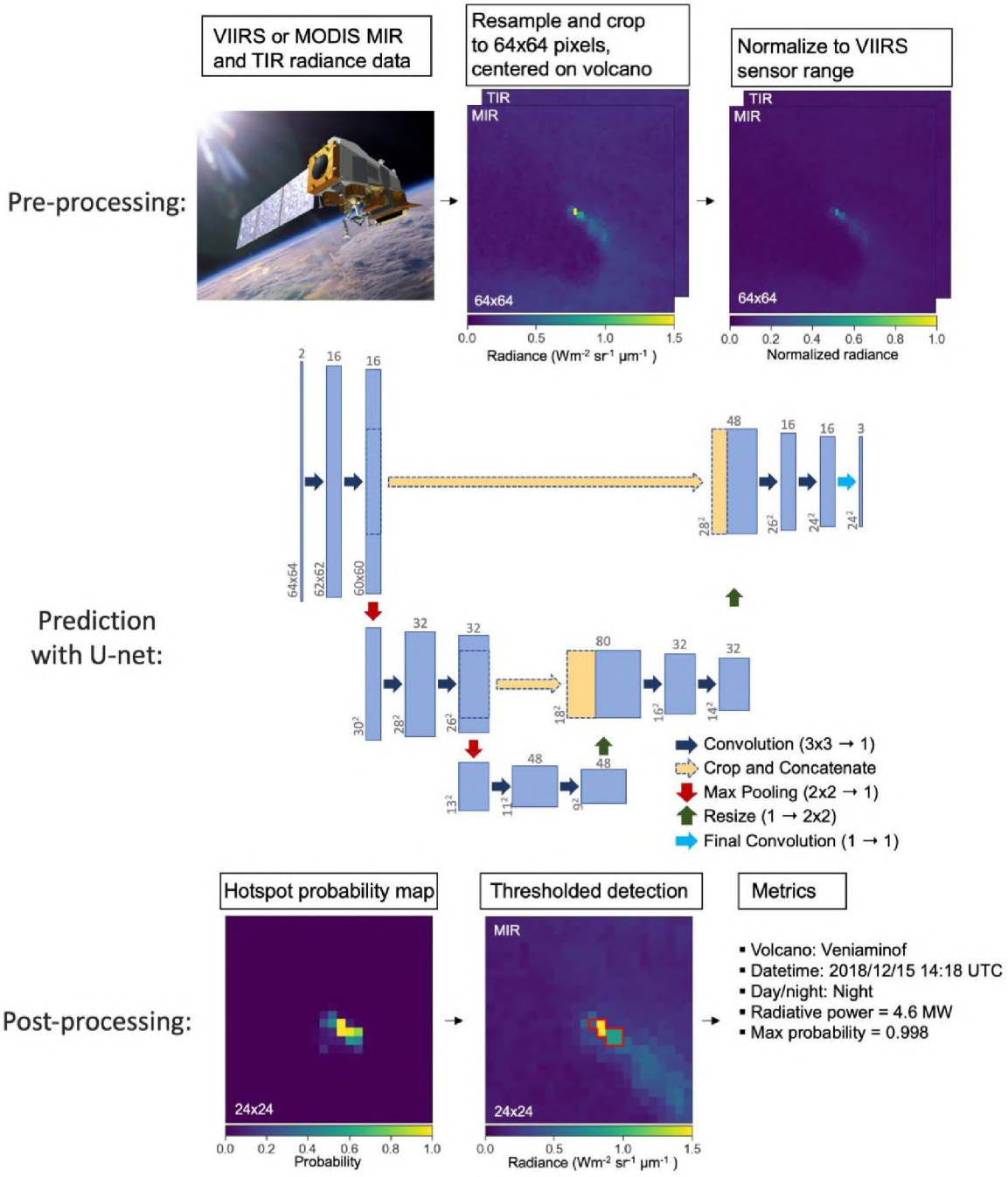

Figure 2.3: Steps of HotLINK processing: pre-processing, prediction with the U-net, and post-processing of a hotspot detection. Blue and tan rectangles of the U-net diagram represent data, the dimensions of which are labeled and denoted by the shape of the rectangles. For example, the input is [64 x 64 x 2] pixels and the output is [24 x 24 x 3] pixels. Note that at each convolution step the height and width of the data are decreased by two, since convolutions on the perimeter pixels are undefined. This progressive loss of perimeter pixels results in a prediction area significantly smaller than the input area. For further description of the motivation and function of the U-net architecture, see Ronneberger et al. (2015).

Many additional parameters can be adjusted in order to alter the architecture, training, or functionality of the model – these are referred to as hyperparameters. We experimented with many of these, selecting the values which result in the best performance (as measured by the validation dataset). Parameters that we tested include the random seed and distribution used to initialize the kernels (Glorot and Bengio, 2010), the number of convolutional filters used in each layer (i.e. the width of each rectangle in Figure 2.3), the gradient descent algorithm (Kingma and Ba, 2014), and the number of training epochs. We also tried many techniques to address the class imbalance in our training dataset. In the VIIRS training dataset approximately 25% of images contain a hotspot, while the remaining 75% do not. We explored several methods to mitigate the effects of the class imbalance, including: oversampling images with hotspots, undersampling the background images, using class weights, and using image augmentation to generate more training samples (details in the appendix). Out of the methods explored, only the image augmentation resulted in an increase in model performance. The rest of this paper only describes the final model, referred to as HotLINK, which uses the best hyperparameters found through dozens of training iterations.

HotLINK is trained on the VIIRS training dataset for 250 epochs, which is the point when the loss ceases to decrease for the validation dataset. During training, input images are augmented using 90° rotations and flips applied randomly after each epoch using the Albumentations library (Buslaev et al., 2020). This produces eight unique orientations for each original input image, which helps the model to learn only the most relevant features for prediction. The model is trained using the Adam optimizer (Kingma and Ba, 2014) with a sparse categorical cross entropy loss function, both of which are a part of the TensorFlow Python library (Abadi et al., 2015). Our U-net took ~2 hours to train on a 6-core Intel i7 processor, and after training makes predictions at an average rate of ~5 images per second. Further details on the specific hyperparameters used in the training of the HotLINK model can be found in the code itself, available in the appendix and on GitHub (https://github.com/csaundersshultz/HotLINK). Although we found these hyperparameters to work best for our problem, they may require modification for other hotspot detection applications.

2.4.3 Validation and testing

During the training process, we use the validation dataset to try out many different versions of the model in order to test which architectures, hyperparameters, etc., result in the best hotspot predictions. This process also helps to ensure that the model is learning patterns that are applicable to unseen data and not overfitting. Validation data are also used to tune threshold parameters applied to the output probability maps, and to compare HotLINK and our optimized application of the existing threshold-based algorithm, MIROVA (Coppola et al., 2016). To assess how the trained and validated model performs on new data, we use the test dataset, which is composed entirely of images from volcanoes the model has not seen during training.

We use two main metrics during validation and testing to evaluate HotLINK and MIROVA's performance: accuracy and F1-score. Accuracy is simply the percentage of images correctly identified by the model. It is defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.3}$$

where TP, TN, FP, and FN refer to the number of true positives (true hotspot detections), true negatives (true background detections), false positives (erroneous hotspot detections), and false negatives (missed volcanic hotspot detections), respectively, generated by the model. However, accuracy may not be the most appropriate metric for imbalanced datasets, which have higher proportions of some classes than others. For example, in this study a high percentage of images do not contain a volcanic hotspot. Therefore, a high accuracy could be achieved simply by predicting no hotspots in any image. A better metric for evaluating model performance in cases with imbalanced datasets is the F1-score (Ferri et al., 2009), defined as:

$$F1 = \frac{TP}{TP + \frac{1}{2}(FN + FP)} \tag{2.4}$$

The F1-score rewards true positive results and equally punishes false positives and false negatives, while true negatives have no impact on the score. Although our model predicts whether or not each pixel comprises a hotspot, accuracy and F1-scores are calculated on an

image-wise basis. Image-wise metrics are used to evaluate the model's ability to detect a hotspot, image-wise labeling is faster allowing us to create larger test datasets (Table 2.2). The training dataset is labeled for each pixel, since the U-net requires every pixel to be labeled in order to train.

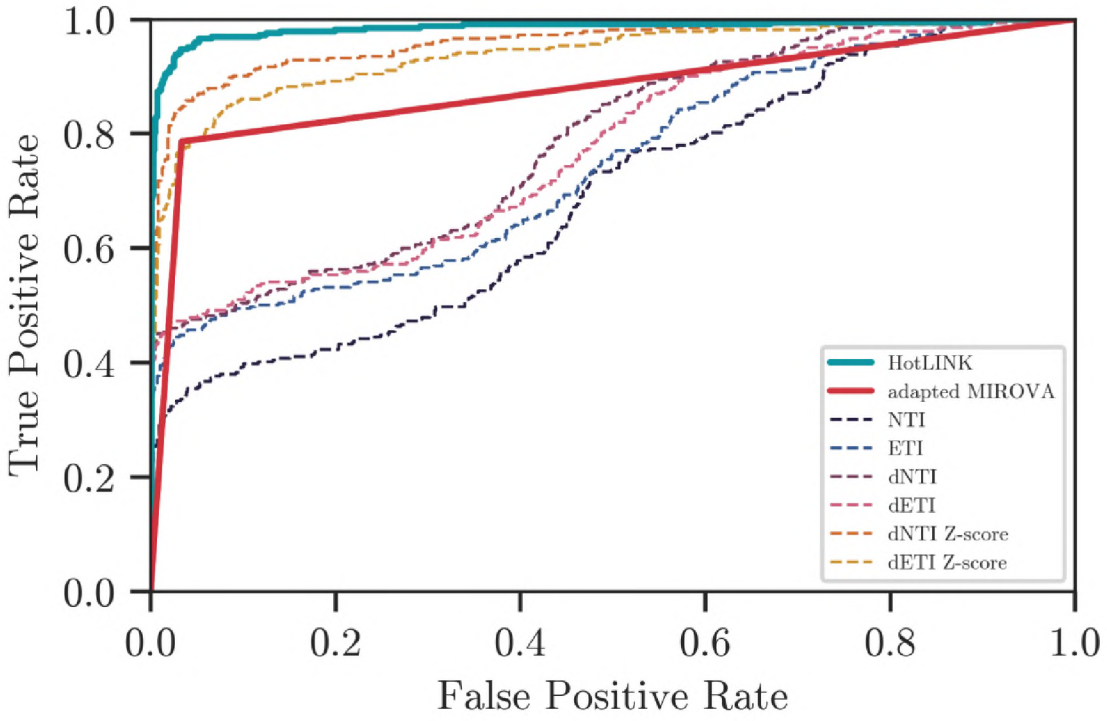

Figure 2.4: Receiver Operating Characteristic (ROC) curve applied to HotLINK and the adapted MIROVA algorithm. HotLINK probabilities are shown in blue, MIROVA prediction is red, and the different indices used in MIROVA are the thinner dashed lines. Preferred classifiers have a high true positive rate (TPR) and low false positive rate (FPR). Note that MIROVA consists of two straight lines because it produces just a binary output.

Another way to compare HotLINK and our optimized MIROVA algorithm is by using receiver operating characteristic (ROC) curves, which provide a graphical means to characterize the effectiveness of binary classification models (Figure 2.4). For a given index or predicted probability, an ROC curve plots the true positive rate against the false positive rate achieved by thresholding at different values. In this way it shows the tradeoff between false positives and true positives. For example, setting a low threshold will achieve a high true positive rate at the expense of more false positives, and setting a high threshold will achieve a low true positive rate while providing fewer false positives. ROC curves plot a model's performance at all possible

thresholds, thereby showing a particular model's ability to identify hotspots with low FP and FN rates. The ROC curve comparison of HotLINK and MIROVA is further discussed in section 2.5.2.

2.4.4 MIROVA optimization on the VIIRS training dataset

In order to test the performance of HotLINK, we compare our results to the MIROVA algorithm, which was originally developed for use with MODIS data (Coppola et al. 2016). The MIROVA algorithm has already been applied to VIIRS data (Campus et al., 2022, using moderate resolution bands; Aveni et al., 2023, using the same image bands used here). However, these studies use the original thresholds of the MIROVA algorithm that were designed for use with MODIS data. Since VIIRS and MODIS have different spatial resolutions and slightly different spectral bands, it is possible that the original thresholds could be improved for use with VIIRS data. To make a fair comparison between MIROVA's threshold methodology and our model, we optimize the thresholds of the MIROVA algorithm using a grid search over the same VIIRS training dataset that is used to train HotLINK. This allows for an unbiased comparison, ensuring that any observed performance differences can be attributed to the inherent capabilities of each model rather than variations in the data used.

MIROVA employs three thresholds (C1, C2, and K) on multiple indices calculated from the MIR and TIR spectral bands. These indices are the Normalized Thermal Index (NTI), Enhanced Thermal Index (ETI), spatially filtered versions of the first two indices called dNTI and dETI, and the Z-scores of dNTI and dETI. These indices are designed to increase the contrast between hotspot and background pixels, by combining spectral information at each pixel (indices NTI and ETI) with spatial information from surrounding pixels (indices dNTI and dETI) and the scene as a whole (Z_{dNTI} and Z_{dETI}). A full description of the algorithm and definitions of indices are presented in Coppola et al. (2016). In brief, pixels are flagged as active if the index NTI is greater than the threshold K, or if the indices dNTI, dETI, and the Z-scores of both, surpass the C1 and C2 thresholds, respectively:

$$(NTI > K)\ or$$
$$\left((\mathrm{dNTI} > C_1)\ \text{or}\ (Z_{dNTI} > C_2)\right)\ \text{and}\ \left((\mathrm{dETI} > C_1)\ \text{or}\ (Z_{dETI} > C_2)\right) \tag{2.5}$$

In order to optimize MIROVA for use with VIIRS data, we conudct separate grid-searches for nighttime and daytime data to define new threshold values for C1 and C2, which minimize the error rate on images within the VIIRS training dataset. The daytime grid search is conducted between C1 values of 0.0 – 0.29 with a stepsize of 0.01, and C2 values of 2.0 – 11.75 with a stepsize of 0.25. The nighttime samples are more sensitive to the C1 threshold so we use a finer stepsize of 0.005 and smaller range of 0.0 – 0.095. The C2 range and stepsize remain the same for the nighttime grid search. At each step the accuracy of MIROVA using specific thresholds is calculated. The K threshold was not optimized because it was found to have little effect on the pixel selections made by the algorithm, so it was left as the default value of -0.8 for nighttime images and -0.6 for daytime images. Default MIROVA values for daytime data are C1=0.02 and C2=15, and for nighttime data are C1=0.003 and C2=5. With our grid search we found the highest accuracy using values of C1=0.11 and C2=6.25 for daytime data, and C1=0.075 and C2=5.25 for nighttime data (see Figures A.3 and A.4 in the appendix for visualization of both grid searches). The grid searches demonstrate that slight changes to threshold values can result in slight increases in the performance of MIROVA, at least when applied to our particular dataset.

2.4.5 Hysteresis thresholding and Radiative Power calculation

Some final considerations for implementing the model are choosing how to threshold pixels in the output probability map (Figure 2.3), and then calculating useful metrics for each detection to better track changes in volcanic thermal emissions over time. Although each pixel is predicted with an individual probability, we recognize that a pixel is more likely to be a hotspot if it is adjacent to a hotspot pixel. For that reason, we implement hysteresis thresholding, in which a high threshold is used to initialize hotspot detections and a lower threshold is used to continue them. Here, all pixels with a probability greater than 0.5 are classified as hotspots, and pixels with a probability greater than 0.4 are classified as hotspot pixels if they are adjacent to other hotspot pixels. The high threshold is set by optimizing the validation dataset for image F1-score, and then the low threshold is set by optimizing for pixel-wise F1-score. To clarify, these metrics are chosen because only the high threshold determines which images are active, while the low threshold determines which pixels within the image are active.

Once active images are detected and all hotspot pixels within those images are identified, radiative power (RP) is calculated following the method of Wooster et al., (2003), using the following formula:

$$RP = C \times A_{pix} \times \sum^{n} L_{pix} - L_{BG} \quad (2.6)$$

where RP is the radiative power measured in Watts, C is a constant of proportionality that is specific to the sensor ($sr^{-1}\mu m^{-1}$, 18.9 for MODIS and 17.34 for VIIRS), A_{pix} is the area of the pixel in kilometers squared (1 km^2 for MODIS, 0.14 km^2 for VIIRS), n is the number of pixels in the hotspot, L_{pix} is the radiance of each hotspot pixel ($Wm^{-2}sr^{-1}\mu m^{-1}$), and L_{BG} is the mean radiance of pixels directly surrounding the hotspot detection ($Wm^{-2}sr^{-1}\mu m^{-1}$, following the established methods of Wooster et al., 2003). RP is a measure of how much energy is released over the entire hotspot, and includes corrections for pixel size, central wavelength, and background radiance. Since pixel size and central wavelengths are different for VIIRS and MODIS, using RP allows us to make direct comparisons between the two sensors.

2.5 Results

2.5.1 Validation and test results

Results on the VIIRS validation dataset (Table 2.3) show that the final model works well when applied to data that has not been seen during training but comes from the same volcanoes. Specifically, both Mount Veniaminof and Mount Cleveland validation data yield model accuracies >95% and F1-scores >0.9.

On the VIIRS test dataset, which includes data from the six volcanoes that the model has not seen previously, HotLINK achieves a relatively low F1-score of 0.667 (Table 2.3). This seemingly poor performance is best explained by the lack of true hotspots in the dataset used; out of the six volcanoes, only Bogoslof Island erupted during the sampling period of the test dataset (Table 2.2). Since F1-score is mainly a function of true positive detections we achieve a poor score on most of the volcanoes since there were not many true hotspots to detect. False negative

and false positive rates on all datasets do not exceed 4%, except for the Augustine Volcano false negative rate, which is 7.9%.

Dataset	Accuracy	F1-score	TN	TP	FN	FP	Count
VIIRS Training	0.952	0.914	0.698	0.254	0.031	0.017	3781
Cleveland	0.962	0.898	0.795	0.167	0.017	0.021	1551
Veniaminof	0.945	0.920	0.631	0.314	0.041	0.014	2230
VIIRS Validation	0.962	0.923	0.731	0.231	0.022	0.016	1275
Cleveland	0.977	0.933	0.820	0.157	0.011	0.011	527
Veniaminof	0.951	0.919	0.668	0.282	0.029	0.020	748
VIIRS Test	0.947	0.667	0.908	0.049	0.024	0.019	2956
Augustine	0.914	0.172	0.901	0.009	0.079	0.011	547
Bogoslof	0.955	0.892	0.765	0.189	0.024	0.022	460
Okmok	0.956	0.512	0.927	0.024	0.024	0.026	468
Pavlof	0.974	0.723	0.936	0.037	0.008	0.019	483
Redoubt	0.919	0.608	0.940	0.040	0.002	0.019	530
Shishaldin	0.979	0.444	0.970	0.009	0.002	0.019	468
MODIS Test (Veniaminof)	0.981	0.954	0.786	0.195	0.019	0.0	646

Table 2.3: HotLINK results on training, validation, and test datasets. Each row shows the average of all volcanoes first, and then indented below specific values for each volcano in the dataset. Note that ambiguous images (195 total) are removed prior to this analysis.

2.5.1 HotLINK results on MODIS test data

The MODIS test dataset consists of all Mount Veniaminof data from the Aqua satellite in 2018, including 634 images in total. HotLINK achieves an accuracy of 98% on the MODIS test dataset, and an F1-score of 0.95 (Table 2.3). Unexpectedly, this performance is better than the model performs on VIIRS data. In section 2.6.3 we discuss a possible explanation for this.

2.5.2 HotLINK and adapted MIROVA results on the VIIRS validation dataset

The VIIRS validation dataset is used to compare the results of HotLINK and the optimized MIROVA algorithm after both models are trained/optimized with the VIIRS training dataset. On the validation dataset, we find that HotLINK outperforms our implementation of the MIROVA algorithm in all metrics (Table 2.4). Specifically, HotLINK produces more true positives (fewer missed detections), and more true negatives (fewer false detections) than the

MIROVA approach. Both methods score higher on nighttime data than daytime data. The conditions under which each model performs best is further discussed in section 2.6.4.

Model	Accuracy	F1-score	Night F1-score	Day F1-score	TN	TP	FN	FP
HotLINK	0.962	0.923	0.929	0.916	0.731	0.231	0.022	0.016
Adapted MIROVA algorithm	0.921	0.834	0.894	0.765	0.722	0.198	0.054	0.025

Table 2.4: Comparison of HotLINK and the adapted MIROVA algorithm on the VIIRS validation dataset. Metrics shown are: accuracy, day/night/combined F1-scores, and ratio of True Negatives, True Positives, False Negatives, and False Positive detections.

The ROC curve (Figure 2.4) further demonstrates that HotLINK (blue line) outperforms the MIROVA algorithm implementation (red line) with respect to true and false positives. In this plot, preferred classifiers have a high true positive rate (TPR) and low false positive rate (FPR). So better classifiers are those which plot further into the top left corner. These results show that HotLINK performs better than the overall optimized MIROVA algorithm, as well as all of the individual indices used by the MIROVA algorithm (thin dashed lines) with respect to TPR and FPR. This indicates that HotLINK is able to better differentiate hotspot and background pixels in comparison with individual indices, regardless of threshold selection. This is due to the CNN's ability to extract additional spatial information compared to manually tuned spatial filters.

2.5.3 Time series results

After applying HotLINK to the validation and test datasets, we apply HotLINK to the VIIRS and MODIS analysis datasets. This provides 10 years of VIIRS and 22 years of MODIS hotspot detections for the eight target Alaska volcanoes. These results can be found in Figure 2.5. Despite being unlabeled, these results can help provide a qualitative check on the effectiveness of the model when applied to different volcanoes experiencing background, unrest, or eruptive behavior. All detections found in this dataset are plotted as time series in Figure 2.5, with the Alaska Volcano Observatory (AVO) Aviation Color Code as the background color. In this analysis we use the AVO Aviation Color Code as a proxy for the state of activity of the volcano. A color code of "green" is used to indicate that a volcano is at a background non-eruptive state, "yellow" indicates increasing unrest with the possibility of an eruption in the future, "orange"

indicates that effusive or low-level explosive eruptions are occurring or are expected in the immediate future, "red" indicates a significant explosive eruption is occurring or imminent, and "unassigned" (colored as gray in Figures 5 and 7) indicates that there is insufficient ground-based monitoring data to assign a color code (Guffanti and Miller, 2013). While accuracy metrics are useful, the time series plots demonstrate the utility of HotLINK in practical applications. Figure 2.5 illustrates that HotLINK succeeds at detecting eruptions, which are accompanied by significant increases in the frequency and RP of detected hotspots. This figure also shows patterns of potential false positive detections during non-eruptive periods at all volcanoes, which are discussed in the following paragraphs.

Mount Cleveland erupts frequently, as indicated by many periods of orange color code in the timeline (Figure 2.5), which represent lava dome eruptions and other elevated activity (e.g., Werner et al., 2017). The Mount Cleveland time series shows numerous hotspot detections, which are much more frequent during periods of orange color code compared to when the color code is unassigned.

Okmok Caldera had only one eruption during our analysis period, in 2008. Only MODIS data is available for this eruption, from which there was one nighttime and three daytime detections during the eruptive period all with RP values >5 MW. Steady detections occur in VIIRS night and daytime data at Okmok Caldera, which we infer may be due to the presence of lakes within the caldera.

At Bogoslof Island we see a strong seasonal trend, in which VIIRS daytime detections and associated RP increase in the summer and decrease during winter. These seasonal trends are observable both before and after the 2016 – 2017 eruption, but are stronger post-eruption. The 2017 Bogoslof Island eruption is captured well, with VIIRS nighttime detections producing higher RP values than at any other time.

At Shishaldin Volcano, extended eruption periods from 2014 – 2016 and 2019 – 2020 are tracked well by HotLINK detections. The onset of these eruptions are accompanied by significant increases in the rate and RP of detections, and the end of eruptions are accompanied by a return to background values.

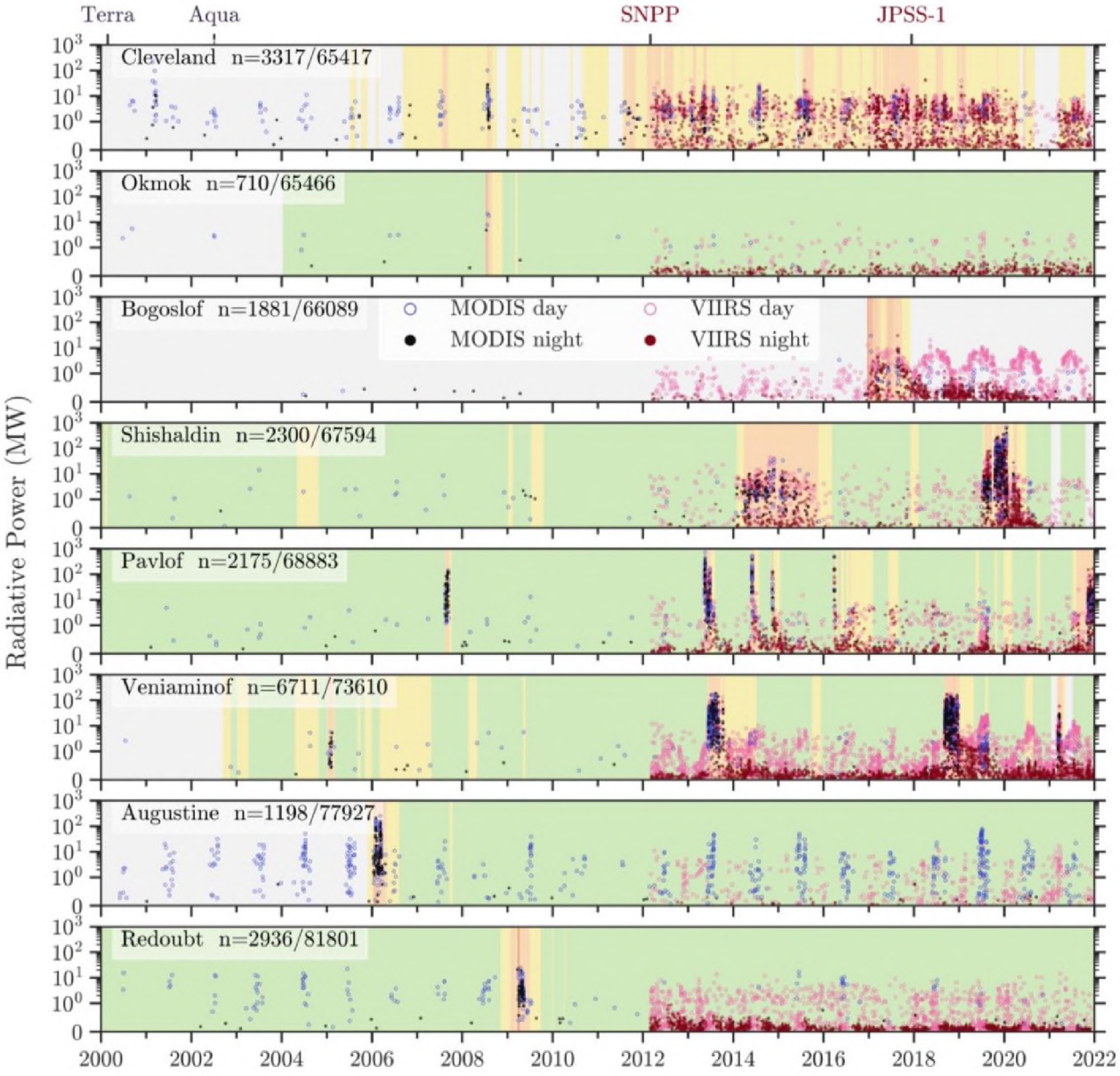

Figure 2.5: Time series results of HotLINK detections and calculated radiative powers for all eight target volcanoes: Mount Cleveland, Okmok Caldera, Bogoslof Island, Shishaldin Volcano, Pavlof Volcano, Mount Veniaminof, Augustine Volcano, Redoubt Volcano. The AVO color code at each volcano is shown as the background color of each figure for general context on the state of activity at the volcano (see section 2.5.3 for description of color codes), with gray indicating a period with insufficient monitoring data for AVO to assign a color code ("Unassigned;" Guffanti and Miller, 2013). The RP of individual hotspot detections are shown as points (MODIS = black, VIIRS = red), with lighter and darker shades representing day and nighttime image acquisitions. Next to the name of each volcano is the number of total detections at each volcano, and the number of total images for each volcano in both the VIIRS and MODIS analysis datasets. Note that for all plots the y-axis scale is linear between 0–1 MW, and logarithmic >1MW. The top axis shows the start of data acquisition from satellites used.

Pavlof Volcano eruptions are detected well by the HotLINK system, with RP values during eruptive episodes significantly higher than during non-eruptive periods. The 2007 eruption is captured well in MODIS data, and subsequent eruptions are captured well in both VIIRS and MODIS data.

At Mount Veniaminof there have been multiple eruptions that are detected by HotLINK, but there is also a high rate of background detections, which could either be indicative of background heat output or potentially the emissivity and thermal inertia differences between the active cone and surrounding glacier. In section 2.6.4 we further discuss the nature of these signals.

Augustine Volcano had one observed eruption in 2006. Augustine Volcano has infrequent VIIRS nighttime detections, but does show a seasonal signal with increased VIIRS daytime detections during winter and increased MODIS daytime detections during summer.

Redoubt Volcano also had only one eruption during our analysis period, in 2009, which was detected well in MODIS data. Since then, no anomalous thermal activity has been detected but there have been frequent hotspot detections in VIIRS nighttime and daytime data, which may be attributed to localized persistent degassing and snow melt on the 2009 lava dome.

2.6 Discussion

In this section we discuss the time series results at all volcanoes to investigate the strengths and weaknesses of our model. We also discuss the probabilistic output of HotLINK, and our finding that probabilities are well calibrated. Next, we compare VIIRS and MODIS applications of HotLINK, and estimate detection limits for each sensor. Finally, we advance our comparison of HotLINK and the threshold-based MIROVA algorithm by looking at a case study of the Mount Veniaminof time series.

2.6.1 Analysis of time series results from all volcanoes

Based on the time series of detections at all volcanoes (Figure 2.5), we find that (1) the HotLINK model, as currently trained, works well for many, but not all volcano morphologies/settings, (2) the VIIRS sensor has a lower detection limit than MODIS due to a finer spatial resolution, which also results in a slightly higher false positive rate for VIIRS, and

(3) the RP and relative frequency of daytime and nighttime detections reveals distinct periods in the eruptive chronologies at many volcanoes, which can be used to further discern true and false detections. We discuss how we can discern true and false hotspot detections during non-eruptive periods at volcanoes, why false positive detections appear more often in some volcanoes during certain times of the day and year than others, and how results can be further filtered to remove many of the false detections.

Although HotLINK has a lower false positive rate than MIROVA in the validation dataset (Table 2.4), in the analysis dataset we still see nearly continuous hotspot detections at all volcanoes even between eruptive periods (Figure 2.5). Even though HotLINK makes many detections when volcanoes are at "green," or a background state (e.g. Okmok Caldera 2012-2022), that does not mean that all of those detections are false positives as it is common for many volcanoes to be persistently degassing and producing heat at the surface even in absence of an eruption. In this case, increases in the rate and RP of detections, rather than the detection of a single hotspot, may indicate volcanic unrest or eruption. However, as testing shows (Table 2.3), we expect HotLINK to have a false positive rate ~2%, such that some of the detections during background periods are likely not true volcanic hotspots.

In our analysis of Figure 2.5, we expect true volcanic hotspot detections to be those which are spaced closely together in time and at higher RP than other detections observed during periods with no eruptive activity. At all volcanoes, likely false positives seem to occur in VIIRS daytime images with RP in the range of ~1 – 10 MW, and in VIIRS nighttime images with RP ~0 – 0.5 MW. We determine that most detections with RP above these thresholds are true positives, but that does not preclude the possibility of true (but weak) volcanic hotspot detections within those ranges.

At some volcanoes (Bogoslof Island and Augustine Volcano) there are notable seasonal variations in the number of detections and the RP of those detections. At these volcanoes we believe the source of these detections is primarily from diurnal effects on land/water boundaries. For example, both Bogoslof Island and Augustine Volcano are island volcanoes, which means that during the day the land surface regularly heats up more than the surrounding ocean, creating a temperature difference that is visible in infrared images and to our model looks like a volcanic hotspot. Since Bogoslof Island is ~1.5 km in diameter while Augustine Island is ~12 km in diameter, Bogoslof Island tends to appear more like a hotspot in daytime VIIRS data while

Augustine Volcano Island regularly is identified as a hotspot in daytime, summer, MODIS data (Figure 2.5). Similarly, clouds frequently develop during the daytime on land, creating localized solar reflections.

A similar effect occurs at volcanoes that have crater lakes/lagoons (e.g., Okmok Caldera and Bogoslof Island). Since water has a higher thermal inertia than land, it preserves solar heat longer into the night than land and is commonly warmer than land at night, particularly when the land is snow-covered. Volcanic lakes are commonly connected to hydrothermal systems and increasing lake temperature can be linked to volcanic activity (Hurst et al., 1991; Rouwet et al., 2014). However, increasing lake temperatures due to volcanic thermal input are difficult to distinguish from increasing temperatures due to diurnal patterns. With that in mind, a hotspot detection of a lake is not necessarily indicative of increased volcanic or hydrothermal activity. By looking at trends in detections and RP over time, however, HotLINK may have the capability to characterize background lake temperatures and thus detect deviations above background. In our data we did find clear examples of diurnal and seasonal cycles in hotspot detections at Okmok Caldera and Bogoslof Island. However, in neither case did we observe clear deviations in the background radiative power that might have been caused by increased volcanic activity. Example images of false detections at Okmok Caldera and Bogoslof Island and comparison with high resolution true color imagery are available in the supplementary materials (Figures A.5 and A.6). Other common effects producing non-volcanic hotspot detections are snow melting off rocky areas that then become solar-heated (Mount Veniaminof), and clouds or volcanic plumes reflecting solar radiation.

While these non-volcanic sources of apparent hotspots are considered in our study to be false-positives, they highlight the capability of HotLINK to detect subtle warming signals that could be successfully applied to other research problems. Fundamentally there will always be a tradeoff between the sensitivity of the method to detect real volcanic hotspots, and the number of false positives produced. With this in mind, there are simple ways to minimize the occurrence of the false positives in the dataset through filtering. One easy approach is to only use the nighttime data, which is much less susceptible to false positives, especially those occurring on exposed rocks surrounded by snow and ice fields and solar reflection off clouds or plumes. Another way is to set a specific probabilistic threshold. In Figure 2.5, we calculated radiative power for all images containing any pixels whose probability exceeds 0.5. However, this probability could be

adjusted for different contexts. For example, if conducting a long-term historical analysis, it may be better to set a high confidence threshold and remove as many false positives as possible. Conversely, for near-real-time monitoring it may be important to incorporate as many detections as possible, even if a greater percentage of them might be false.

To illustrate the effects of further filtering the data, we look at time series from Bogoslof Island, Okmok Caldera, Redoubt Volcano, and Augustine Volcanoes, each of which only had one eruption during the time period of study. At all four of these volcanoes combined there are 6,725 total detections made out of 291,283 total images analyzed (Figure 2.5). These statistics yield a combined detection rate of 2.3% (>97% of images are non-detections). However, if we use only night time data and set a probabilistic threshold of 0.75 at the same volcanoes, HotLINK detects 2,661 hotspots out of 168,400 total images, which is a detection rate of 1.6%. So, with a higher threshold and only using nighttime images HotLINK removes >98% of images as non-detections. These statistics also help us estimate an upper bound on the false positive rate of HotLINK at around 2%, which is similar to what we calculated earlier with the VIIRS test dataset. For comparison to detection rates during eruptions see section 2.7.3 in which detection rates of VIIRS and MODIS sensors at Mount Veniaminof during eruptive periods are discussed.

2.6.2 Analysis of HotLINK probability estimates

In order to use probabilistic predictions from HotLINK for filtering hotspot detections, or for future incorporation into forecasting methods, we must verify that the probabilistic predictions of the model are meaningful. This is especially relevant since modern neural networks have shown a tendency to be overconfident (Guo et al., 2017). Although the model outputs a probability prediction for each pixel in the image, we are most interested in whether the image contains a hotspot at all. Therefore, for the purposes of this analysis we refer to 'image probability' as the highest probability of all pixels in the image, since it only takes one hotspot pixel for an image to be classified as active. We evaluate our probability outputs using a reliability diagram, adapted from Hamill (1997; Figure 2.6A).

For image probabilities to be well calibrated, we want the accuracy of a thresholded prediction to scale with its probability (Hamill, 1997). For example, if a well-calibrated model predicts five images to contain hotspots at a probability of 80%, four of the images would contain hotspots while one would not. While this may seem counterintuitive, we want some

images with high probabilities to be wrong in order to confirm that probabilistic predictions are reliable. We find a strong correlation between the probabilities of HotLINK predictions and whether images contain a hotspot, since they align with the ideal distribution (black line) shown in the reliability diagram below (Figure 2.6A). This demonstrates that the probabilistic output of HotLINK can be considered a well-calibrated estimate.

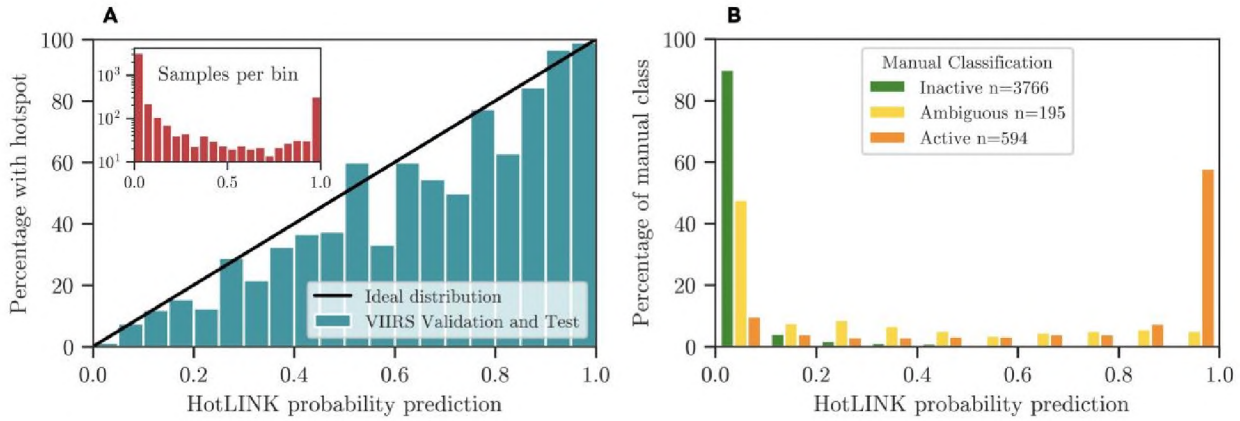

Figure 2.6: Reliability diagram and histogram of VIIRS validation and test datasets. A) Reliability diagram of the HotLINK model applied to the VIIRS training and validation dataset (unambiguous images only). Blue bars represent the proportion of images manually identified as active in 5 percentile bins. The black line represents the ideal probability distribution, indicating that probability predictions are accurate to the true classification. Bars below the black line are overconfident (probability prediction of hotspots is higher than the true probability), and bars above are under-confident (probability prediction of hotspots is lower than true probability). The inset figure shows the number of samples per bin on a logarithmic scale. B) Histogram of the VIIRS validation and VIIRS test datasets, showing the percentage of each class - inactive (green), active (orange), and ambiguous (yellow) - in 10 percentile bins. Ambiguous images are the most represented class at intermediate probabilities (0.1-0.8).

While the reliability diagram (Figure 2.6A) demonstrates that probabilities are well calibrated, we can expand our probabilistic analysis by including the ambiguous images identified by human visual inspection. The ambiguous images contained in the VIIRS validation and test datasets present a great opportunity to compare HotLINK's probability predictions to images we could not confidently classify as volcanic or not. Figure 2.6B shows that ambiguous images are skewed toward low probabilities, with ~50% of ambiguous images predicted at a probability <0.1. However, ambiguous images are proportionally more represented than each other class in all bins from 0.1 – 0.8. In other words, ambiguous images are much more likely to

be predicted at intermediate probabilities than images labeled as inactive or active. This finding supports the idea that CNNs mimic the visual learning of human experts. It also provides more confidence in the quality of probabilistic predictions, since images that appear ambiguous to analysts are likely to be predicted at intermediate probabilities by the network.

2.6.3 Comparison and detection limits of MODIS and VIIRS data

We speculate that the higher accuracy of HotLINK on the MODIS test dataset relative to the VIIRS test and validation datasets is due to the larger pixel size of MODIS preventing small hotspots from being identified by either HotLINK or manual analysis, resulting in an increased number of true negatives for MODIS compared to VIIRS. Similarly, the larger pixel size blurs out smaller scale background variance that is visible in VIIRS data, such that MODIS has a lower false positive rate than VIIRS and a higher F1-score. The larger pixel size of MODIS data results in fewer detections overall than VIIRS.

HotLINK shows a slightly better accuracy on MODIS data than on VIIRS because the MODIS data contains a greater proportion of true negatives and a smaller proportion of false positives. Despite this, the VIIRS data has a higher true positive rate and is able to see smaller and weaker hotspots. To further support this conclusion we compare VIIRS and MODIS detections during three eruptive events at Mount Veniaminof from the analyzed datasets. From these eruptions we also attempt to quantify a night and daytime detection limit for HotLINK when applied to VIIRS and MODIS data.

Mount Veniaminof had three eruptions between 2012 – 2022, the time period when both VIIRS and MODIS data are available. These eruptions were effusive-explosive in nature, characterized by lava effusion into and within the intra-caldera glacier, and sporadic ash emissions (Waythomas et al., 2021, 2023; Loewen et al., 2021). Start and end dates for these eruptions are taken from Loewen et al. (2021). During these eruptive periods, both VIIRS and MODIS agree well on RP estimates in our analysis. For the 2013 eruption (June 13 – Oct 17), both MODIS and VIIRS retrieved an average RP of 27.8 MW. During the 2018 eruption (Sep 4 – Dec 27) MODIS retrieved an average of 27.6 MW and VIIRS 30.2 MW, and for the 2021 eruption (Feb 28 – Apr 21) MODIS retrieved an average of 6.0 MW and VIIRS 5.0 MW (Figure 2.7).

Although the average RP retrieved by both sensors is comparable, the VIIRS sensor had a much higher rate of detections during the same eruptive periods. Across all three eruptions, VIIRS had 1,553 detections out of 2,874 total images, for an active percentage of 54%. Meanwhile MODIS had 536 detections out of 1,902 total images, for an active percentage of 28%. We hypothesize VIIRS had a greater active percentage because it was able to capture significantly weaker signals, due to its finer spatial resolution (0.137 km^2 compared to 1 km^2 pixel area at nadir). In future work, this hypothesis could be tested through a more robust analysis of the relative detection rate of VIIRS and MODIS images that are captured at nearly the same time.

To approximate detection limits for both sensors using HotLINK, we use the 5^{th} percentile radiative power of all hotspots detected during the 2013, 2018, and 2021 eruptions at Mount Veniaminof. It is important to acknowledge the possibility of false positives in these data, constituting approximately 2% of samples according to the labeled VIIRS validation and test datasets (Table 2.3). To mitigate the impact of false positives on the detection limit estimate, we opt for a conservative approach by using the 5th percentile, which is more than twice the estimate for the percentage of false positives in the dataset. This ensures that potential low RP false positives do not artificially lower the detection limit estimate. Still, our estimate for detection limit is not the threshold at which signals are missed, but approximates this by indicating the weakest signals retrieved by HotLINK. This estimate allows us to compare the relative detection limits between sensors. For VIIRS data, we find the 5th percentile of daytime detections to be 0.69 MW, and nighttime detections to be 0.26 MW. For MODIS data, we find the 5th percentile of daytime detections to be 1.4 MW, and nighttime detections to be 0.79 MW. These results demonstrate that HotLINK is 1.8 – 3 times more sensitive to nighttime observations than daytime observations, and that HotLINK is 2 – 3x more sensitive when applied to VIIRS data compared to MODIS. To compare with literature values, the MIROVA algorithm applied to MODIS data cites a detection limit of ~1 MW irrespective of the time of day (Coppola et al., 2020). This is the first time the authors are aware of a comparison of the detection limits between MODIS and VIIRS I-bands, although the radiative power between MODIS and VIIRS M-bands (750 m at nadir) have been previously compared, finding that the VIIRS M-bands are more sensitive than MODIS bands to thermal signals (Li et al. 2018, Campus et al., 2022). We caution that these detection limits are only approximations, since we are only using one volcano

for this analysis and are not looking at the radiative power of missed detections. Detection limits could be more rigorously ascertained by comparing the radiative power of true positive and false negative detections across many volcanoes. Here we only calculated the radiative power for images that were detected as hotspots by HotLINK and statistical analysis of the RP of false negative detections was not done.

2.6.4 Analysis of HotLINK and adapted MIROVA on the Veniaminof time series

Table 2.4 shows a higher true positive rate of HotLINK relative to our implementation of the MIROVA algorithm, indicating a greater sensitivity to smaller and lower temperature hotspots. Similarly, the high true negative rate of HotLINK relative to this adapted MIROVA indicates that HotLINK is less susceptible to false positive detections. We can expand on this analysis by examining the Mount Veniaminof time series from 2017 – 2021 to further compare results during eruptive and inter-eruptive periods (Figure 2.7). During this time period there were two eruptions, one in 2018 and one in 2021. The main difference between HotLINK and the optimized MIROVA detections during this period is that HotLINK detects more hotspots. From an eruption tracking perspective, the MIROVA algorithm does well as it has a similar detection rate as HotLINK during eruptions. In contrast, during non-eruptive periods HotLINK makes a greater number of detections than MIROVA, which may represent volcanic thermal output associated with volcanic unrest, as well as false positives. Therefore, while both models perform well for eruption detection and tracking, HotLINK is able to detect weaker signals that may be relevant for monitoring unrest at Mount Veniaminof.

Figure 2.7 shows an increase in HotLINK detected RP prior to the 2018 eruption, and more peaks in 2019 and 2020 that are not seen in MIROVA data. These HotLINK detections are consistent with Alaska Volcano Observatory analyst checks of VIIRS MIR images, where analysts observed weakly to moderately elevated surface temperatures qualitatively prior to eruption at Mount Veniaminof, and again during discrete time periods in the summers of 2019 and 2020 (Figure 2.7C,D; Cameron et al., 2023; Orr et al., 2023). We therefore find that the HotLINK detections are real, capturing weaker, but notable above-background thermal signals as seen in both the rate and radiative power of detections. These HotLINK results also have the advantage of providing quantitative information in comparison to the qualitative AVO remote

sensing database classifications of "barely elevated," "moderately elevated," and "saturated/incandescent."

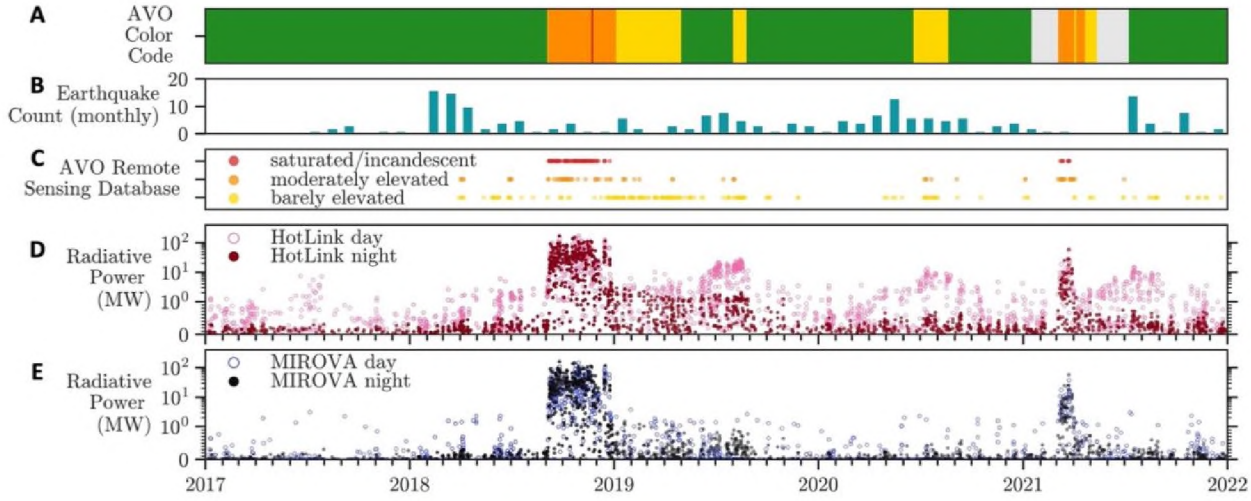

Figure 2.7: Multidisciplinary observations at Mount Veniaminof. Subplots show: A) Alaska Volcano Observatory Aviation Color Code timeseries, with color code levels indicated by their respective colors, and gray indicating periods with insufficient monitoring data for AVO to designate a color code (Guffanti and Miller, 2013). B) Monthly earthquake counts within 20 km of Mount Veniaminof, assembled from the USGS ComCat earthquake database (https://earthquake.usgs.gov/earthquakes/search/). C) Analyst flags from the AVO remote sensing database, showing analyst identified thermal signals in VIIRS images, characterized as being "saturated," "moderately elevated," and "barely elevated." D) Mount Veniaminof hotspot detections in VIIRS images from 2017–2022 using HotLINK and E) hotspot detections from the adapted MIROVA algorithm.

Inspection of these signals through complementary high resolution optical satellite imagery (e.g. Sentinel-2, Maxar) suggests that they comprise a combination of subtle surface heating, potentially due to increased vent degassing behavior at the volcano, as well as a seasonal signature due to the still-warm 2018 lavas readily melting the overlying snow cover in spring. The 2018 pre-eruptive hotspot signals suggest increased thermal output, perhaps via increased degassing or ground surface temperatures of the active cone (Orr et al. 2019). The 2019 and 2020 peaks in RP coincide with seasonal snow melting that exposed the large and relatively-warm lava flow field, but these signals also coincide with seismic unrest noted by AVO that prompted AVO to raise the color code from green to yellow on 1 August 2019 for 24 days and on 18 June 2020 for 64 days (Orr et al., 2023; Cameron et al. 2023). Further analysis of the detected radiative power in comparison with complementary multiparameter datasets and higher resolution infrared

images (e.g., Figure 1.1) could help tease out the origins and processes associated with these detections.

Our analysis shows that while both HotLINK and MIROVA are able to detect large and high temperature hotspots (e.g. Figure 2.8A), more subtle hotspots (Figure 2.8C) are only detected by HotLINK. The MIROVA system struggles to disregard bright and dispersed signals, such as solar reflections off of clouds, which exceed thresholds defined in the algorithm, but are visibly not hotspots in context (Figure 2.8B). HotLINK is able to detect more subtle hotspots that may be weak but still match the spatial patterns of a discrete thermal signal. The detection capabilities of HotLINK are similar to what an analyst can detect by eye.

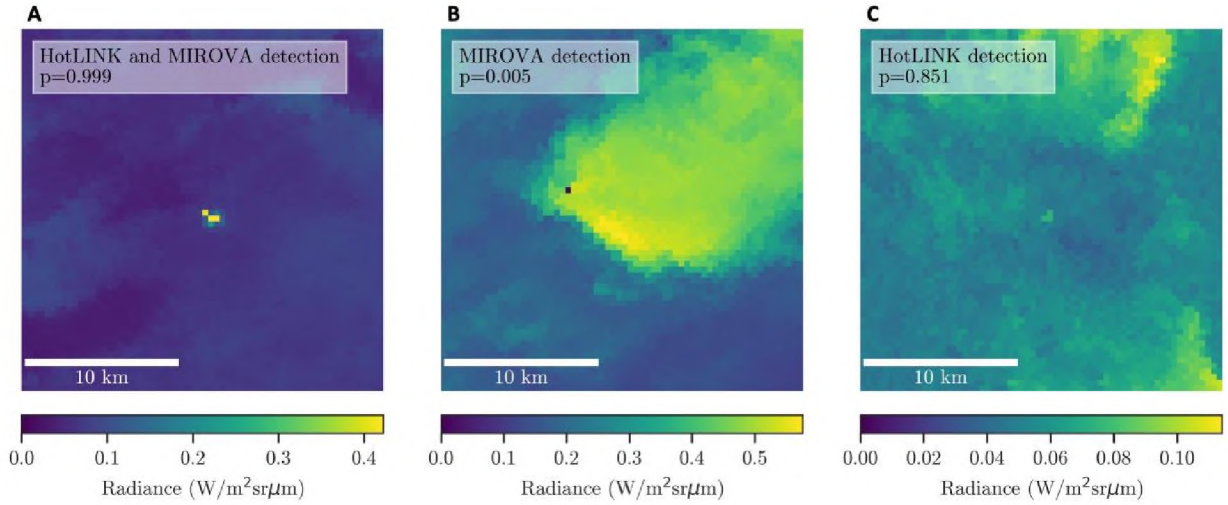

Figure 2.8: Example images from the VIIRS validation dataset. All images show MIR spectral radiance (Wm^{-2} sr^{-1} μm^{-1}) at Mount Veniaminof. (A) a true hotspot detection made by both HotLINK and the adapted MIROVA algorithm (nighttime image), (B) a false positive detection of a bright cloud made by the adapted MIROVA algorithm (daytime image), and (C) a true positive detection of a more subtle hotspot made by HotLINK, which is missed by the adapted MIROVA algorithm (night image). All images are 64 x 64 pixels, or roughly 24 x 24 km. Note that each image has its own colorbar scale in order to show the maximum contrast within each image.

2.7 Conclusions

This study confirms the capability of machine learning, specifically convolutional neural networks (CNNs) to automate remote sensing tasks usually designated to human experts (Corradino et al., 2023). This technology provides three main improvements relative to threshold-based algorithms: (1) the model is more sensitive to subtle signals and can detect a

larger number of hotspots while also detecting fewer false positive hotspots, (2) the probabilistic nature of the detections makes the model useful for different monitoring contexts, and (3) the same model performs well on data from different sensors (MODIS and VIIRS) and different Alaska volcanoes (with some caveats for volcanoes that are islands or have crater lakes).

The ability to detect more and weaker hotspots opens up the possibility of detecting precursory as well as eruptive hotspot signals. Specifically, our network detects subtle increases in volcanic surface temperature from Mount Veniaminof that correspond with both increased number of analyst detections of thermal signals and elevated seismicity. The capability to detect subtle signals associated with volcanic unrest, as well as eruptions, may aid in eruption forecasting efforts. Another advantage of our network is the probabilistic output. This expands the amount of information available to human analysts and will facilitate incorporation into statistical eruption forecasting models.

We found that HotLINK was able to detect hotspots in MODIS data with an even higher accuracy than for VIIRS data. Our model is therefore directly applicable to both VIIRS and MODIS data and is shown to work well on multiple volcanoes, only producing large errors in cases with crater lakes or small island volcanoes, which are especially susceptible to seasonal false detections. These errors could be minimized in the future using a detection threshold that exceeds the seasonal background signals at relevant volcanoes and/or by filtering out daytime images.

In conclusion, with a labeled training dataset of less than 4,000 VIIRS images from two volcanoes we were able to train a model to detect hotspots in both VIIRS and MODIS data that is applicable to many volcanoes. The time series for the eight volcanoes analyzed here captures volcanic unrest and eruption and thus can provide critical input into data-driven volcano monitoring and forecasting studies, as well as valuable insight into the magmatic and eruptive processes occurring in active volcanic systems across Alaska. The model itself is also readily applicable for near-real-time or historical hotspot detection efforts by volcano observatories.

Chapter 3: Overall conclusions

The goal of this study was to engineer an automated volcanic hotspot detection model, based on modern computer vision principles and implementing a convolutional neural network (CNN) with a U-net architecture. In this, we succeeded; the model was successfully trained, validated, and tested on VIIRS infrared satellite data, and also tested on an additional MODIS dataset. Not only did HotLINK perform well on the test datasets, we found that it also outperformed an earlier automated volcanic hotspot detection approach, an optimized version of the MIROVA algorithm. We believe the basis of this improvement is in the algorithm we chose; by applying a CNN we were able to leverage better spatial pattern recognition and so improve automated detections of volcanic hotspots. To our knowledge, this was the first time these tools have been applied to the task of hotspot detection using data from VIIRS and MODIS satellite sensors.

Secondary goals of this project were to see what type of volcanic signals could be observed in HotLINK detections and if HotLINK could detect subtle warming signals that may be potential precursors to eruptions. In pursuit of these goals, we applied HotLINK to over 20 Terabytes of satellite data by processing 22 years of MODIS and 10 years of VIIRS data for eight target volcanoes in Alaska. In the processed time series data for these eight volcanoes, there were over 15 discrete eruption periods observed with activity ranging from lava flows (Mount Veniaminof, Shishaldin Volcano) to lava dome growth (Mount Cleveland, Redoubt Volcano, Augustine Volcano, Bogoslof Island), lahars (Redoubt Volcano, Augustine Volcano), ash emissions (observed at all volcanoes), and explosive events (observed at all volcanoes).

We found that HotLINK succeeds at detecting and tracking eruptive activity. In addition to detecting eruptive signals, there is some evidence of HotLINKs capability to detect lower temperature, non-eruptive thermal signals (see section 2.6.4). However, while pre-eruptive detections at Mount Veniaminof in 2018 are likely evidence of new thermal activity (i.e., increased surface warming or increased degassing), subtle detections in the summer of 2019 and 2020 could also have been residual heat from the 2018 eruption. Still, these detections demonstrate that HotLINK is able to detect subtle thermal signals which might be similar in nature to precursory thermal signals. Although we did not look in great detail at thermal activity prior to eruptions, our brief analysis did show many instances where HotLINK detected hotspots

prior to Aviation Color Code changes assigned by the Alaska Volcano Observatory (AVO). A more robust analysis of precursors would look at the timing and trend of hotspot detections relative to eruption onset, and comparison to high-resolution imagery and geophysical data to characterize the type of thermal feature and process. Unfortunately, the exact timing of eruption onset can sometimes be difficult to constrain for Alaska volcanoes due to their remote locations and often cloudy weather, and are not always well captured by AVO's color code changes (Cameron et al. 2018). In the future, HotLINK could be used to help identify eruption onset, once the radiative power and/or brightness temperature values of lava spatter, domes, or flows for the target volcanoes are better characterized.

Future work could also apply HotLINK with complementary spatial and temporal analysis to distinguish different volcanic features. For example, Kaneko et al. (2002) distinguished between exogenous and endogenous dome growth by their thermal signatures. One way this could be approached is through single image and temporal analysis of the radiative power (RP), brightness temperature (BT), and area of hotspots. For example, we would expect a lava flow to have a sudden onset of high RP and high BT over a large area, cooling slowly over time, whereas surface warming of a volcanic vent or geothermal area might have a slow but steady onset with overall lower RP and BT and cover a much smaller area.

Although for now HotLINK has only been applied retrospectively to historical data, it can easily be integrated into near-real-time operational analysis to aid with monitoring efforts at volcano observatories. To this end, we have published a tutorial and the code for the model and pre-processing pipeline on GitHub (Saunders-Shultz, 2023). Implementation in real time would require server space to conduct analysis and archive imagery, access real time VIIRS and MODIS data, an alarm system to output hotspot detections, and a database to store and analyze the time series of results.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., … Zheng, X. (2015). TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. https://www.tensorflow.org/

El Adoui, M., Mahmoudi, S. A., Larhmam, M. A., & Benjelloun, M. (2019). MRI breast tumor segmentation using different encoder and decoder CNN architectures. Computers, 8(3), 52.

Aggarwal, S. (2004). Principles of remote sensing. *Satellite remote sensing and GIS applications in agricultural meteorology*, *23*(2), 23-28.

Allaby, M. (Ed.). (2013). A dictionary of geology and earth sciences. Oxford University Press, USA.

Apple (2023) About Face ID advance technology. https://support.apple.com/en-us/102381 Published August 22, 2023. [Accessed November 22, 2023]

Baxter, P. J. (2005). Human impacts of volcanoes. *Volcanoes and the Environment*, 273-303.

Blackett, M. (2013). Review of the utility of infrared remote sensing for detecting and monitoring volcanic activity with the case study of shortwave infrared data for Lascar Volcano from 2001–2005. *Geological Society, London, Special Publications*, *380*, 107–135. https://doi.org/10.1144/SP380.10

Blackett, M. (2017). An overview of infrared remote sensing of volcanic activity. *Journal of Imaging*, *3*(2), 13. https://doi.org/10.3390/jimaging3020013

Bleick, H. A., Coombs, M. L., Cervelli, P. F., Bull, K. F., & Wessels, R. L. (2013). Volcano–ice interactions precursory to the 2009 eruption of Redoubt Volcano, Alaska. *Journal of Volcanology and Geothermal Research*, *259*, 373–388. https://doi.org/10.1016/j.jvolgeores.2012.10.008

Brown, S. K., Jenkins, S. F., Sparks, R. S. J., Odbert, H., & Auker, M. R. (2017). Volcanic fatalities database: analysis of volcanic threat with distance and victim classification. *Journal of Applied Volcanology*, *6*, 1-20. https://doi.org/10.1186/s13617-017-0067-4

Buslaev, A., Iglovikov, V. I., Khvedchenya, E., Parinov, A., Druzhinin, M., & Kalinin, A. A. (2020). Albumentations: Fast and Flexible Image Augmentations. *Information*, *11*(2), 125. https://doi.org/10.3390/info11020125

Cameron, C. E., Prejean, S. G., Coombs, M. L., Wallace, K. L., Power, J. A., & Roman, D. C. (2018). Alaska Volcano Observatory Alert and Forecasting Timeliness: 1989–2017. *Frontiers in Earth Science*, *6*, 1–16. https://doi.org/10.3389/feart.2018.00086

Cameron, C. E., & Staff, S. A. (2022). Geologic database of information on volcanoes in Alaska (GeoDIVA). https://doi.org/10.14509/geodiva. https://doi.org/10.14509/30901

Campus, A., Laiolo, M., Massimetti, F., & Coppola, D. (2022). The Transition from MODIS to VIIRS for Global Volcano Thermal Monitoring. *Sensors*, *22*, 1713. https://doi.org/10.3390/s22051713

Carter, A. J., Ramsey, M. S., & Belousov, A. B. (2007). Detection of a new summit crater on Bezymianny Volcano lava dome: Satellite and field-based thermal data. *Bulletin of Volcanology*, *69*(7), 811–815. https://doi.org/10.1007/s00445-007-0113-x

Cashman, K. V., & Giordano, G. (2008). Volcanoes and human history. *Journal of Volcanology and Geothermal Research*, *176*(3), 325-329. https://doi.org/10.1016/j.jvolgeores.2008.01.036

Cassidy, M., Sandberg, A., & Mani, L. (2023). The ethics of volcano geoengineering. *Earth's Future*, *11*(10), e2023EF003714. https://doi.org/10.1029/2023EF003714

Castaño, L. M., Ospina, C. A., Cadena, O. E., Galvis-Arenas, B., Londono, J. M., Laverde, C. A., Kaneko, T., & Ichihara, M. (2020). Continuous monitoring of the 2015–2018 Nevado del Ruiz activity, Colombia, using satellite infrared images and local infrasound records. *Earth, Planets and Space*, *72*, 81. https://doi.org/10.1186/s40623-020-01197-z

Chevrel, M. O., Villeneuve, N., Grandin, R., Froger, J. L., Coppola, D., Massimetti, F., ... & Peltier, A. (2023). Lava flow daily monitoring: the case of the 19 September–5 October 2022 eruption at Piton de la Fournaise. *Volcanica*, *6*(2), 391-404. https://doi.org/10.30909/vol.06.02.391404

Coombs, M. L., Wech, A. G., Haney, M. M., Lyons, J. J., Schneider, D. J., Schwaiger, H. F., Wallace, K. L., Fee, D., Freymueller, J. T., Schaefer, J. R., & Tepp, G. (2018). Short-Term Forecasting and Detection of Explosions During the 2016–2017 Eruption of Bogoslof Volcano, Alaska. *Frontiers in Earth Science*, *6*, 1–17. https://doi.org/10.3389/feart.2018.00122

Coppola, D., Piscopo, D., Laiolo, M., Cigolini, C., Delle Donne, D., & Ripepe, M. (2012). Radiative heat power at Stromboli volcano during 2000–2011: Twelve years of MODIS observations. *Journal of Volcanology and Geothermal Research*, *215*, 48-60. https://doi.org/10.1016/j.jvolgeores.2011.12.001

Coppola, D., Laiolo, M., Delle Donne, D., Ripepe, M., & Cigolini, C. (2014). Hot-spot detection and characterization of strombolian activity from MODIS infrared data. *International Journal of Remote Sensing*, *35*(9), 3403-3426. https://doi.org/10.1080/01431161.2014.903354

Coppola, D., Laiolo, M., Cigolini, C., Donne, D. D., & Ripepe, M. (2016). Enhanced volcanic hot-spot detection using MODIS IR data: results from the MIROVA system. *Geological Society, London, Special Publications*, *426*, 181–205. https://doi.org/10.1144/SP426.5

Coppola, D., Laiolo, M., Cigolini, C., Massimetti, F., Delle Donne, D., Ripepe, M., ... & William, R. (2020). Thermal remote sensing for global volcano monitoring: experiences from the MIROVA system. *Frontiers in Earth Science*, *7*, 362. https://doi.org/10.3389/feart.2019.00362

Coppola, D., Valade, S., Masias, P., Laiolo, M., Massimetti, F., Campus, A., ... & Valdivia, D. (2022). Shallow magma convection evidenced by excess degassing and thermal radiation during the dome-forming Sabancaya eruption (2012–2020). *Bulletin of Volcanology*, *84*(2), 16. https://doi.org/10.1007/s00445-022-01523-1

Coppola, D., Cardone, D., Laiolo, M., Aveni, S., Campus, A., & Massimetti, F. (2023). Global radiant flux from active volcanoes: the 2000–2019 MIROVA database. *Frontiers in Earth Science*, *11*.

Corradino, C., Ramsey, M. S., Pailot-Bonnetat, S., Harris, A. J. L., & Negro, C. Del. (2023). Detection of Subtle Thermal Anomalies: Deep Learning Applied to the ASTER Global Volcano Dataset. *IEEE Transactions on Geoscience and Remote Sensing*, *61*, 1–15. https://doi.org/10.1109/TGRS.2023.3241085

Crutzen, P. J. (2006). The "anthropocene". In *Earth system science in the anthropocene* (pp. 13-18). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-26590-2_3

Dehn, J., Dean, K., & Engle, K. (2000). Thermal monitoring of North Pacific volcanoes from space. *Geology*, *28*(8), 755-758. https://doi.org/10.1130/0091-7613(2000)28<755:TMONPV>2.0.CO;2

Dehn, J., Dean, K., Engle, K., & Izbekov, P. (2002). Thermal precursors in satellite images of the 1999 eruption of Shishaldin Volcano. *Bulletin of Volcanology*, *64*, 525–534. https://doi.org/10.1007/s00445-002-0227-0

Edmonds, M., & Woods, A. W. (2018). Exsolved volatiles in magma reservoirs. *Journal of Volcanology and Geothermal Research*, *368*, 13-30. https://doi.org/10.1016/j.jvolgeores.2018.10.018

Ernst, R. E. (2014). *Large igneous provinces*. Cambridge University Press.

Ferri, C., Hernández-Orallo, J., & Modroiu, R. (2009). An experimental comparison of performance measures for classification. *Pattern Recognition Letters*, *30*(1), 27–38. https://doi.org/10.1016/j.patrec.2008.08.010

Ganci, G., Vicari, A., Fortuna, L., & Del Negro, C. (2011). The HOTSAT volcano monitoring system based on combined use of SEVIRI and MODIS multispectral data. *Annals of Geophysics*, *54*, 544–550. https://doi.org/10.4401/ag-5338

Genzano, N., Pergola, N., & Marchese, F. (2020). A Google Earth Engine Tool to Investigate, Map and Monitor Volcanic Thermal Anomalies at Global Scale by Means of Mid-High Spatial Resolution Satellite Data. *Remote Sensing*, *12*, 3232. https://doi.org/10.3390/rs12193232

Girona, T., Realmuto, V., & Lundgren, P. (2021). Large-scale thermal unrest of volcanoes for years prior to eruption. Nature Geoscience, 14(4), 238-241.

Gouhier, M., Guéhenneux, Y., Labazuy, P., Cacault, P., Decriem, J., & Rivet, S. (2016). HOTVOLC: a web-based monitoring system for volcanic hot spots. *Geological Society, London, Special Publications*, *426*, 223–241. https://doi.org/10.1144/SP426.31

Guffanti, M., Diefenbach, A. K., Ewert, J. W., Ramsey, D. W., Cervelli, P. F., & Schilling, S. P. (2009). Volcano-monitoring instrumentation in the United States, 2008. *US Geological Survey Open-File Report, 1165*.

Guffanti, M., & Miller, T. P. (2013). A volcanic activity alert-level system for aviation: review of its development and application in Alaska. *Natural hazards*, *69*, 1519-1533. https://doi.org/10.1007/s11069-013-0761-4

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On Calibration of Modern Neural Networks. *Proceedings of the 34th International Conference on Machine Learning*, *PMLR*, 1321–1330. http://arxiv.org/abs/1706.04599

Hamill, T. M. (1997). Reliability Diagrams for Multicategory Probabilistic Forecasts. *Weather and Forecasting*, *12*(4), 736–741. https://doi.org/10.1175/1520-0434(1997)012<0736:RDFMPF>2.0.CO;2

Harris, A. J., & Stevenson, D. S. (1997). Thermal observations of degassing open conduits and fumaroles at Stromboli and Vulcano using remotely sensed data. *Journal of Volcanology and Geothermal Research*, *76*(3-4), 175-198. https://doi.org/10.1016/S0377-0273(96)00097-2

Harris, A. J., & Baloga, S. M. (2009). Lava discharge rates from satellite-measured heat flux. *Geophysical Research Letters*, *36*(19). https://doi.org/10.1029/2009GL039717

Harris, A. (2013). *Thermal remote sensing of active volcanoes: a user's manual*. Cambridge university press.

Harris, A. J., De Groeve, T., Garel, F., & Carn, S. A. (Eds.). (2016). Detecting, modelling and responding to effusive eruptions. Geological Society of London. https://www.lyellcollection.org/doi/10.1144/SP426.0

Harris, A. J., Villeneuve, N., Di Muro, A., Ferrazzini, V., Peltier, A., Coppola, D., ... & Arellano, S. (2017). Effusive crises at Piton de la Fournaise 2014–2015: a review of a multi-national response model. *Journal of Applied Volcanology*, *6*(1), 1-29. https://doi.org/10.1186/s13617-017-0062-9

Higgins, J., & Harris, A. (1997). Vast: A program to locate and analyse volcanic thermal anomalies automatically from remotely sensed data. *Computers and Geosciences*, *23*, 627–645. https://doi.org/10.1016/S0098-3004(97)00039-3

Hirn, B., Di Bartola, C., & Ferrucci, F. (2009). Combined use of SEVIRI and MODIS for detecting, measuring, and monitoring active lava flows at erupting volcanoes. *IEEE Transactions on Geoscience and Remote Sensing*, *47*(8), 2923-2930. https://doi.org/10.1109/TGRS.2009.2014224

Hurst, A. W., Bibby, H. M., Scott, B. J., & McGuinness, M. J. (1991). The heat source of Ruapehu crater lake; deductions from the energy and mass balances. *Journal of Volcanology and Geothermal Research*, *46*, 1–20. https://doi.org/10.1016/0377-0273(91)90072-8

Jeni, L. A., Cohn, J. F., & De La Torre, F. (2013). Facing Imbalanced Data--Recommendations for the Use of Performance Metrics. *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, 245–251. https://doi.org/10.1109/ACII.2013.47

Kaneko, T., Wooster, M. J., & Nakada, S. (2002). Exogenous and endogenous growth of the Unzen lava dome examined by satellite infrared image analysis. Journal of Volcanology and Geothermal Research, 116(1-2), 151-160. https://doi.org/10.1016/S0377-0273(02)00216-0

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. https://doi.org/10.48550/arXiv.1412.6980

Laiolo, M., Coppola, D., Barahona, F., Benítez, J. E., Cigolini, C., Escobar, D., ... & Finizola, A. (2017). Evidences of volcanic unrest on high-temperature fumaroles by satellite thermal monitoring: The case of Santa Ana volcano, El Salvador. *Journal of Volcanology and Geothermal Research*, *340*, 170-179. https://doi.org/10.1016/j.jvolgeores.2017.04.013

Layana, S., Aguilera, F., Rojo, G., Vergara, Á., Salazar, P., Quispe, J., ... & Urrutia, D. (2020). Volcanic Anomalies monitoring System (VOLCANOMS), a low-cost volcanic monitoring system based on Landsat images. *Remote Sensing*, *12*(10), 1589. https://doi.org/10.3390/rs12101589

LeCun, Y., Kavukcuoglu, K., & Farabet, C. (2010, May). Convolutional networks and applications in vision. In *Proceedings of 2010 IEEE international symposium on circuits and systems* (pp. 253-256). IEEE. https://doi.org/10.1109/ISCAS.2010.5537907.

Li, F., Zhang, X., Kondragunta, S., & Csiszar, I. (2018). Comparison of Fire Radiative Power Estimates From VIIRS and MODIS Observations. *Journal of Geophysical Research: Atmospheres*, *123*, 4545–4563. https://doi.org/10.1029/2017JD027823

Loewen, M. W., Dietterich, H. R., Graham, N., & Izbekov, P. (2021). Evolution in eruptive style of the 2018 eruption of Veniaminof volcano, Alaska, reflected in groundmass textures and remote sensing. *Bulletin of Volcanology*, *83*(11), 72. https://doi.org/10.1007/s00445-021-01489-6

Lombardo, V. (2016). AVHotRR: near-real time routine for volcano monitoring using IR satellite data. *Geological Society, London, Special Publications*, *426*, 73–92. https://doi.org/10.1144/SP426.18

Loughlin, S. C., Vye-Brown, C., Sparks, R. S. J., Brown, S. K., Barclay, J., Calder, E., Cottrell, E., Jolly, G., Komorowski, J. C., Mandeville, C., Newhall, C., Palma, J., Potter, S., & Valentine, G. (2015). An introduction to global volcanic hazard and risk. In *Global Volcanic Hazards and Risk* (Issue 2015). https://doi.org/10.1017/CBO9781316276273.003

Massimetti, F., Coppola, D., Laiolo, M., Valade, S., Cigolini, C., & Ripepe, M. (2020). Volcanic hot-spot detection using SENTINEL-2: a comparison with MODIS–MIROVA thermal data series. *Remote Sensing*, *12*(5), 820. https://doi.org/10.3390/rs12050820

Mazzeo, G., Ramsey, M. S., Marchese, F., Genzano, N., & Pergola, N. (2021). Implementation of the NHI (Normalized hot spot indices) algorithm on infrared aster data: Results and future perspectives. *Sensors*, *21*, 1–16. https://doi.org/10.3390/s21041538

Moran, S. C., Newhall, C., & Roman, D. C. (2011). Failed magmatic eruptions: Late-stage cessation of magma ascent. *Bulletin of Volcanology*, *73*, 115–122. https://doi.org/10.1007/s00445-010-0444-x

Murphy, S. W., de Souza Filho, C. R., Wright, R., Sabatino, G., & Correa Pabon, R. (2016). HOTMAP: Global hot target detection at moderate spatial resolution. *Remote Sensing of Environment*, *177*, 78–88. https://doi.org/10.1016/j.rse.2016.02.027

NASA Goddard Space Flight Center. (2018). NASA Visible Infrared Imaging Radiometer Suite Level-1B Product User Guide (Version 1.0). https://landweb.modaps.eosdis.nasa.gov/NPP/forPage/NPPguide/NASAVIIRSL1BUGMay2018.pdf

Oppenheimer, C., Rothery, D. A., & Francis, P. W. (1993). Thermal distributions at fumarole fields: implications for infrared remote sensing of active volcanoes. *Journal of Volcanology and Geothermal research*, *55*(1-2), 97-115. https://doi.org/10.1016/0377-0273(93)90092-6

Orr, T. R., Cameron, C. E., Dietterich, H. R., Dixon, J. P., Enders, M. L., Grapenthin, R., ... & Wech, A. G. (2023). *2019 Volcanic activity in Alaska—Summary of events and response of the Alaska Volcano Observatory* (No. 2023-5039). US Geological Survey. https://doi.org/10.3133/sir20235039

Pergola, N., Coviello, I., Filizzola, C., Lacava, T., Marchese, F., Paciello, R., & Tramutoli, V. (2016). A review of RSTVOLC, an original algorithm for automatic detection and near-real-time monitoring of volcanic hotspots from space. *Geological Society Special Publication*, *426*, 55–72. https://doi.org/10.1144/SP426.1

Pergola, N., Marchese, F., & Tramutoli, V. (2004). Automated detection of thermal features of active volcanoes by means of infrared AVHRR records. *Remote Sensing of Environment*, *93*, 311–327. https://doi.org/10.1016/j.rse.2004.07.010

Philpotts, A. R., & Ague, J. J. (2022). Principles of igneous and metamorphic petrology. Cambridge University Press.

Pieri, D., & Abrams, M. (2005). ASTER observations of thermal anomalies preceding the April 2003 eruption of Chikurachki volcano, Kurile Islands, Russia. *Remote Sensing of Environment*, *99*, 84–94. https://doi.org/10.1016/j.rse.2005.06.012

Pipolo, S., Salanne, M., Ferlat, G., Klotz, S., Saitta, A. M., & Pietrucci, F. (2017). Navigating at will on the water phase diagram. *Physical review letters*, *119*(24), 245701. https://doi.org/10.1103/PhysRevLett.119.245701

Planck, M. (1914). *The theory of heat radiation*. Blakiston.

Poland, M. P., & Anderson, K. R. (2020). Partly cloudy with a chance of lava flows: Forecasting volcanic eruptions in the twenty-first century. *Journal of Geophysical Research: Solid Earth*, *125*(1), e2018JB016974. https://doi.org/10.1029/2018JB016974

Pritchard, M. E., Poland, M., Reath, K., Andrews, B., Bagnardi, M., Biggs, J., ... & Roman, A. (2022). Optimizing satellite resources for the global assessment and mitigation of volcanic hazards—Suggestions from the USGS Powell Center Volcano Remote Sensing Working Group (No. 2022-5116). US Geological Survey.

Racki, G. (2020). A volcanic scenario for the Frasnian–Famennian major biotic crisis and other Late Devonian global changes: More answers than questions?. *Global and Planetary Change*, *189*, 103174. https://doi.org/10.1016/j.gloplacha.2020.103174

Ramsey, M. S., Wessels, R. L., & Anderson, S. W. (2012). Surface textures and dynamics of the 2005 lava dome at Shiveluch Volcano, Kamchatka. *Bulletin*, *124*(5-6), 678-689. https://doi.org/10.1130/B30580.1

Ramsey, M. S., Corradino, C., Thompson, J. O., & Leggett, T. N. (2023). Statistical retrieval of volcanic activity in long time series orbital data: Implications for forecasting future activity. *Remote Sensing of Environment*, *295*, 113704. https://doi.org/10.1016/j.rse.2023.113704

Reath, K. A., Ramsey, M. S., Dehn, J., & Webley, P. W. (2016). Predicting eruptions from precursory activity using remote sensing data hybridization. *Journal of Volcanology and Geothermal Research*, *321*, 18–30. https://doi.org/10.1016/j.jvolgeores.2016.04.027

Robock, A. (2000). Volcanic eruptions and climate. *Reviews of geophysics*, *38*(2), 191-219. https://doi.org/10.1029/1998RG000054

Rogers, J. J. (1996). A history of continents in the past three billion years. *The journal of geology*, *104*(1), 91-107. https://doi.org/10.1086/629803

Rogic, N., Cappello, A., & Ferrucci, F. (2019). Role of emissivity in lava flow 'Distance-to-Run'estimates from satellite-based volcano monitoring. *Remote Sensing*, *11*(6), 662. https://doi.org/10.3390/rs11060662

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. *IEEE Access*, *9*, 16591–16603. https://doi.org/10.1109/ACCESS.2021.3053408

Rouwet, D., Tassi, F., Mora-Amador, R., Sandri, L., & Chiarini, V. (2014). Past, present and future of volcanic lake monitoring. *Journal of Volcanology and Geothermal Research*, *272*, 78–97. https://doi.org/10.1016/j.jvolgeores.2013.12.009

Saunders-Shultz, P. (2023) Hotspot Learning and Identification Network (HotLINK) [Computer software]. https://github.com/csaundersshultz/HotLINK

Segall, P. (2013). Volcano deformation and eruption forecasting. *Geological Society, London, Special Publications*, *380*(1), 85-106. https://doi.org/10.1144/SP380.4

Sola, J., & Sevilla, J. (1997). Importance of input data normalization for the application of neural networks to complex industrial problems. *IEEE Transactions on Nuclear Science*, *44*(3), 1464–1468. https://doi.org/10.1109/23.589532

Sparks, R. S. J. (2003). Forecasting volcanic eruptions. *Earth and Planetary Science Letters*, *210*(1-2), 1-15. https://doi.org/10.1016/S0012-821X(03)00124-9

Tilling, R. I. (2008). The critical role of volcano monitoring in risk reduction. *Advances in Geosciences*, *14*, 3-11. https://doi.org/10.5194/adgeo-14-3-2008

Valade, S., Ley, A., Massimetti, F., D'Hondt, O., Laiolo, M., Coppola, D., Loibl, D., Hellwich, O., & Walter, T. R. (2019). Towards global volcano monitoring using multisensor sentinel missions and artificial intelligence: The MOUNTS monitoring system. Remote Sensing, 11, 1–31. https://doi.org/10.3390/rs11131528

Wan, J., Wang, D., Hoi, S. C. H., Wu, P., Zhu, J., Zhang, Y., & Li, J. (2014, November). Deep learning for content-based image retrieval: A comprehensive study. In Proceedings of the 22nd ACM international conference on Multimedia (pp. 157-166). https://doi.org/10.1145/2647868.2654948

Waythomas, C. (2021). Simultaneous effusive and explosive cinder cone eruptions at Veniaminof Volcano, Alaska. *Volcanica*, *4*(2), 295-307. https://doi.org/10.30909/vol.04.02.295307

Waythomas, C. F., Edwards, B. R., Miller, T. P., & McGimsey, R. G. (2023). Lava-ice interactions during historical eruptions of Veniaminof Volcano, Alaska and the potential for meltwater floods and lahars. *Natural Hazards*, *115*(1), 73-106. https://doi.org/10.1007/s11069-022-05523-4

Werner, C., Kern, C., Coppola, D., Lyons, J. J., Kelly, P. J., Wallace, K. L., Schneider, D. J., & Wessels, R. L. (2017). Magmatic degassing, lava dome extrusion, and explosions from Mount Cleveland volcano, Alaska, 2011–2015: Insight into the continuous nature of volcanic activity over multi-year timescales. *Journal of Volcanology and Geothermal Research*, *337*, 98-110. https://doi.org/10.1016/j.jvolgeores.2017.03.001

Wood, H. O. (1913). The Hawaiian Volcano Observatory. *Bulletin of the Seismological Society of America*, *3*(1), 14-19. https://doi.org/10.1785/BSSA0030010014

Wooster, M. (2003). Fire radiative energy for quantitative study of biomass burning: derivation from the BIRD experimental satellite and comparison to MODIS fire products. *Remote Sensing of Environment*, *86*(1), 83–107. https://doi.org/10.1016/S0034-4257(03)00070-1

Wright, R., Flynn, L. P., Garbeil, H., Harris, A. J. L., & Pilger, E. (2004). MODVOLC: Near-real-time thermal monitoring of global volcanism. *Journal of Volcanology and Geothermal Research*, *135*, 29–49. https://doi.org/10.1016/j.jvolgeores.2003.12.008

Wright, R. (2016). MODVOLC: 14 years of autonomous observations of effusive volcanism from space. *Geological Society Special Publication*, *426*, 23–53. https://doi.org/10.1144/SP426.12

Appendix

A. U-net code

Python code to generate the U-net model architecture used in the final application. Note that this code just describes the architecture of the network prior to any training, and does not have the weights of the final trained model. The final trained model can be accessed via https://github.com/csaundersshultz/HotLINK

```
#U_NET ARCHITECTURE
#imports
from tensorflow.keras import layers

#set up input image shape (64x64) and number of channels (2).
img_input = layers.Input(shape=(64, 64, 2)) #2 channels MIR and TIR inputs
kern='glorot_normal'
pad = 'valid'
act = 'relu'
kern_reg = None #None used in final model
kern_con =  None #None used in final model
dropout=0.05
#DOWN BLOCK
x = layers.Conv2D(filters=16, kernel_size=3, activation=act, kernel_initializer=kern, padding=pad,
kernel_constraint=kern_con)(img_input)
x = layers.Dropout(dropout)(x)
x = layers.Conv2D(filters=16, kernel_size=3, activation=act, kernel_initializer=kern, padding=pad, kernel_constraint=kern_con)(x)
resid1 = x #shape 60
x = layers.MaxPooling2D(pool_size=(2, 2), strides=2)(x)
x = layers.Conv2D(filters=32, kernel_size=3, activation=act, kernel_initializer=kern, padding=pad, kernel_constraint=kern_con)(x)
x = layers.Dropout(dropout)(x)
x = layers.Conv2D(filters=32, kernel_size=3, activation=act, kernel_initializer=kern, padding=pad, kernel_constraint=kern_con)(x)
resid2 = x #shape 26
x = layers.MaxPooling2D(pool_size=(2, 2), strides=2)(x)
x = layers.Conv2D(filters=48, kernel_size=3, activation=act, kernel_initializer=kern, padding=pad, kernel_constraint=kern_con)(x)
x = layers.Dropout(dropout)(x)
x = layers.Conv2D(filters=48, kernel_size=3, activation=act, kernel_initializer=kern, padding=pad, kernel_constraint=kern_con)(x)

#UP BLOCK
x = tf.image.resize(x, [18,18] , method='nearest')
resid2 = layers.Cropping2D(cropping=4)(resid2) #crop to 18 from 26
x = layers.Concatenate(axis=3)([x, resid2])
x = layers.Conv2D(filters=32, kernel_size=3, activation=act, kernel_initializer=kern, padding=pad, kernel_constraint=kern_con)(x)
x = layers.Dropout(dropout)(x)
x = layers.Conv2D(filters=32, kernel_size=3, activation=act, kernel_initializer=kern, padding=pad, kernel_constraint=kern_con)(x)
x = tf.image.resize(x, [28,28] , method='nearest')
resid1 = layers.Cropping2D(cropping=16)(resid1) #crop to 36 from 60 #crop to 28
x = layers.Concatenate(axis=3)([x, resid1])
x = layers.Conv2D(filters=16, kernel_size=3, activation=act, kernel_initializer=kern, padding=pad, kernel_constraint=kern_con)(x)
x = layers.Dropout(dropout)(x)
x = layers.Conv2D(filters=16, kernel_size=3, activation=act, kernel_initializer=kern, padding=pad, kernel_constraint=kern_con)(x)

#OUTPUT LAYER (note softmax activation)
output = layers.Conv2D(filters=3, kernel_size=1, activation="softmax", padding="valid", kernel_initializer=kern)(x) #3 output
channels

model = Model(img_input, output)
```

B. Image Augmentation Validation

The table below shows the different results on the VIIRS validation dataset when a U-net model is trained without any augmentations, and when using random 90° rotations and vertical/horizontal flips. All other variables and hyperparameters remain the same. The model with image augmentations is the final HotLINK model.

Model	Accuracy	F1-score	True Negative rate	False Positive rate
			False Negative rate	True Positive rate
U-net with no augmentations	0.95	0.89	0.725	0.022
			0.032	0.221
U-net with random 90° rotations and vertical/horizontal flips	0.96	0.93	0.730	0.017
			0.021	0.232

Table A.1: Results of model training with and without image augmentations.

C. Optimizing Hysteresis thresholds

Hysteresis thresholding is a two step process used to identify pixels of interest, in this case hotspot pixels. First, all pixels with probabilities greater than the high threshold are considered hotspots. Next, the hotspots are expanded to all nearby pixels whose probabilities exceed the low threshold. Only the high threshold determines which images contain hotspots, and the low threshold is more important for determining which pixels within the images are flagged as hotspots. For this reason, we optimize the high hysteresis threshold to image F1-score, and the low hysteresis threshold to pixel F1-score. This is done over the VIIRS validation dataset.

The maximum image F1-score is achieved at H=0.53, while maximum accuracy is achieved at multiple values ranging from 0.53-0.65 (Supplementary Figure 1). Horizontal lines show argmax thresholding, which chooses the highest probability from the 3 classes (background, hotspot, hotspot-adjacent), so it occasionally allows pixels to be active with

probabilities <0.5. During prediction we used a high threshold of 0.50. This is done partly for simplicity, but also because the small size of the VIIRS validation dataset limits the confidence that 0.53 is the true optimum threshold value.

The low threshold is set using pixel-wise F1-score (Supplementary Figure 2). The optimum is found to be 0.4 with clear drop-offs at higher and lower thresholds. The pixel-wise accuracy seems to favor higher thresholds ~>0.44. This demonstrates that a threshold of 0.4 has less true negatives (more false positives) and slightly lower accuracy overall than higher thresholds, but with more true positive predictions as well. During prediction we use a low threshold of 0.4 since it optimizes the F1-score. This optimization has many more samples than the image-wise F1-score, as each image prediction contains 24x24=576 pixels, for a total dataset of 734,400 pixels.

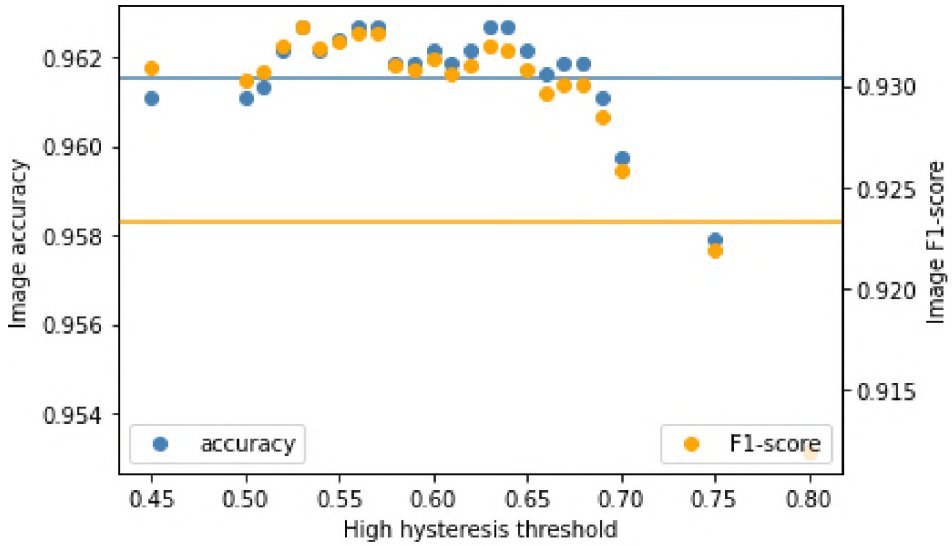

Figure A.1: Optimizing for the high hysteresis threshold.

Figure A.1 shows the process of optimizing for the high hysteresis threshold. The maximum image F1-score is achieved at H=0.53, while maximum accuracy is achieved at multiple values ranging from 0.53 – 0.65. Horizontal lines show argmax thresholding, which chooses the highest probability from the 3 classes (background, hotspot, hotspot-adjacent), so it occasionally allows pixels to be active with probabilities <0.5. During prediction we used a high threshold of 0.50. This is done partly for simplicity, but also because the small size of the VIIRS validation dataset limits the confidence that 0.53 is the true optimum threshold value.

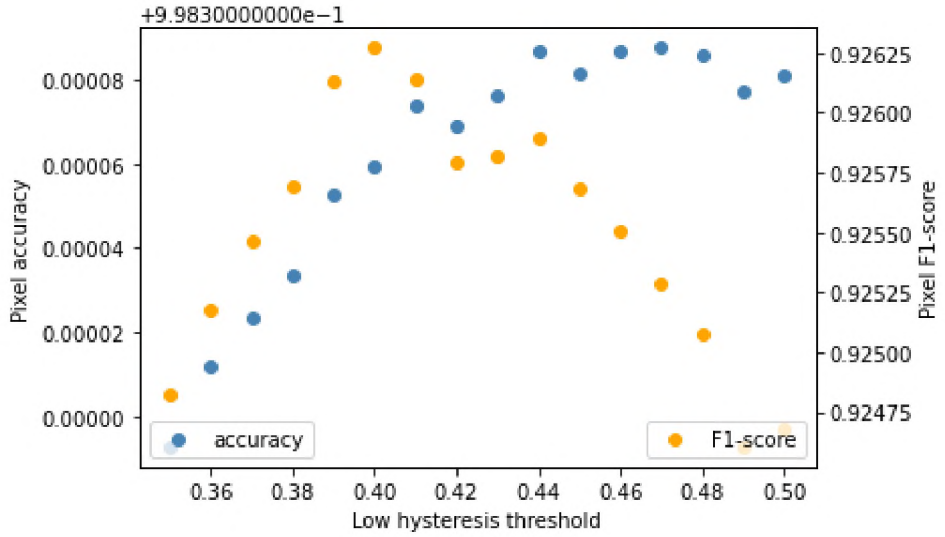

Figure A.2: Optimizing for the low hysteresis threshold.

D. Optimizing MIROVA thresholds

MIROVA thresholds are optimized over the VIIRS training dataset. First the dataset is split into night and daytime imagery. Then, the maximum value of indices: NTI, dNTI, dETI, dNTI Z-Score, and dETI Z-Score are saved for each image. This simplifying assumption saves computational resources, since instead of having to evaluate MIROVA over an entire image we can just use the maximum values of indices from each image. Next a grid search of C1 and C2 thresholds are tried for each sub dataset. The C1 and C2 values which maximize accuracy (minimize error rate) are found for both night and daytime images (Supplementary Figures 3 and 4). The complete MIROVA algorithm including the K threshold is used during the grid search process, but K is set to the default values (K=-0.8 for nighttime data, and -0.6 for daytime data) specified in Coppola et al. (2016). Changing the K value did not result in significantly different results, so it was left as is during the optimization process of the other two thresholds.

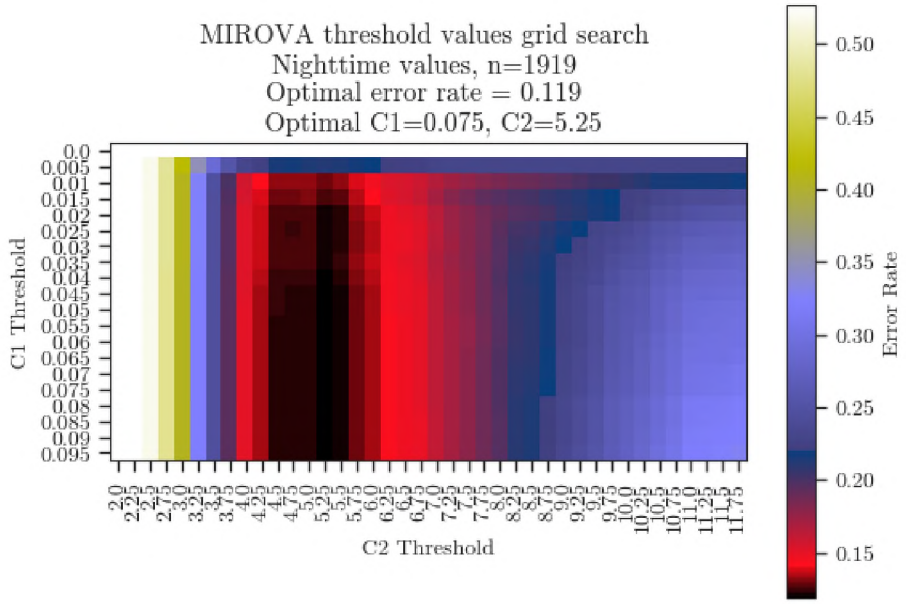

Figure A.3: Grid search for nighttime MIROVA thresholds C1 and C2.

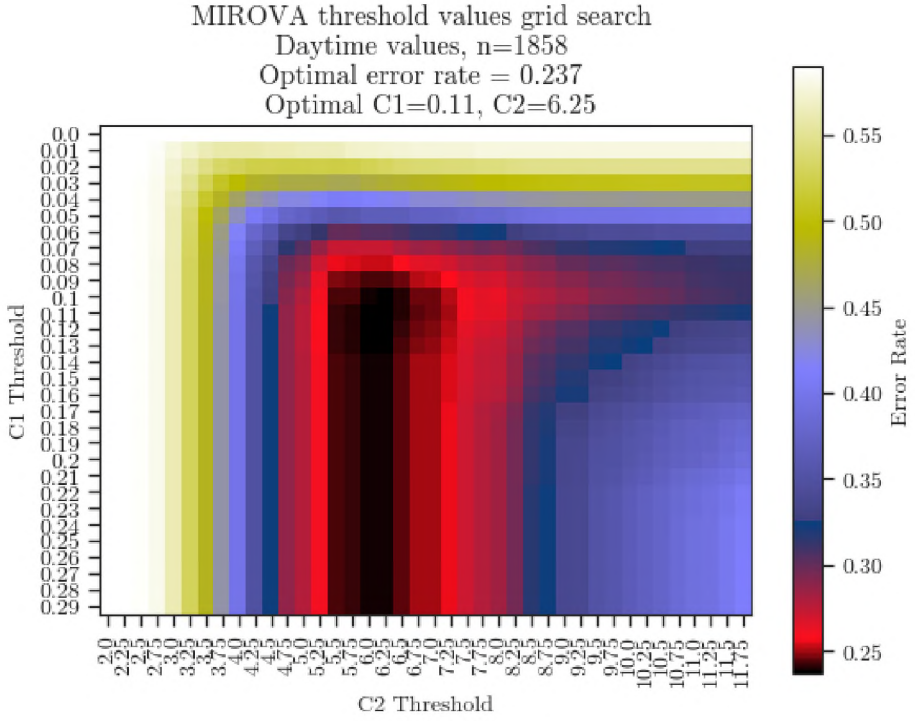

Figure A.4: Grid search for daytime MIROVA thresholds C1 and C2.

E. Additional HotLINK detection examples

Additional examples are provided showing hotspot detections at various volcanoes in VIIRS MIR images (band I04).

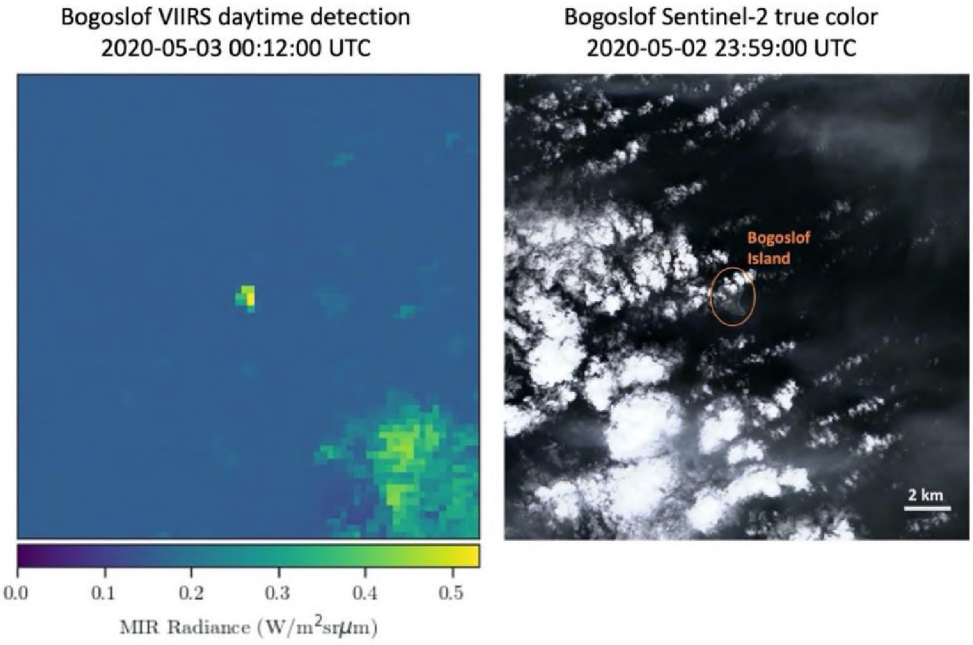

Figure A.5: A Hotlink detection at Bogoslof Island from 2020-05-03. Sentinel-2 true color imagery captured minutes earlier shows no evidence of volcanic activity. Comparison of the two images shows that the size of the hotspot in the MIR imagery is roughly the same as the entire island.

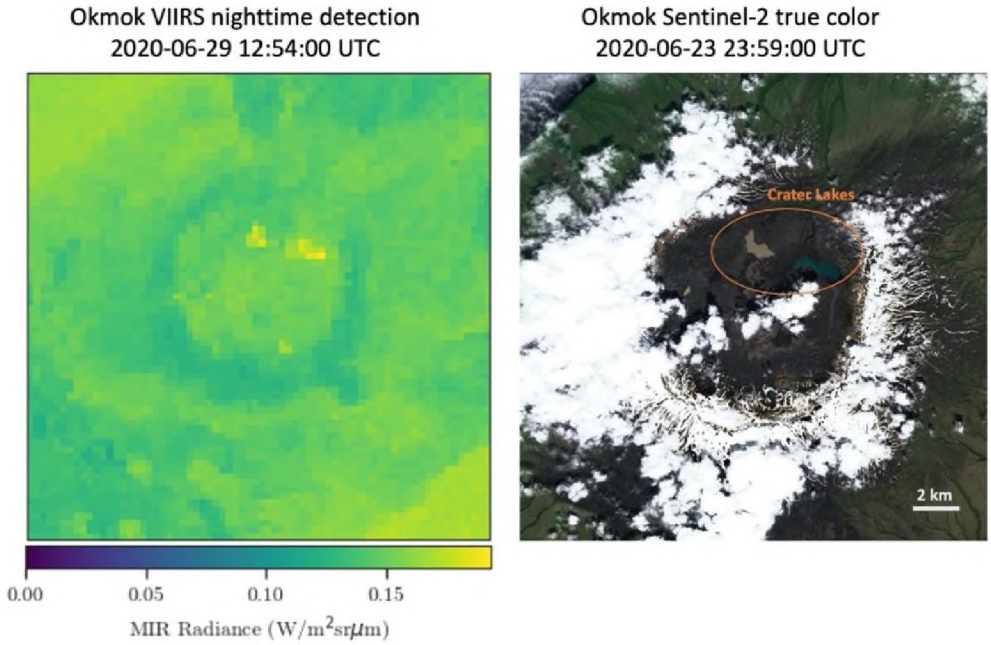

Figure A.6: A Hotlink detection at Okmok Caldera from 2020-06-29. Sentinel-2 true color imagery a few days earlier (closest clear sky conditions) shows no evidence of volcanic activity. Comparison of the two images shows that the two bright spots in the MIR image are coincident with crater lakes inside the Okmok Caldera.