Evaluating Terrain-Dependent Performance for Martian Frost Detection in Visible Satellite Observations

Gary Doran
Serina Diniega
Steven Lu
Mark Wronkiewicz
Kiri L. Wagstaff
gary.b.doran.jr@jpl.nasa.gov
Jet Propulsion Laboratory,
California Institute of Technology
Pasadena, California, USA

Jacob Widmer
University of California, Los Angeles
Los Angeles, California, USA

ABSTRACT

Seasonal frosting and defrosting on the surface of Mars is hypothesized to drive both climate processes and the formation and evolution of geomorphological features such as gullies. Past studies have focused on manually analyzing the behavior of the frost cycle in the northern mid-latitude region of Mars using high-resolution visible observations from orbit. Extending these studies globally requires automating the detection of frost using data science techniques such as convolutional neural networks. However, visible indications of frost presence can vary significantly depending on the geologic context on which the frost is superimposed. In this study, we (1) present a novel approach for spatially partitioning data to reduce biases in model performance estimation, (2) illustrate how geologic context affects automated frost detection, and (3) propose future work to further mitigate observed biases in automated frost detection work.

KEYWORDS

planetary science, remote sensing, deep learning

ACM Reference Format:
Gary Doran, Serina Diniega, Steven Lu, Mark Wronkiewicz, Kiri L. Wagstaff, and Jacob Widmer. 2022. Evaluating Terrain-Dependent Performance for Martian Frost Detection in Visible Satellite Observations. In *Proceedings of 3rd ACM SIGKDD Workshop on Deep Learning for Spatiotemporal Data (SIGKDD DeepSpatial '22).* ACM, New York, NY, USA, 5 pages.

1 INTRODUCTION

As on Earth, frost will accumulate on the Martian surface from the poles towards the equator each winter. This frost is an important driver for surface geological and climate processes [2] and provides a key observable constraint for studies of how volatiles are transported around Mars in the present climate [3]. Unlike the Earth, the atmosphere of Mars is comprised primarily of carbon dioxide (CO_2) and this volatile constitutes most of the frost, falling as snow or condensing at the surface due to surface temperatures falling to the CO_2 frost point. A small amount of water frost will also form when temperatures are below the water (H_2O) frost point, but only if the local concentration of H_2O vapor in the atmosphere is high enough. A global, high-resolution map of where and when specific types of frost form around Mars would be a great aid towards generation of a comprehensive view of the Martian global frost cycle and an important input for many studies focused on understanding Mars' atmospheric dynamics, volatile budget, landscape and landform evolution, and surface operations of robotic and human explorers.

Confident detection of Martian frost and characterization of its type (i.e., H_2O or CO_2, snowfall or surface condensate) requires the combination of information across multiple remote sensing instruments, including high-resolution visible imaging systems. Over 100 TB of high resolution surface image data has been returned from Mars, making it infeasible to manually analyze these images for the presence of frost. Therefore, we have trained a convolutional neural network (CNN) model to detect frost using a corpus of labeled data from previously studied frost sites. The model can be deployed across the entire image dataset to automatically detect frost and enable global-scale scientific analysis of the frost cycle.

This paper describes our initial efforts to train and validate a Martian frost detection model for visible images. We describe some domain-specific challenges and approaches related to label collection, validation, and bias characterization. We find that detection recall is biased against certain underrepresented terrain types such as dunes, and we propose future work to mitigate this bias.

2 BACKGROUND

The scientific community is interested in better understanding the Martian global frost cycle, its effects on surface evolution, and its role in the larger Martian climate system. The extent of the contiguous seasonal frost cap and broad presence of frost has been mapped using low-resolution (0.1–6 km/pixel) thermal instruments [11], while frost within specific small areas (including areas with patchier frost) has been investigated using medium-resolution (18 m/pixel) spectral instruments [12] and high-resolution (25–50 cm/pixel) visible instruments [5]. Due to the large volumes of data, extending the latter type of focused site studies globally, so as to combine that view with the global, low-resolution results, requires an automated approach. In this paper, we focus on training and evaluating a CNN model for frost detection within observations acquired using the High Resolution Imaging Science Experiment (HiRISE) instrument [10], which provides visible band, high-resolution surface images.

SIGKDD DeepSpatial '22, August 15, 2022, Washington, DC
2022.

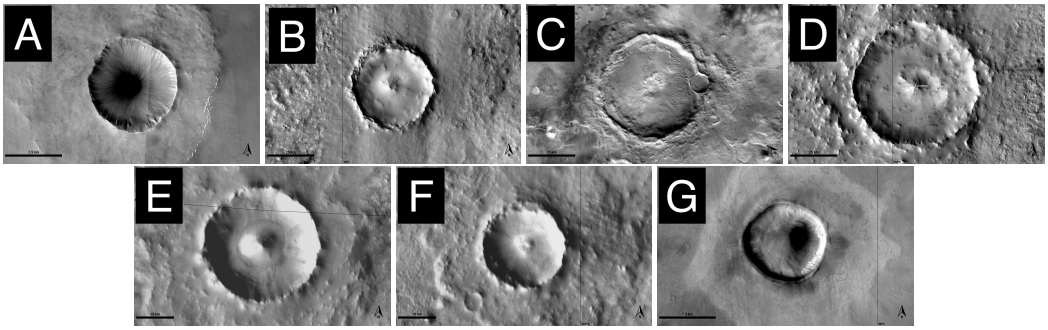

Figure 1: Previously studied northern mid-latitude frost sites used for training: A: 64.550°N, 315.907°E, **B:** 58.236°N, 89.607°E, **C:** 63.738°N, 11.035°E, **D:** 42.572°N, 67.332°E, **E:** 56.847°N, 350.401°E, **F:** 59.839°N, 135.999°E, **G:** 64.829°N, 209.406°E.

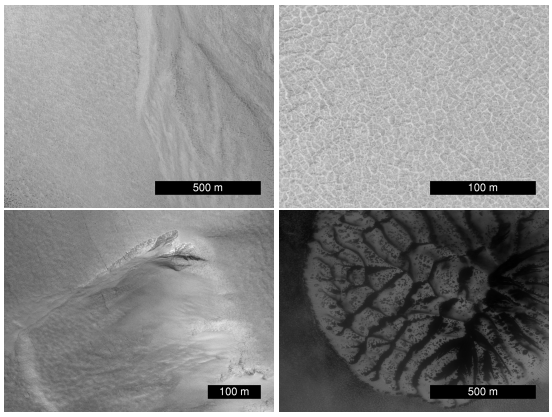

Figure 2: Visible indications of frost, including uniform albedo (top left), polygonal features (top right), halos (bottom left), and defrosting marks on dunes (bottom right).

In order to train a CNN frost detection model, we used HiRISE observations collected for a previous frost study in the northern mid-latitude region of Mars [15]. The 7 sites we focused on (Figure 1) are impact craters containing dark-colored basalt dune fields on which frost is visually apparent in the winter. The presence of frost on dunes indicates that regional conditions are favorable for frost formation, which helps to disambiguate whether frost may be present on nearby terrains. This aspect of the manual frost detection methodology highlights a key challenge for traditional machine learning models: confident frost detection often requires the use of larger scale contextual information not available to CNN models using only local image information. To address this challenge, we focused on detecting *visible indications* of frost, which include a uniform bright albedo, polygonal features, halos, and defrosting marks (see Figure 2) but excludes frost that is only detectable at other wavelengths [8]. This mirrors as closely as possible the manual frost detection problem, excluding the final step of assimilating information across scales and imaging modalities.

Our dataset consists of repeated observations of the same locations, which introduces another challenge when training models; we must account for the fact that the same locations are observed at different times throughout the seasonal cycle. In addition, because the data are highly clustered into discrete sites, it is necessary to account for these correlations during the validation process to prevent "data leakage" across the training, validation, and testing sets [6, 14]. Below, we describe a novel spatial partitioning to address this challenge. One benefit of the overlapping images is that observations during summer months can be used to provide frost-free (negative) training examples, whereas observations during winter months provide candidate frost (positive) examples.

3 METHODOLOGY

In this section, we provide an overview of our methodology from data generation to evaluation, with a focus on aspects specific to our dataset and problem domain.

3.1 Creating a Machine Learning Ready Dataset

Starting from HiRISE observations of the northern mid-latitude sites shown in Figure 1, we produced a set of labeled image tiles for the "frost" (positive) and "background" (negative) classes. Since determining frost composition is not straightforward using visible image data alone, we did not attempt to differentiate CO_2 and H_2O frost. HiRISE observations are typically around 10,000 pixels across, and of variable length depending on along-track exposure during "push-broom" imaging. We used the map-projected HiRISE products available from the Planetary Data System (PDS), which are between 25–100 cm/pixel resolution [9]. Because entire observations are too large for most labeling tools, we break each observation into a "subframe" that is 5,120 pixels along each dimension (except for partial subframes remaining near observation edges). Any subframe containing more than 75% of pixels outside the valid map-projected data area were discarded. We randomly selected 15 subframes containing frost identified from previous studies and 15 without frost from each of the 7 sites, for a total of 210 subframes. Only the 105 frosted subframes required more detailed labeling.

We used the Labelbox[1] platform to annotate polygonal boundaries around regions with visible evidence of frost. For each polygon, we collected additional information from the labeler including the applicable visible indicators as well as geologic context, which is either "dunes," "gullies," "crater rim/wall," or "other." Here, the geologic context categories are mutually exclusive, so labelers could

[1] https://labelbox.com/

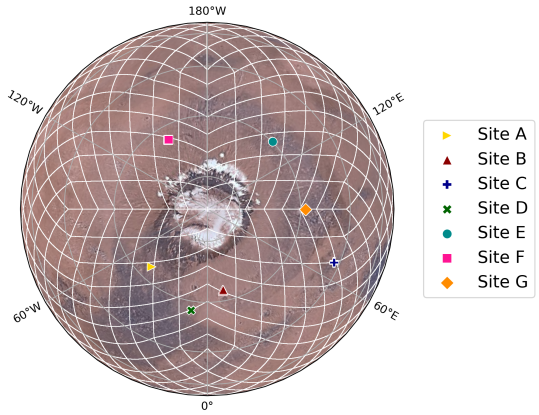

Figure 3: Site locations within the HEALPix partition (white grid) of the surface (North pole orthographic projection).

only pick one geologic context per frost polygon. We used the geologic context information to investigate terrain-dependent bias in classifier performance. To document the labeling process, we performed an iterative series of labeling sessions with both data science and science domain experts. Domain expert labeling guidance and clarifications at each iteration were captured in a labeling guide, included with the publicly available dataset[2]. A total of 6 subframes, detailed in the released dataset, were excluded due to contamination with excess instrument noise or cloud cover. Each subframe was labeled by three different annotators.

Finally, we split each subframe into 299×299 pixel tiles to generate a labeled dataset for training and evaluation. For each tile, the class label was determined through majority vote by comparing the set of overlapping polygons across the three annotators. If the number of polygons overlapping a tile is fewer than the required majority threshold, it was discarded to avoid ambiguous examples. The tiling process also discards any frost tiles that contain more than 10% black pixels, which would indicate that they fall partially outside the valid map-projected image data.

3.2 Spatial Validation

We used a standard train, validation, and test split methodology to evaluate the classifier both during and after training. Because the tiles derived from HiRISE observations are spatially clustered, it is important to split the data in such a way that does not mix tiles from the same repeatedly imaged region across train, validation, and test folds. Here, we present a novel application of Hierarchical Equal Area isoLatitude Pixelation (HEALPix) to spatially divide the globe into equal-area regions for model validation [4]. The advantages of HEALPix over other partitioning approaches is that it can be parameterized to flexibly subdivide the surface to arbitrary granularity ($N_{\text{side}} = 8$ for this application, meaning that each base-resolution pixel is divided into 8 along each side). The equal-area nature of the pixelization also ensures that no regions are disproportionately represented. The pixelization used for our study, along with the location of the northern mid-latitude sites, is shown in Figure 3.

[2]http://doi.org/10.5281/zenodo.6561241

Context	Other	Crater Rim/Wall	Gully	Dune
Train	83.1%	10.0%	2.6%	4.3%
Test	98.3%	—	—	1.7%

Table 1: Distribution of Geologic Context in Train/Test Sets

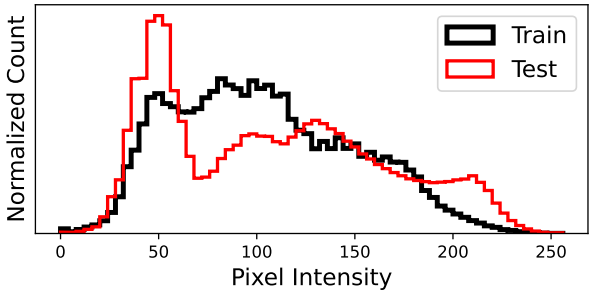

Figure 4: Comparison of pixel intensity distributions across all train and test tile pixels.

3.3 Model Training and Evaluation

We fine-tuned the InceptionV3 model [13] for frost detection by adjusting the weights in the final fully-connected layer using TensorFlow [1]. The learning rate was fixed to 10^{-3}, and the batch size was set to 1. Training was performed for 100 epochs using the Adam optimizer [7] and cross-entropy loss, and the model with the best validation set accuracy was selected. We used classification accuracy to evaluate overall model performance on the training, validation, and test sets. To understand how geologic context affects detection for the positive frost class, we evaluated performance on the frost tiles using recall separately for each context.

4 RESULTS

The full labeled dataset contains 23,767 tiles, nearly balanced with 12,657 frost tiles and 11,110 background tiles (53.3% frost tiles). During tile generation, 6471 potential frost tiles were excluded due to label ambiguity. Due to the HEALPix-based spatial partitioning, there is some unevenness in splitting the data, so the training, validation, and test sets comprise 65.3%, 14.3%, and 20.4% of the data set, respectively. Total labeling time across all three annotators was 11.5 hours, which corresponds to 1.7 seconds of labeling effort per tile produced. The validation set accuracy was maximized on the 97^{th} training epoch with a value of 99.4%. The selected model also has a training set accuracy of 99.4% and a test set accuracy of 92.0%. We hypothesize that the drop in accuracy on the test set relative to the validation set is in part due to inter-site variation in overall image tile characteristics.

Table 1 shows the distribution of geologic contexts (determined by majority vote across annotations) in both the train and test sets. The splitting of data by spatial partitioning induces a significant shift in this distribution in which two of the contexts (Crater Rim/Wall and Gully) are not represented in the test set. This suggests a relatively large degree of variability in terrain types covered by observations at each site. Figure 4 shows the overall differences in pixel intensity distributions across the train and test sets. While

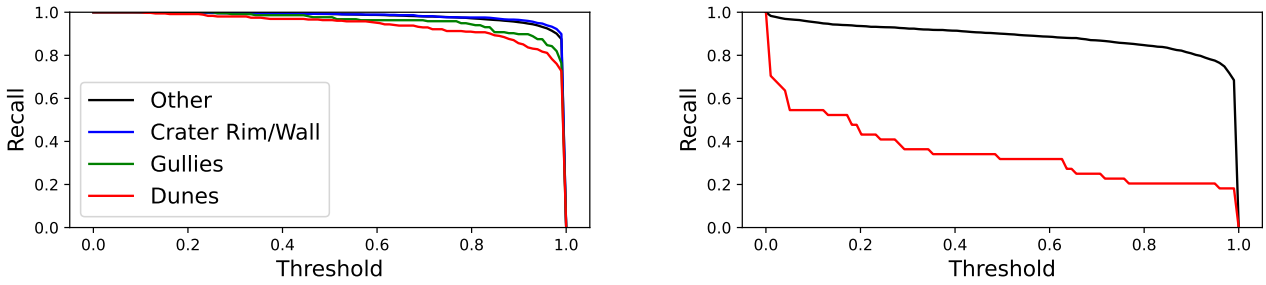

Figure 5: Frost detection recall scores as a function classification threshold for train (left) and test (right) sets.

these distributions are similar, they do show some degree of overall covariate shift in addition to the shift in geologic context.

Focusing on context-dependent performance, Figure 5 shows classifier recall on the train and test sets for each individual geologic context. Recall is plotted as a function of classification threshold. Even within the training set, there is reduced recall for some contexts, particularly gullies and dunes. Within the test set, there is a significant drop in recall for dunes relative to other contexts.

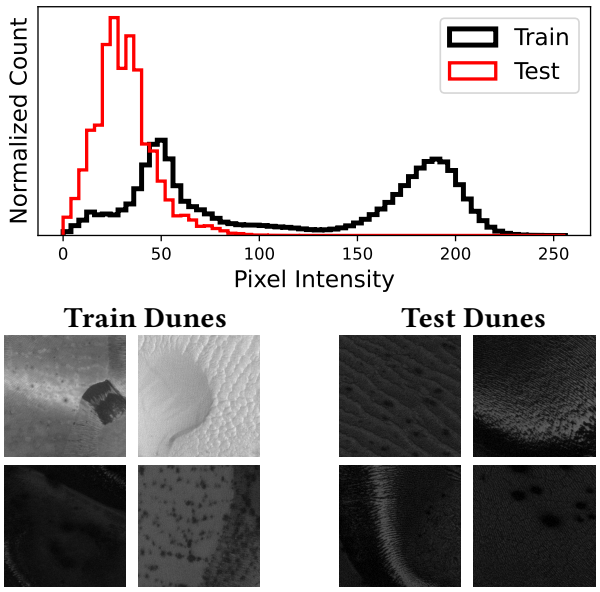

Figure 6: (Top) Difference in pixel intensities across dune images within train and test sets. (Bottom) Representative examples of dune tiles in train and test sets.

To better understand the drop in frost recall scores on dunes, we visually inspect the differences in these tiles across the train and test sets, shown in Figure 6. Overall, we see that the distribution of pixel intensities is multi-modal in the training set, with some overall bright and dark observations, whereas in the test set, only one mode is represented with darker pixels overall. The lower half of Figure 6 shows some representative examples from each set and shows a difference in the diversity of frost appearance across the two sets. These overall differences in brightness may be due to some combination of exposure and illumination across sites.

5 CONCLUSION AND FUTURE WORK

In this work, we explore the use of ML models to automate the detection of surface frost in high-resolution Martian images. Specifically, we propose a new application of HEALPix for partitioning spatial data to mitigate artificially inflated estimates of generalization performance across geologically varying sites on the surface of Mars. We explore the biases in model performance induced by the variability of observed geologic and observational characteristics across sites. In order to quantify this bias, collecting information about geologic context during labeling was an essential component of building a machine-learning-ready dataset for this domain.

We found that geologic context bias is present and significant for this model's performance on the test set, specifically for dune fields often found in northern mid-latitude craters. Interestingly, for human annotators, dunes often provide strong evidence of frost due to the striking visual appearance of defrosting marks which expose dark basalt sand beneath light-color frost. However, there is also a large degree of diversity in frost appearance on this underrepresented terrain type, both inherently and due to differing illumination and observational conditions, perhaps making the concept challenging for the classification model to learn.

To improve model performance and generalizability in future work, we propose (1) expanding the training set to include more diverse examples of underrepresented terrain types, (2) expanding the number of sites used to improve representation of all terrain types in the validation and test sets, and (3) performing contrast- and brightness-based augmentation to promote generalization under varying observational conditions. We expect these improvements will permit the training of models better suited for full-planet frost detection, thereby facilitating the creation of global frost maps.

ACKNOWLEDGMENTS

Part of this research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

REFERENCES

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. https://www.tensorflow.org/ Software available from tensorflow.org.

[2] Serina Diniega, Ali M. Bramson, Bonnie Buratti, Peter Buhler, Devon M. Burr, Matthew Chojnacki, Susan J. Conway, Colin M. Dundas, Candice J. Hansen, Alfred S. McEwen, Mathieu G.A. Lapôtre, Joseph Levy, Lauren Mc Keown, Sylvain Piqueux, Ganna Portyankina, Christy Swann, Timothy N. Titus, and Jacob M. Widmer. 2021. Modern Mars' geomorphological activity, driven by wind, frost, and gravity. *Geomorphology* 380 (2021), 107627. https://doi.org/10.1016/j.geomorph.2021.107627

[3] Serina Diniega and Isaac B Smith. 2020. High-priority science questions identified at the Mars Workshop on Amazonian and Present-Day Climate. *Planetary and Space Science* 182 (2020), 104813.

[4] Krzysztof M Gorski, Eric Hivon, Anthony J Banday, Benjamin D Wandelt, Frode K Hansen, Mstvos Reinecke, and Matthia Bartelmann. 2005. HEALPix: A framework for high-resolution discretization and fast analysis of data distributed on the sphere. *The Astrophysical Journal* 622, 2 (2005), 759.

[5] C.J. Hansen, N. Thomas, G. Portyankina, A. McEwen, T. Becker, S. Byrne, K. Herkenhoff, H. Kieffer, and M. Mellon. 2010. HiRISE observations of gas sublimation-driven activity in Mars' southern polar regions: I. Erosion of the surface. *Icarus* 205, 1 (2010), 283–295. https://doi.org/10.1016/j.icarus.2009.07.021 MRO/HiRISE Studies of Mars.

[6] Shachar Kaufman, Saharon Rosset, Claudia Perlich, and Ori Stitelman. 2012. Leakage in Data Mining: Formulation, Detection, and Avoidance. *ACM Trans. Knowl. Discov. Data* 6, 4, Article 15 (dec 2012), 21 pages. https://doi.org/10.1145/2382577.2382579

[7] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[8] L. Lange, S. Piqueux, and C. S. Edwards. 2022. Gardening of the Martian Regolith by Diurnal CO_2 Frost and the Formation of Slope Streaks. *Journal of Geophysical Research: Planets* 127, 4 (2022). https://doi.org/10.1029/2021JE006988 arXiv:https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2021JE006988

[9] Alfred McEwen. 2007. Mars Reconnaissance Orbiter High Resolution Imaging Science Experiment, Reduced Data Record, MRO-M-HIRISE-3-RDR-V1.1. NASA Planetary Data System.

[10] Alfred S. McEwen, Eric M. Eliason, James W. Bergstrom, Nathan T. Bridges, Candice J. Hansen, W. Alan Delamere, John A. Grant, Virginia C. Gulick, Kenneth E. Herkenhoff, Laszlo Keszthelyi, Randolph L. Kirk, Michael T. Mellon, Steven W. Squyres, Nicolas Thomas, and Catherine M. Weitz. 2007. Mars Reconnaissance Orbiter's High Resolution Imaging Science Experiment (HiRISE). *Journal of Geophysical Research: Planets* 112, E5 (2007). https://doi.org/10.1029/2005JE002605

[11] Sylvain Piqueux, Armin Kleinböhl, Paul O. Hayne, Nicholas G. Heavens, David M. Kass, Daniel J. McCleese, John T. Schofield, and James H. Shirley. 2016. Discovery of a widespread low-latitude diurnal CO_2 frost cycle on Mars. *Journal of Geophysical Research: Planets* 121, 7 (2016), 1174–1189. https://doi.org/10.1002/2016JE005034

[12] A. Pommerol, G. Portyankina, N. Thomas, K.-M. Aye, C. J. Hansen, M. Vincendon, and Y. Langevin. 2011. Evolution of south seasonal cap during Martian spring: Insights from high-resolution observations by HiRISE and CRISM on Mars Reconnaissance Orbiter. *Journal of Geophysical Research: Planets* 116, E8 (2011). https://doi.org/10.1029/2010JE003790

[13] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going Deeper with Convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/CVPR.2015.7298594

[14] Alexandre M.J.-C. Wadoux, Gerard B.M. Heuvelink, Sytze de Bruin, and Dick J. Brus. 2021. Spatial cross-validation is not the right way to evaluate map accuracy. *Ecological Modelling* 457 (2021), 109692. https://doi.org/10.1016/j.ecolmodel.2021.109692

[15] Jacob M. Widmer and Serina Diniega. 2019. Constraining the Spatial Extent and Timing of Local-Scale Seasonal Frost in the Northern Mid Latitude Region of Mars. In *50^{th} Annual Lunar and Planetary Science Conference*.