# Toward Foundation Models for Earth Monitoring: Proposal for a Climate Change Benchmark

**Alexandre Lacoste** [1] **Evan David Sherwin** [2] **Hannah Kerner** [3] **Hamed Alemohammad** [4] **Björn Lütjens** [5] **Jeremy Irvin** [2] **David Dao** [6] **Alex Chang** [1] **Mehmet Gunturkun** [1] **Alexandre Drouin** [1] **Pau Rodriguez** [1] **David Vazquez** [1]

## Abstract

Recent progress in self-supervision shows that pre-training large neural networks on vast amounts of unsupervised data can lead to impressive increases in generalisation for downstream tasks. Such models, recently coined as *foundation models*, have been transformational to the field of natural language processing. While similar models have also been trained on large corpuses of images, they are not well suited for remote sensing data. To stimulate the development of foundation models for Earth monitoring, we propose to develop a new benchmark comprised of a variety of downstream tasks related to climate change. We believe that this can lead to substantial improvements in many existing applications and facilitate the development of new applications. This proposal is also a call for collaboration with the aim of developing a better evaluation process to mitigate potential downsides of foundation models for Earth monitoring.

## 1. Introduction

Earth monitoring with machine learning-based methods plays an increasing role in climate change mitigation and adaptation as well as climate science (Rolnick et al., 2019). Applications include methane source detection (Sheng et al., 2020; Dileep et al., 2020), forest carbon quantification (Lütjens et al., 2019), deforestation monitoring (Finer et al., 2018; Dao et al.), flood detection (Mateo-Garcia et al., 2021), extreme weather prediction (McGovern et al., 2017), wildfire detection (Jain et al., 2020), and crop monitoring (Kerner et al., 2020; Dado et al., 2020). Across many of these applications, pre-trained models (e.g., a ResNet trained on ImageNet) are used to increase generalisation performance. Improvement of the pre-trained models is shown to reduce the need for large labelled datasets in some contexts (Chen et al., 2020) and can improve model generalisation outside of the training distribution (Hendrycks et al., 2019). Recent studies exploring the scaling of such pre-trained models found that increasing the size of an unsupervised (or weakly supervised) dataset as well as properly scaling the model led to an even greater increase in performances under various metrics (Kaplan et al., 2020; Radford et al., 2021).

While the training of such large-scale models is usually reserved for industrial research labs with very large computer clusters, the publication of the pre-trained models opens opportunities to the rest of the community. These pre-trained models were recently coined as *foundation models* (Bommasani et al., 2021) as they might serve as foundations for sub-fields of machine learning. Specifically, the publication of large pre-trained models like BERT (Devlin et al., 2018), and GPT-3 (Brown et al., 2020) led to a paradigm shift in the field of natural language processing (NLP). This inspired a similar shift in the field of computer vision with the release of models like CLIP (Radford et al., 2021) and DINO (Caron et al., 2021). While CLIP performs well on various types of vision tasks, it is still under-performing on Earth monitoring tasks (Radford et al., 2021). This is not surprising as it is trained mainly on RGB images taken from a ground perspective, rather than multispectral bands taken from an overhead perspective prevalent in remote sensing data. This suggests that there is still untapped potential for foundation models to benefit the field Earth monitoring as it has done for NLP and computer vision.

Foundation models also come with downsides. Specifically, large language models are known to amplify and perpetuate biases (Bender et al., 2021) and have high $CO_2$e emissions associated with their training (Strubell et al., 2019; Patterson et al., 2021). Recently, an interdisciplinary group of researchers published a collective work discussing the risks and opportunities of foundation models (Bommasani

[1]Element AI / Service Now [2]Stanford University [3]University of Maryland [4]Radiant Earth Foundation [5]MIT [6]ETH Zurich. Correspondence to: Alexandre Lacoste <alexandre.lacoste@servicenow.com>.

*Proceedings of the $38^{th}$ International Conference on Machine Learning*, PMLR 139, 2021. 

et al., 2021). This study highlighted that the relevant stakeholders are often not well represented during the design of foundation models. In addition, the increased accessibility of foundation models can lead to the development of unexpected applications with potential positive and negative impacts. To mitigate potential negative impacts, we suggest an open evaluation procedure early in the process. To this end, we propose a benchmark dataset and evaluation process to facilitate the development of foundation models in Earth monitoring. We will aggregate a collection of downstream tasks such as classification or semantic segmentation to identify ground-based features, provide corresponding labelled datasets, and define a transparent evaluation procedure with open-source code. To highlight the importance of working on climate change, benchmark datasets and tasks will focus on multiple areas related to understanding, mitigating, and adapting to climate change. The advantages of such a benchmark are numerous, as they:

- stimulate and facilitate the development of foundation models for Earth monitoring,
- provide a systematic way of measuring the quality of models for better scientific progress,
- provide insights into which pre-trained models work best for specific climate-related tasks, and
- preemptively reduce negative impacts of foundation models through an appropriate evaluation procedure.

This work is a proposal and a call to action. We ask the community to engage by proposing suitable datasets, flagging potential concerns, and proposing modifications to the evaluation procedure. In Appendix A, we review the potential positive and negative societal impacts of this work.

## 2. Remote sensing data for self-supervision

The development of foundation models does not typically rely on a specific dataset for the pre-training phase. The choice of data is part of the design of the model, e.g., a very large corpus of text from the internet (Mitchell et al., 2018) or pairs of text associated with images from the web (Radford et al., 2021). To follow this trend, the data for training foundation models will not be provided with the benchmark. Potential sources of data are listed below.

**Multispectral with revisits** Data sources such as Sentinel 2 (Drusch et al., 2012; ESA, 2021) and Landsat 8 (USGS, 2021) provide images in multiple spectral bands with periodic revisits. This yields a 4-dimensional array of structured data (longitude, latitude, wavelength, time) which can be used to perform various forms of self-supervision, e.g., predicting adjacent tiles (Jean et al., 2019) or contrasting the different seasons for the same region (Mañas et al., 2021).

**Other sensors** Synthetic Aperture Radar (SAR) and terrain elevation are also frequently available and can be matched to other sources of data through geolocalisation (Pepin et al., 2020). Such data are complementary to spectral bands and may encourage the model to learn higher-level semantic representations.

**Semantic data** Through georeferencing, text-based data such as Wikipedia articles can be linked to satellite images (Uzkent et al., 2019). It is also possible to join content from non-image data layers like OpenStreetMap (Li et al., 2020). By predicting or contrasting information from these sources, the model may learn useful and transferable semantic representations.

## 3. The Benchmark

### 3.1. Climate Change Downstream Tasks

The aim is to provide a variety of downstream tasks to evaluate different aspects of foundation models pre-trained on other datasets. To go beyond simple image classification, since it is often not representative of real-world tasks, we include segmentation, regression, and counting tasks. However, for the dataset to be useful in this benchmark, several other criteria need to be met:

**Not too big** Remote sensing datasets can be comprised of millions of samples totalling terabytes of data. Benchmark datasets should be small enough to easily download onto a personal computer, roughly 100 to a few thousand labelled samples per task. If the license permits it, the dataset can be sub-sampled.

**Permissive license** Most datasets need to be adapted to fit a conventional machine learning pipeline. In such cases, a permissive license (e.g., Creative Commons) is required.

**Multispectral and SAR** One of the main reasons to have a foundation model tailored to remote sensing is to learn how to better interpret multispectral and SAR data. To evaluate their ability to do so, a substantial fraction of the benchmark datasets must contain multispectral and SAR data on tasks that can leverage such information.

**Meta information for distribution shift evaluation** We also aim to evaluate model performance under distribution shift, when the model is applied to data from a different distribution than the training data (Koh et al., 2021). Of specific interest are downstream tasks in which the training set and the testing set are in different countries. Other variables such as date, sun elevation, and spatial resolution can also provide insightful distribution shift evaluations.

We present current candidate datasets that we are considering for this benchmark in Appendix Table 1. We encourage the community to contact us to propose additional datasets.

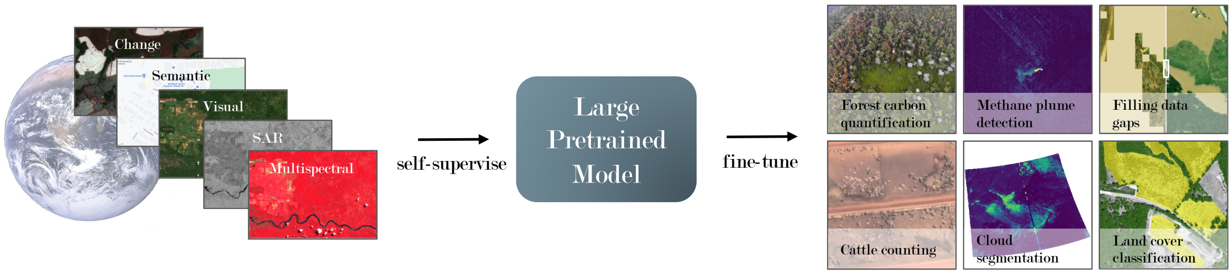


*Figure 1.* Foundation models encapsulate multimodal data streams through self-supervised training. The trained models can then be fine-tuned for a variety of climate-related remote sensing tasks. Image sources: quantification (Lütjens et al., 2019), detection (Jongaram-rungruang et al., 2021), generation (Lütjens et al., 2021), counting (Laradji et al., 2020), segmentation (Zantedeschi et al., 2019), and multi-class classification (Pallai and Wesson, 2017).

### 3.2. Automatic fine-tuning of the model

To evaluate a pre-trained model, it is common to simply "probe" the model, i.e., use the learned representations from the model as the input features to another model (Jean et al., 2019). However, fine-tuning the model to the given task has proven to generalize better and is closer to the needs of practitioners (Mañas et al., 2021; Chen et al., 2020). Adapting a pre-trained architecture to a variety of types of tasks for fine-tuning comes with significant technical challenges. To this end, we will provide a GitHub codebase with the necessary tools to facilitate and standardise the evaluation procedure. The codebase will provide the following features:

**Fine-tuning code** To facilitate and standardise fine-tuning, the benchmark will provide code for adapting popular architectures such as ResNet (He et al., 2016) and Visual Transformer (Kolesnikov et al., 2021) to the supported types of task such as classification, segmentation and detection.

**Fine-tuning API** When the pre-trained network is not compatible with existing fine-tuning methods, we encourage the users to submit a pull request to grow the library.

**Evaluation of representations** Often called probing, this approach does not require fine-tuning. The pre-trained model encode every images of all tasks and predictions are made from the fixed features. This approach requires less computations and is less likely to have compatibility issues.

### 3.3. Evaluation Metrics

We propose to include a variety of metrics to enable rigorous evaluation of the pre-trained models:

**Task-specific metrics** We propose to report a few metrics that are natural to each task being evaluated, e.g., F1 for classification tasks and mIoU for semantic segmentation.

**Aggregated metric** For a valid comparison of a pre-trained model across multiple tasks, we will use the pairwise sign test (Lacoste et al., 2012). This simply counts the number of times one model outperforms a baseline and assesses if the difference is significant. When a few strong baselines are compared, Friedman's test (Friedman, 1937) can be used to provide a more powerful test.

**Distribution shift** As specified in Section 3.1, we are collecting metadata for distribution shift evaluation. This is done by partitioning the train, validation, and test sets of each dataset based on specific values of a selected metadata variable such as country or date. Each partition yields a different evaluation with potentially different insights.

**Energy efficiency and CO2 equivalent emissions** We will also report energy consumption, and $tCO_2e$ emissions during the benchmarking phase for each model(Lacoste et al., 2019; Schmidt et al., 2021). These emissions are expected to be significantly smaller than that of the pre-training phase, which we do not have access. However, this evaluation will provide a good comparison, highlighting which model is more energy efficient.

## 4. Conclusion

We propose to develop a new benchmark for evaluating foundation models on climate change downstream tasks. This involves adapting a variety of remote sensing datasets to a more conventional machine learning pipeline and providing code for fine-tuning and evaluating on individual tasks. We expect that this benchmark will stimulate the development of new foundation models that could lead to better generalisation on a variety of climate-related downstream tasks and could open up opportunities for new applications.

This proposal is also a call for collaboration. We hope to receive recommendations to include additional public datasets as well as datasets that have not yet been released. We also welcome any recommendations about the evaluation procedure that could improve the validation of foundation models for Earth monitoring and mitigate their potential downsides.

## References

H. Alemohammad. The case for open-access ML-ready geospatial training data. In *International Geoscience and Remote Sensing Symposium*. IEEE, 2021.

H. Alemohammad and K. Booth. LandCoverNet: A global benchmark land cover classification training dataset. *arXiv:2012.03111 [cs]*, Dec. 2020. URL http://arxiv.org/abs/2012.03111. arXiv: 2012.03111.

E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, 2021.

R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

M. Burke, A. Driscoll, D. B. Lobell, and S. Ermon. Using satellite imagery to understand and promote sustainable development. *Science*, 371(6535), 2021.

M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294*, 2021.

T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

M. T. Chiu, X. Xu, Y. Wei, Z. Huang, A. G. Schwing, R. Brunner, H. Khachatrian, H. Karapetyan, I. Dozier, G. Rose, D. Wilson, A. Tudor, N. Hovakimyan, T. S. Huang, and H. Shi. Agriculture-Vision: A Large Aerial Image Database for Agricultural Pattern Analysis. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2825–2835, Seattle, WA, USA, June 2020. IEEE. ISBN 978-1-72817-168-5. doi: 10.1109/CVPR42600.2020.00290. URL https://ieeexplore.ieee.org/document/9157397/.

W. T. Dado, J. M. Deines, R. Patel, S.-Z. Liang, and D. B. Lobell. High-resolution soybean yield mapping across the us midwest using subfield harvester data. *Remote Sensing*, 12(21):3471, 2020.

D. Dao, C. Cang, C. Fung, M. Zhang, N. Pawlowski, R. Gonzales, N. Beglinger, and C. Zhang. Gainforest: Scaling climate finance for forest conservation using interpretable machine learning on satellite imagery.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

S. Dileep, D. Zimmerle, J. R. Beveridge, and T. Vaughn. Automated identification of oil field features using cnns. 2020.

M. Drusch, U. Del Bello, S. Carlier, O. Colin, V. Fernandez, F. Gascon, B. Hoersch, C. Isola, P. Laberinti, P. Martimort, et al. Sentinel-2: Esa's optical high-resolution mission for gmes operational services. *Remote sensing of Environment*, 120:25–36, 2012.

R. M. Duren, A. K. Thorpe, K. T. Foster, T. Rafiq, F. M. Hopkins, V. Yadav, B. D. Bue, D. R. Thompson, S. Conley, N. K. Colombi, C. Frankenberg, I. B. McCubbin, M. L. Eastwood, M. Falk, J. D. Herner, B. E. Croes, R. O. Green, and C. E. Miller. California's methane super-emitters. *Nature*, 575(7781):180–184, Nov. 2019. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-019-1720-3. URL http://www.nature.com/articles/s41586-019-1720-3.

EPA. Greenhouse Gas Emissions: Understanding Global Warming Potentials. Technical report, US Environmental Protection Agency, Feb. 2017. URL https://www.epa.gov/ghgemissions/understanding-global-warming-potentials.

ESA. Sentinel-2. Technical report, European Space Agency, Paris, France, 2021. URL https://sentinel.esa.int/web/sentinel/missions/sentinel-2.

M. Finer, S. Novoa, M. Weisse, R. Petersen, J. Mascaro, T. Souto, F. Stearns, and R. Martinez. Combating deforestation: From satellite to intervention. *Science*, 360: 1303 – 1305, 2018.

M. Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the american statistical association*, 32(200):675–701, 1937.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

D. Hendrycks, M. Mazeika, S. Kadavath, and D. Song. Using self-supervised learning can improve model robustness and uncertainty. *arXiv preprint arXiv:1906.12340*, 2019.

P. Jain, S. C. Coogan, S. G. Subramanian, M. Crowley, S. Taylor, and M. D. Flannigan. A review of machine learning applications in wildfire science and management. *Environmental Reviews*, 28(4):478–505, 2020.

N. Jean, S. Wang, A. Samar, G. Azzari, D. Lobell, and S. Ermon. Tile2vec: Unsupervised representation learning for spatially distributed data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3967–3974, 2019.

F. Johnson, A. Wlazlo, R. Keys, V. Desai, E. Wetherley, R. Calvert, and E. Berman. Airborne methane surveys pay for themselves: An economic case study of increased revenue from emissions control. preprint, Environmental Monitoring, July 2021. URL http://eartharxiv.org/repository/view/2532/.

S. Jongaramrungruang, C. Frankenberg, A. K. Thorpe, and G. Matheou. Methanet - an ai-driven approach to quantifying methane point-source emission from high-resolution 2-d plume imagery. *ICML Workshop on Tackling Climate Change with AI*, 2021.

J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

H. Kerner, G. Tseng, I. Becker-Reshef, C. Nakalembe, B. Barker, B. Munshell, M. Paliyam, and M. Hosseini. Rapid response crop maps in data sparse regions. *arXiv preprint arXiv:2006.16866*, 2020.

P. W. Koh, S. Sagawa, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao, T. Lee, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021.

A. Kolesnikov, A. Dosovitskiy, D. Weissenborn, G. Heigold, J. Uszkoreit, L. Beyer, M. Minderer, M. Dehghani, N. Houlsby, S. Gelly, T. Unterthiner, and X. Zhai. An image is worth 16x16 words: Transformers for image recognition at scale. 2021.

A. Lacoste, F. Laviolette, and M. Marchand. Bayesian comparison of machine learning algorithms on single and multiple datasets. In *Artificial Intelligence and Statistics*, pages 665–675. PMLR, 2012.

A. Lacoste, A. Luccioni, V. Schmidt, and T. Dandres. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*, 2019.

I. Laradji, P. Rodriguez, F. Kalaitzis, D. Vazquez, R. Young, E. Davey, and A. Lacoste. Counting cows: Tracking illegal cattle ranching from high-resolution satellite imagery. *arXiv preprint arXiv:2011.07369*, 2020.

H. Li, X. Dou, C. Tao, Z. Wu, J. Chen, J. Peng, M. Deng, and L. Zhao. Rsi-cb: A large-scale remote sensing image classification benchmark using crowdsourced data. *Sensors*, 20(6):1594, 2020.

B. Lütjens, L. Liebenwein, and K. Kramer. Machine learning-based estimation of forest carbon stocks to increase transparency of forest preservation efforts. *2019 NeurIPS Workshop on Tackling Climate Change with AI (CCAI)*, 2019.

B. Lütjens, B. Leshchinskiy, C. Requena-Mesa, F. Chishtie, N. Díaz-Rodríguez, O. Boulais, A. Sankaranarayanan, A. Pina, Y. Gal, C. Raissi, A. Lavin, and D. Newman. Physically-consistent generative adversarial networks for coastal flood visualization. *ICML Workshop on AI for Modeling Oceans and Climate Change (AIMOCC)*, 2021.

L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson. Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS journal of photogrammetry and remote sensing*, 152:166–177, 2019.

O. Mañas, A. Lacoste, X. Giro-i Nieto, D. Vazquez, and P. Rodriguez. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. *arXiv preprint arXiv:2103.16607*, 2021.

M. Maskey, H. Alemohammad, K. Murphy, and R. Ramachandran. Advancing ai for earth science: A data systems perspective. *Eos*, 101, 2020a.

M. Maskey, R. Ramachandran, M. Ramasubramanian, I. Gurung, B. Freitag, A. Kaulfus, D. Bollinger, D. J. Cecil, and J. Miller. Deepti: Deep-learning-based tropical cyclone intensity estimation system. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13:4271–4281, 2020b. doi: 10.1109/JSTARS.2020.3011907.

G. Mateo-Garcia, J. Veitch-Michaelis, L. Smith, S. V. Oprea, G. Schumann, Y. Gal, A. G. Baydin, and D. Backes. Towards global flood mapping onboard low cost satellites with machine learning. *Scientific Reports*, 11, 2021.

A. McGovern, K. L. Elmore, D. J. Gagne, S. E. Haupt, C. D. Karstens, R. Lagerquist, T. Smith, and J. K. Williams. Using artificial intelligence to improve real-time decision-making for high-impact weather. *Bulletin of the American Meteorological Society*, 98(10), 2017.

T. Mitchell, W. Cohen, E. Hruschka, P. Talukdar, B. Yang, J. Betteridge, A. Carlson, B. Dalvi, M. Gardner, B. Kisiel, J. Krishnamurthy, N. Lao, K. Mazaitis, T. Mohamed, N. Nakashole, E. Platanios, A. Ritter, M. Samadi, B. Settles, R. Wang, D. Wijaya, A. Gupta, X. Chen, A. Saparov, M. Greaves, and J. Welling. Never-ending learning. *Communications of the ACM*, 61(5):103–115, Apr. 2018. ISSN 0001-0782, 1557-7317. doi: 10.1145/3191513. URL https://dl.acm.org/doi/10.1145/3191513.

C. Pallai and K. Wesson. Chesapeake bay program partnership high-resolution land cover classification accuracy assessment methodology, 2017. URL https://chesapeakeconservancy.org/wp-content/uploads/2017/01/Chesapeake_Conservancy_Accuracy_Assessment_Methodology.pdf.

D. Patterson, J. Gonzalez, Q. Le, C. Liang, L.-M. Munguia, D. Rothchild, D. So, M. Texier, and J. Dean. Carbon emissions and large neural network training. *arXiv preprint arXiv:2104.10350*, 2021.

K. Pepin, H. A. Zebker, and W. Ellsworth. High-Pass Filters to Reduce the Effects of Broad Atmospheric Contributions in Sbas Inversions: A Case Study in the Delaware Basin. In *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, pages 1030–1033, Waikoloa, HI, USA, Sept. 2020. IEEE. ISBN 978-1-72816-374-1. doi: 10.1109/IGARSS39084.2020.9324656. URL https://ieeexplore.ieee.org/document/9324656/.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

C. Rambour, N. Audebert, E. Koeniguer, B. Le Saux, M. Crucianu, and M. Datcu. Flood detection in time series of optical and SAR images. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIII-B2-2020:1343–1346, Aug. 2020. ISSN 2194-9034. doi: 10.5194/isprs-archives-XLIII-B2-2020-1343-2020. URL https://www.int-arch-photogramm-remote-sens-spatial-inf-sci.net/XLIII-B2-2020/1343/2020/.

B. Rausch, K. Mayer, M.-L. Arlt, G. Gust, P. Staudt, C. Weinhardt, D. Neumann, and R. Rajagopal. An enriched automated pv registry: Combining image recognition and 3d building data. *arXiv preprint arXiv:2012.03690*, 2020.

D. Rolnick, P. L. Donti, L. H. Kaack, K. Kochanski, A. Lacoste, K. Sankaran, A. S. Ross, N. Milojevic-Dupont, N. Jaques, A. Waldman-Brown, et al. Tackling climate change with machine learning. *arXiv preprint arXiv:1906.05433*, 2019.

M. R. V. Ross, S. N. Topp, A. P. Appling, X. Yang, C. Kuhn, D. Butman, M. Simard, and T. M. Pavelsky. AquaSat: A Data Set to Enable Remote Sensing of Water Quality for Inland Waters. *Water Resources Research*, 55(11):10012–10025, Nov. 2019. ISSN 0043-1397, 1944-7973. doi: 10.1029/2019WR024883. URL https://onlinelibrary.wiley.com/doi/10.1029/2019WR024883.

V. Schmidt, K. Goyal, A. Joshi, B. Feld, L. Conell, N. Laskaris, D. Blank, J. Wilson, S. Friedler, and S. Luccioni. CodeCarbon: Estimate and Track Carbon Emissions from Machine Learning Computing. 2021. doi: 10.5281/zenodo.4658424.

R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni. Green ai. *Communications of the ACM*, 63(12):54–63, 2020.

H. Sheng, J. Irvin, S. Munukutla, S. Zhang, C. Cross, K. Story, R. Rustowicz, C. Elsworth, Z. Yang, M. Omara, et al. Ognet: Towards a global oil and gas infrastructure database using deep learning on remotely sensed imagery. *arXiv preprint arXiv:2011.07227*, 2020.

E. Strubell, A. Ganesh, and A. McCallum. Energy and policy considerations for deep learning in nlp. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, 2019.

USGS. Landsat 8. Technical report, United States Geological Survey, Reston, Virginia, USA, 2021. URL https://www.usgs.gov/core-science-systems/nli/landsat/landsat-8?qt-science_support_page_related_con=0#qt-science_support_page_related_con.

B. Uzkent, E. Sheehan, C. Meng, Z. Tang, M. Burke, D. Lobell, and S. Ermon. Learning to interpret satellite images using wikipedia. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019.

V. Zantedeschi, F. Falasca, A. Douglas, R. Strange, M. J. Kusner, and D. Watson-Parris. Cumulo: A dataset for learning cloud classes. *arXiv preprint arXiv:1911.04227*, 2019.

X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, 5(4):8–36, 2017.

# A. Societal Impact of Foundation Models for Earth Monitoring

Remote sensing and Earth monitoring have been transformational in the past decades. Applications include military, insurance, market forecasting, climate science, and more. Much of this impact is not directly attributed to deep learning nor large pre-trained networks and its review extends beyond the scope of this section. In this section, our focus is on the impact of bringing foundation models to Earth monitoring.

## A.1. Climate mitigation and adaptation

Machine learning on remote sensing data is widely used to develop solutions for a variety of problems relevant to climate change (Burke et al., 2021; Rolnick et al., 2019; Zhu et al., 2017; Ma et al., 2019). The vast majority of these solutions are built by curating datasets for a specific task and require significant resources to develop. Furthermore, the solutions are often tailored to specific regions as extending approaches to new geographies remains a significant challenge, primarily due to the lack of labeled data (Zhu et al., 2017). Less-economically developed regions of the world are no less susceptible to the impacts of climate change, yet suffer from the lack of effective remote sensing-based solutions (Burke et al., 2021). Foundation models for Earth monitoring have the potential to address many of these issues and substantially accelerate and enable the development of new remote sensing solutions for climate change.

## A.2. Increased accessibility

Reducing the need for curating a large labeled dataset for each task could democratize access to the development of machine learning models for remote sensing, specifically for groups or organisations with limited budgets (Maskey et al., 2020a; Alemohammad, 2021). In particular, foundation models may especially benefit non-profit organisations, academic universities, startups, and developing countries. It may also open opportunities for applications that were not previously profitable. Although we believe that increased accessibility to these models will have a largely net positive impact, we acknowledge that this accessibility may lead to unexpected applications with potentially negative impacts (Bommasani et al., 2021). We also note that such models may have dual-use applications, where, for example, they may help oil and gas industries in their operations in ways that increase (or reduce) overall emissions.

## A.3. Emissions of large pre-trained models

Recent work has investigated emissions of large neural networks (Strubell et al., 2019; Schwartz et al., 2020; Schmidt et al., 2021; Lacoste et al., 2019; Patterson et al., 2021). Specifically, training a large transformer can emit 284 $tCO_2e$ when trained on computers using largely fossil fuel energy (US national average) (Strubell et al., 2019). When put in perspective with individual actions, such emissions are large—e.g., a roundtrip passenger flight from San Francisco to London is 2.8 $tCO_2e$ , about 100× smaller. However, the extensive reusability of pre-trained models and their potential for helping efforts to mitigate climate change (Rolnick et al., 2019) calls for a different perspective.

When evaluating new tools and systems, it is important to consider the likely net impact on emissions of both the creation and testing of the tool and its eventual deployment. For example, evaluating the performance of airborne methane sensing tools at emission levels commonly found in oil and gas operations can emit about 7 metric tonnes of methane, roughly 600 $tCO_2e$ equivalent using a 20-year global warming potential (EPA, 2017). However, in a single day of flying, such a single instrument can survey hundreds of sites, often identifying leaks for repair that emit well over 7 metric tonnes of methane per day (Johnson et al., 2021). Similarly, foundation models may significantly advance our ability to leverage enormous quantities of passively collected satellite data to massively reduce emissions, qualitatively advance our understanding of climate science, or improve our ability to adapt to climate change.

In sum, the potential benefits for climate change mitigation with improved Earth monitoring methods likely outweigh the emissions associated with foundational models. Moreover, various actions can be taken to reduce and mitigate emissions related to the training of your model (Lacoste et al., 2019):

- Select data centers that are certified carbon neutral or largely powered by renewable energy, with good power usage effectiveness (PUE). Such measures can reduce emissions dramatically  50× reduction in emissions (Lacoste et al., 2019).
- Design your code development pipeline to minimize the number of computationally-intensive runs required, e.g. employ modular development and testing when possible.
- Make your code more efficient and sparsify your network when possible (Patterson et al., 2021). This can reduce emissions up to  10-fold.
- Favour more energy-efficient hardware, e.g., TPUs or GPUs.
- Monitor (Schmidt et al., 2021) and report your emissions (Lacoste et al., 2019). Better communication about climate change is fundamental for systemic changes. Better documentation will help other coders pick up where you left off, potentially bypassing some

computationally intensive runs.
- Offset the cumulative emissions of your projects.

### A.4. Fairness and biases

Large language models are known to amplify and perpetuate biases (Bender et al., 2021). While this can lead to serious societal issues, we believe that biases in remote sensing models are likely to have much less impact. We do however anticipate potential biases and fairness issues.

**Data coverage and resolution** Some satellites cover the whole Earth with standard spatial resolution and revisit rate (e.g., Sentinel-2 covers the whole Earth at 10-60 m/pixel resolution every 5 days). This makes imagery freely available uniformly across the planet. Other satellite data providers such as Maxar acquire images on-demand and have higher spatial resolution (up to 0.3m per pixel), but also have lower revisit rates and high costs. Some countries, such as New Zealand, freely provide aerial imagery with resolution up to 0.1m per pixel[1]. Finally, it is worth noting that cloudy seasons in some climates may limit data availability for some countries. Overall, while the coverage is fairly uniform, some regions have much higher coverage than others and money can be a limiting factor to access the data. This can lead to some level of biases and fairness issues.

## B. List of Downstream Tasks

[1] https://data.linz.govt.nz/

[h]

| Name | Task | Sector | # labels | Resolution | Spectral bands |
|---|---|---|---|---|---|
| AgricultureVision (Chiu et al., 2020) | Multi-class classification or segmentation of agricultural patterns important to farmers (e.g., planter skip or nutrient deficiency) in aerial images. | Agriculture | 94,986 | 10cm | RGB + near infrared |
| AquaSat (Ross et al., 2019) | Per-pixel regression to predict water quality (e.g., total suspended sediments) in satellite images. | Water quality | 600,000 | 30m | Multispectral + RGB |
| CalMethane Survey (Duren et al., 2019) | Methane plume classification | Energy | 60-1000 | 3m | Hyperspectral |
| CUMULO (Zantedeschi et al., 2019) | Detecting clouds to reduce uncertainties in climate models | Climate | 300,000 | 1km | Hyperspectral (36-band) |
| LandCoverNet (Alemohammad and Booth, 2020) | Segmentation via multispectral satellite imagery with annual land cover class per pixel | Land use | $\tilde{2}$,000 | 10m | Multispectral |
| SEN12-FLOOD (Rambour et al., 2020) | Image classification of multispectral and radar satellite imagery to identify flooded regions | Climate/ Adaptation | 5,567 | 10m | Multispectral +SAR |
| Tropical cyclone wind speed (Maskey et al., 2020b) | Regression-based estimation of surface wind speed of tropical cyclones using satellite imagery | Climate | 114,634 | 4km | Single-band microwave |
| 3D PV Locator (Rausch et al., 2020) | Classification and segmentation of solar panels in satellite imagery | Energy | 100,000 | 10cm | RGB |

*Table 1.* Example datasets for benchmarking climate-focused Erath monitoring foundation models. All listed datasets satisfy the criteria presented in Section 3.1