# A global inventory of photovoltaic solar energy generating units

Kruitwagen, L.,[1,*] Story, K.,[2] Friedrich, J.,[3] Byers, L.,[3] Skillman, S.,[2] & Hepburn, C.[1]

[1] *Smith School of Enterprise and the Environment, University of Oxford, Oxford, United Kingdom*

[2] *Descartes Labs Inc., San Francisco, CA, United States*

[3] *World Resources Institute, Washington, DC, United States*

* *Corresponding Author*

**Photovoltaic (PV) solar energy generating capacity has grown by 41% per year since 2009[1]. Energy system projections that mitigate climate change and facilitate universal energy access show a nearly ten-fold increase in PV solar energy generating capacity by 2040[2,3]. Geospatial data describing the energy system is required to manage generation intermittency, mitigate climate change risks, and identify trade-offs with biodiversity, conservation, and land protection priorities caused by the land use and land cover change necessary for PV deployment. Currently available inventories of solar generating capacity cannot fully address these needs [1–9]. Here, we provide a global inventory of commercial-, industrial-, and utility-scale PV solar energy generation stations (i.e. PV generating stations in excess of 10kW nameplate capacity) using a longitudinal corpus of remote sensing imagery, machine learning, and a large cloud computation infrastructure. We locate and verify 68,661 facilities, an increase of 432% (in number of facilities) on the previously best-available asset-level data. With the help of a hand-labelled test set, we estimate global installed generating capacity to be 423GW [-75GW, +77GW] at the end of 2018. Enrichment of our dataset with estimates of facility installation date, historic land cover classification, and proximity to**



Accepted Manuscript. [2021-10-27] Published in Nature, 598: 604-610.
Available at: https://www.nature.com/articles/s41586-021-03957-7

**vulnerable areas allows us to show that the majority of the PV solar energy facilities are sited on cropland, followed by aridlands and grassland. Our inventory can aid PV delivery aligned with the Sustainable Development Goals.**

In the International Energy Agency's (IEA) Sustainable Development Scenario, 4240GW of PV solar generating capacity is projected to be deployed by 2040[2], a ten-thousand-fold increase from 385 MW in 2000[1]. Fundamental asset-level datasets of the energy system are crucial for the operation of increasingly renewables-based electricity systems, and for the design, implementation, and validation of effective policy regimes and markets to deliver the diffusion of PV solar energy generation aligned with the UN Sustainable Development Goals. Currently-available inventories of solar generating capacity are insufficient to address these needs because they are either aggregated summary statistics (e.g. of the IEA[2], IRENA,[3] or BP[1]); limited in geographical scope (e.g. the World Resources Institute's Global Power Plant Database (GPPD)[4], OpenPV US database[5], Yu et al.'s DeepSolar[6], Huo et al.'s. SolarNet[7]); are not geospatially localised (e.g. S&P Global World Electric Power Plant Database (WEPP)[8]); and/or are not publicly available for the research and policy community (e.g. IHS' Electric Plants[9]). The generating behaviour of PV solar energy (and wind energy) reflects the uncertainty and complexity of the natural world; detailed asset-level data, including the spatial arrangement of installations, are required particularly to address the challenges of generation now-casting and forecasting faced by electricity system operators and electricity markets operators and participants. Decision makers seeking to implement the SDGs must navigate trade-offs between a 1.5°C constraint to anthropogenic climate change and other biodiversity, conservation, and land protection goals.

In this work, we produce such a dataset by mapping commercial-, industrial-, and utility-scale PV solar energy facility footprints in remote sensing imagery using deep learn-

ing. We adapt our definitions of PV installation sizes from the IEA[10], where 'commercial and industrial' installations are defined as those between 10kW and 1MW in nameplate capacity, 'utility-scale' as installations in excess of 1MW. For the purposes of our study, we also include 'off-grid' installations (as defined by the IEA) in excess of 10kW nameplate capacity. Bolinger et al. (2019)[11] categorise 'utility-scale' installation as those in excess of 5MW capacity. We adapt these two categorisations and use bins of 10kW to 1MW; 1MW to 5MW; and >5MW nameplate generating capacity as categories in our analysis and refer to our scope collectively as 'non-residential PV'. (Residential PV systems may, of course, have generating capacities larger than 10kW, however using this size threshold allows comparability with other aggregate statistics of installed solar PV generating capacity.) Our study is comparable to 89% of the capacity reported by the IEA[10] and includes both the conventional generating stations and behind-the-meter installations that electricity system operators might otherwise have poor visibility of. We prepare a machine learning pipeline comprising of three convolutional neural networks, two recurrent neural networks, and a number of heuristical filters and deploy it on a cloud computation infrastructure. The resulting dataset expands the best publicly-available facility-level data for PV solar energy by 432% (in number of facilities), including 18,449 new installations in China, 9,906 in Japan, 4,525 in the United States, 2,021 in India, and 17,918 in the European Economic Area. Our method is agnostic to political borders, and so provides new insight into the diffusion of PV solar energy in reporting-poor areas.

We enrich our data with estimates of generating capacity, installation date, historic land cover, and proximity to protected areas and indigenous and community lands. Policy makers need to properly incentivise the adoption of clean energy in line with economic development and climate goals, while being sensitive to the impacts renewable energy fa-

cilities might have on biodiversity, ecosystem services and vulnerable lands. Our observed installation dates can be used to generate insight into the gap between facility-level final investment decisions, construction start, construction completion, and facility operation. Our land cover analysis provides insight into global trends for PV siting decisions. Sampling from a global landcover map, we observe that non-residential PV is most commonly installed on croplands, followed by deserts and grasslands. We compare PV solar energy land cover to local and national land cover distributions to observe the bias in regional and local PV siting decisions. PV solar energy siting decisions favour agricultural areas and disfavour forests and shrubland. These findings highlight the need for further coherence on policy efforts seeking to navigate trade-offs between climate change mitigation, biodiversity and ecosystem health, conservation and land protection, and food production.

**Machine Learning Pipeline** Earth observations analyzed with machine learning is the only feasible method to produce a dataset like this on a global scale. The maturation of computer vision using convolutional neural networks (CNNs)[12–14] has unlocked new approaches for localising real economy objects in remote sensing imagery. CNNs have proven effective at localising several human-made object classes, including roads[15], buildings[16], automobiles[17], aircraft[18], and ships[19], as well as determining coal-fired power station utilisation rates[20], crop yield prediction[21], and wind power forecasting[22]. Solar PV energy installations are no exception. Imamoglu et al.[23] localise installations in Landsat 8 imagery by adding a class activation map to the shallow CNN architecture of Ishii et al.[16]. Malof et al.[24] and Camilo et al.[25] develop semantic segmentation models to localise installations by using hand-crafted features with the deep CNN SegNet[26]. Yu et al.[6] fine-tune the deep CNN VGG-16[27] and add a class activation map to segment solar PV in aerial and very-high resolution satellite imagery, then map the continental United States. Hou et al.[7] use a U-Net

model with attention, similar to our primary inference models, to map China. While these studies show promise, they are limited in geographic extent and have not released data for further study.

We develop a machine learning pipeline to localize installations and measure installation dates in two globally-available satellite imagery sources: high resolution (1.5m pixel resolution) 4-band SPOT-6/7 imagery and medium resolution (10m pixel resolution) 12-band Sentinel-2 imagery. We incorporate both Sentinel-2 and SPOT to make use of their respective advantages - the wide spectral coverage of Sentinel-2 is sensitive to the spectral signature of PV panels and the high re-visit rate enables measuring installation dates, and the high-resolution imagery from SPOT enables precise installation footprint measurements.

These two data sources are analyzed with the machine learning pipeline shown in Supplementary Figure 1. The pipeline has two stages: an initial global search designed to maximise installation recall, followed by a process to remove false positives and estimate installation dates. These stages are implemented for each imagery source on separate branches, then results are merged into a final combined dataset. This two-branch approach has significant advantages over a unified pipeline: each branch is optimized independently to maximize recall, the dramatically different revisit rates are handled separately, installation dates are measured using the full backcatalogue Sentinel-2, and a separate SPOT pipeline was required to accommodate data licensing conditions.

We perform the initial global search using custom CNN models with U-Net architectures[15,28] to predict which image pixels contain solar PV; see Figure 1. The models are trained with polygon annotations obtained primarily from crowd-sourced OpenStreetMaps (OSM) and

supplemented with negative examples and annotations drawn by the authors. The resulting dataset of 36,882 (38,541) polygons for the Sentinel-2 (SPOT) model has global coverage with the majority from Europe and the United States. See Supplementary Information for discussion.

In the second step, false candidates are eliminated from the Sentinel-2 branch using RNNs to classify timeseries derived from longitudinal inference of the backcatalogue of imagery, and from the SPOT branch using a fine-tuned Imagenet ResNet50[29] binary classifier. Each generating unit's installation date is also estimated from the backcatalogue inferences. The remaining candidates from both branches are merged and hand-verified. Figure 1 shows inference examples from the primary computer vision models.

We develop separate validation and test sets to tune the pipeline hyperparameters and evaluate end-to-end performance. For the validation set, we hand-label polygons seeded from installations in the WRI GPPD, then use this to tune the hyperparameters of the CNNs and RNNs, as well as vectorise SPOT detections. For the test set, we hand-labeled installations in 122 globally distributed sub-regions. This test set is used this to evaluate pipeline precision, recall, Jaccard index (intersection-over-union, or IoU), and to quantify facility area error. Further, we compare aggregates of our dataset against known aggregates for countries around the world - see Table 1 and Supplementary Figure 12. Supplementary Figure 12 shows that for many countries, our estimated aggregate generating capacity approximates the aggregate generating capacity reporting by IRENA[3]. Further work is now required to reduce uncertainty and bias of this bottom-up estimate to enable direct comparison to aggregate statistics and for other policy and planning purposes.

For installations over 10,000m$^2$ (approximately 600kW), we achieve precision of

98.6% relative to our test set, with a modest trade-off in recall which drops to 90%, see Supplementary Figure 6. The IOU of our final dataset for installations over 10,000m$^2$ is 90% – sufficient for the wide range of use cases that our global dataset enables. The precision of our pipeline shows covariate shift across geographies, illustrated in Supplementary Figure 7 and Supplementary Table 7. The pipeline recall is slightly superior in geographies where were better represented in the validation set, suggesting the complete pipeline might be modestly overfit to the validation set. We compare country-level aggregates of our data to IRENA[3] to develop insight into the collective exhaustion of our data, see Supplementary Figure 12. Supplementary Figure 6 and Table 1 provide for more information regarding pipeline performance.

To develop further intuition about the performance of our primary semantic segmentation models, we perform band perturbation, band dropout, and feature activation experiments on our trained models. We find that our primary inference models learn a sophisticated topology of spectral and spatial features. We show that non-visible bands contain important information for the determination of the presence of a PV installation, see Supplementary Figure 8. This suggests that a trained machine might achieve super-human performance at this task. These results and further details on the machine learning pipeline are available in the Supplementary Information and can inform the future design of computer vision models for build-environment object detection.

**A Global Census** The machine learning pipeline is deployed on the global corpus of Sentinel-2 and SPOT6/7 imagery using Descartes Labs cloud computation infrastructure. Assuming that installations will be reasonably proximate to human populations, we define the search area by dilating a global human population-density map, finding that a 7km buffer includes 99.9% of the validation set capacity. We search the resulting 72.1mn km$^2$

or 48.4% of the Earth's land surface area. The date range for our study was 2016-06-01 through 2018-09-30 and we assume that no installations have been decommissioned or relocated. Deployment processed 550 TB of imagery, used in excess of 1mn CPU-hours and 20,000 GPU-hours, and took approximately 2 months in real time.

We add properties to detected polygon to produce a feature-rich dataset. Nominal alternating-current (AC) generating capacity of installations are estimated using simple assumptions about installation tilt angle, ground coverage ratio, inverter loading ratio, and panel efficiency. Facility installation dates are measured by our pipeline using backcatalog of Sentinel-2 detections. Historical land cover classification data is added for each installation by querying a global land cover product. Political administration codes (country-level and state/province-level) are added for faster aggregation and filtering. We also add proximity to protected areas and indigenous and community lands within 10km. Where feasible, we match installations to unique identifiers from other publicly available datasets. The resulting dataset contains 68,661 detections located in 131 countries, 30% of which are complete with installation dates, 43% are matched to existing known installations in WRI's GPPD and the EIA's public data. Our new dataset expands the coverage of publicly available solar PV facility-level data by 38,852 installations, 91 countries, and an estimated 280 GW.

We observe that over the study period, estimated installed capacity of non-residential PV increases by more than 81% to 384 GW, led by increases in China (120%), India (184%), the EU-27+UK (20%), the United States (58%), Japan (119%), South Africa (19%), Thailand (15%), Chile (60%), South Korea (58%), and Turkey (143%). With over half of new installations located in China, the portion of global installed capacity in China grew from 36% to 44%. We detect spatial-temporal hotspots in the deployment of solar

PV, for example in Turkey and the Netherlands through 2017, Mexico, Hungary, and Australia through 2018, and ongoing regional deployments in the United States, India, and China. See Figure 2 for an aggregated arrangement of the dataset, Supplementary Table 10 for country-level aggregated statistics, and the Supplementary Information for a link to an interactive map of our dataset.

We compare aggregate statistics of our data to countries where data for non-residential PV installed capacity is available. Table 1 compares our measured capacity at the end of 2018 for installations >10kW against other well-known bottom-up asset-level inventories and top-down aggregate statistics. We show that with simple assumptions for installation type, tilt angle, and literature-obtained distributions for panel efficiency, ground coverage ratio, and inverter loading ratio, we obtain gross installed capacities that are comparable to top-down aggregate statistics, and far exceed the currently best-available asset-level inventories. Using area-binned recall values from our test set (see Supplementary Table 4), we obtain a best-estimate of 423 GW for global installed non-residential PV generating capacity at the end of 2018, with an estimated uncertainty of [-75GW, +77GW] with 95% confidence, which includes ground coverage ratio, inverter loading ratio, module efficiency, predicted area error, pipeline recall, and pipeline precision. This global capacity estimate approximates the non-residential PV figures published by IRENA[3] (483 GW not disambiguated with residential PV) and the IEA[10] (approximately 420 GW non-residential PV). This analysis highlights the relative advantages of these inventories: while top-down inventories estimate collectively exhaustive capacity, asset-level inventories are ultimately required to bring transparency, veracity, and feature-richness to these important statistics.

Aggregate generating capacity estimates from our bottom-up survey are subject to variation in factors such as detected geometries and and those in the uncertainty esti-

mation above. While we have endeavored to quantify this uncertainty, other sources of epistemic uncertainty remain, and so we advise data users to carefully consider uncertainty in any downstream analysis. Further work is now required to further reduce this uncertainty, including improved coverage in target geographies, adding instillations with $<$10kW, improving geometry estimation, and modelling installation-level estimates of installation type, efficiency, ground coverage, panel orientation, and inverter sizing. These tasks are aided by remote sensing imagery and machine learning, for which our dataset provides an initial training sample. See Supplementary Information for discussion of generating capacity uncertainty and further work required to develop global feature-rich asset-level PV solar energy installation data. We make our validation, test, and final predicted datasets publicly available and invite the research community to join us in answering these questions.

We match our detections to localised data from the WRI GPPD and EIA using location and estimated generating capacity. We compare our estimates of installation dates to matched data from the EIA. We find, on average, our estimated installation dates (which should correspond to the beginning of construction) predates the reported operating date of the EIA by approximately 2 months, matching our intuition and providing insight into the time delay between final investment decisions, construction, and facility operation. We conclude that our dataset provides an initial global census of commercial-, industrial-, and utility-scale solar PV installations, and can be used as a starting point for a more exhaustive, feature-rich inventory of global solar PV. See Supplementary Information for further details.

**Land Cover Analysis** With our dataset of installation geometries we are able to generate novel insight into global land cover patterns of PV solar energy sites. Land use for

renewable energy is an urgent area of study, as the land chosen for the deployment of renewable energy must navigate impacts on and tradeoffs between the costs of renewable energy transitions;[30,31] greenhouse gas emissions due to land cover and land use change;[32] ecosystem health and biodiversity;[33] water resources and food production;[34,35] indigenous and community land use;[36] land and property values;[37] and political acceptability[38]. These trade-offs are captured by the multi-criteria challenge of the SDGs themselves: while renewable energy development might be aligned with SDGs 7 (clean energy), 8 (economic growth), 9 (infrastructure), and 13 (climate action), it might have detrimental affects on SDGs 2 (zero hunger) by displacing croplands, 3 (good health) by impairing ecosystem health benefits, 10 (reduced inequality) by displacing community land use, and 15 (life on land) by impacting biodiversity. We develop insight into impact that the diffusion of PV solar energy has on land use by sampling land cover prior to installation. Our dataset brings transparency to PV solar energy land cover trends at the global scale, and can help policy makers navigate trade-offs in policy objectives at the multilateral, national, and subnational level.

We obtain pre-existing land cover for all installations in our dataset back to 2006, sampling from the European Space Agency Climate Change Initiative 300m land cover data[39]. We reduce the land cover classfication system to 8 classes, 2 anthrome classes: cropland and built-up areas; and 6 biome classes: forests, grassland, shrubland, barren land, wetlands, and other, see Supplementary Table 9. To capture the impact of PV solar energy sitings on vulnerable desert ecosystems, we also add a biome class for aridlands. Following the definition of the World Desertification Atlas,[40] aridlands are defined as a natural biome with an average aridity index less than 20%, which we calculate using ERA5 reanalysis data.[41] Figure 3 shows how land cover used for solar PV deployment has changed over the

study period, Supplementary Figure 10 shows these trends for the top 20 countries in our dataset, and Supplementary Figure 11 shows these trends for our three nominal generating capacity bins.

We observe that, globally, PV solar energy installations are most often sited on land covers indicating significant anthropogenic land use (i.e. 'anthromes'), namely croplands. The second two most frequent land covers for PV solar energy sitings indicate biological and ecosystem land use (i.e. 'biomes'): aridlands and grasslands. We compare the global land cover distribution with the distribution of land cover for PV solar energy sites, finding a significant bias towards siting on cropland. By contrast, the distribution of PV sites compared to the local distribution of land cover, where 'local' is defined as a 0.5°x0.5° grid cell, shows a bias towards grassland and desert and away from cropland, forest, and shrubland. This difference suggests that while rural anthromes are the most frequent targets for PV solar energy sitings, this might be driven by proximity to human populations more than a preference for previously-developed land. We posit that the local bias towards deserts and grasslands may be due to the reduced cost of preparing the land for PV solar energy installation. We note that siting decisions in anthromes may still cause land use and land cover changes in biomes. As an example, cropland displaced by PV solar energy may be re-developed on an biome site elsewhere. These trends vary significantly by country, see Supplementary Figure 10.

Considering European Economic Area countries, for example, we find the distribution of PV solar energy to be similarly biased towards agricultural areas in Germany, Italy, Spain, the United Kingdom, the Czech Republic, and Romania, however *within* PV-containing localities, Germany and France show a heavy bias towards built-up areas for installation, while the others further enforce the bias towards installation on croplands.

Japan and South Korea show almost no bias in the country-level distribution of PV solar energy, but at a local level show considerable bias towards cropland sites. China, the United States, India, Spain, France, South Africa, Mexico, and Chile show significant deployment of PV solar energy on aridlands, uniquely vulnerable ecosystems. Further country-level analysis can be found in the Supplementary Information and Supplementary Figure 10.

These trends highlight the different policy approaches to PV development, developer incentives, and the constraints of geography and existing land cover. As the reduction in the cost of solar PV continues to drive diffusion and adoption, policy makers must carefully consider trade-offs between food supply, ecosystem and biodiversity impacts, land protection for indigenous and community uses, and climate change mitigation. Our dataset and analysis shows the fundamental changes ongoing in the geography of energy resources and is available to help policy makers navigate these trade-offs.

**Conclusion** Our global survey of non-residential PV solar energy installations, using machine learning and remote sensing, has generated a novel public global database of 68,661 spatially-localised facility footprints with installation dates, land cover assessments, estimated generating capacity, and other metadata. This is the first time, to our knowledge, that machine learning and remote sensing has been used to search the entire planet for a specific type of infrastructure asset. We enhance the utility of our dataset with an analysis of pre-existing land cover for PV installations, and show that PV installations most commonly are sited on croplands, followed by deserts and grasslands. Future work to enhance the dataset should include temporal updates, improved installation geometries, better estimates of installation type, and its combination with residential solar PV datasets. Opportunities for subsequent analysis might include solar PV forecasting and now-casting, further land use and land cover impact studies, future infrastructure planning pathways, socioeconomic

spatial-temporal diffusion studies, and policy-targeted counterfactual scenarios. We are encouraged by the success of our pipeline and our use of noisy crowd-sourced training data, and suggest that our method may be applied elsewhere to stimulate the creation of robust public asset-level data. We acknowledge the fundamental public-goods nature of such asset-level data and make dataset publicly available to facilitate future research.

Figure 1: **Solar PV facilities are detected in remote sensing imagery with machine learning**. Here, we show out-of-sample examples showing SPOT6/7 and Sentinel-2 optical imagery, corresponding U-Net prediction maps, and vectorised output. A) a "panda" array in China (lat/lon: 39.98,113.48); B) covered hills in China (34.34,113.38); C) a new array detected in Sentinel-2 but missing from SPOT6/7 in the US (35.46,-79.18); and D) a 120MW array in South Africa (-27.58,22.93). Imagery attributions: Copernicus Sentinel data 2018; ©AIRBUS DS (2018).

Figure 2: **Aggregated arrangement of the global dataset.** Generating capacity by installation date is shown for major countries. Note the last period is only three months.

Figure 3: **Pre-existing land cover for new solar PV installations**. Panels show the time series of installations (b); the distribution of installation sizes by land cover (c); local bias (d) between PV land cover and local land covers, where the local local land covers are those lying within the same 0.5°x0.5° grid pixel; and global bias (e) between the global land cover distribution and the distribution of all PV-containing pixels. A positive bias indicates PV is preferentially installed on this land cover type; a negative bias indicates this land cover type is avoided. Country-level analysis is available in Supplementary Figure 10.

[H]

Table 1: **Comparison of prominent asset-level and aggregated datasets**

| (GW, $N_{inst.}$) | Aggregate | Asset-Level Inventories | | | |
|---|---|---|---|---|---|
| Country | IRENA, IEA[3, 10 A] | IHS[9] | WEPP[8 B] | GPPD[4] | Ours (% improvement)[C] |
| Global | ~420 (n/a) | 90.8 (4,004) | 107.4 (12,915) | 54.3 (5,289) | 384.7 (68,661, +258%) |
| China | 173.5 (n/a) | 28.9 (436) | 13.9 (411) | 0.0 (0) | 167.4 (18,449, +479%) |
| Japan | 46.5 (n/a) | 5.7 (411) | 6.0 (905) | 2.0 (131) | 18.0 (10,504, +200%) |
| United States | 40.2 (n/a) | 14.0 (468) | 27.1 (3,522) | 21.4 (1,790) | 54.1 (7,639, +100%) |
| Germany | 38.7 (n/a) | 3.7 (380) | 6.0 (1,388) | 3.9 (408) | 16.5 (4,702, +175%) |
| Italy | 15.9 (n/a) | 1.4 (179) | 2.8 (1,079) | 0.4 (74) | 12.6 (5,796, +350%) |

A: Solar PV installations >10kW at end 2018.
B: Latest release in March 2018, includes operating plants and planned or in-construction plants for 2018.
C: Assumes mean ground coverage ratio, inverter loading ratio, panel efficiency, and predicted area error. See Supplementary Table 6 for details. Does not include adjustment for pipeline precision and recall (i.e. Supplementary Information Equation 2). Percent improvement relative to generating capacity of most complete dataset.

## Methods

### Energy and Emissions Footprint

Our machine learning pipeline was deployed on computation infrastructure powered by net-zero carbon electricity. The majority of the training was powered by net-zero carbon electricity, the Sentinel-2 model was trained on average generation mix electricity in the Eastern United States. We calculate training and deployment consumed approximately 71 MWh but emitted only 14kg of $CO_2$eq in greenhouse gas emissions. The energy footprint of our study is approximately the same as the energy required to drive an electric vehicle the distance between the Earth and moon. The carbon footprint of our study is approximately the same as driving a petrol vehicle from New York City to Philadelphia.

### Data Availability

The dataset is publicly hosted on Zenodo and is available with DOI 10.5281/zenodo.5005868. It will also be visualised and available for download via the World Resources Institute ResourceWatch, and the Descartes Labs platform.

### Code Availability

The code repository is publicly hosted on Github at https://github.com/Lkruitwagen/solar-pv-global-inventory. The code release for this publication is version 1.0.0 and is also hosted on Zenodo with DOI 10.5281/zenodo.5045001.

### Author Contributions, Information, and Competing Interests Statement

### Author Contributions

L.K. designed and implemented the machine learning pipeline, designed and implemented the dataset analysis, and wrote the paper draft. K.S. designed the machine learning pipeline

and dataset analysis, implemented the SPOT6/7 branch of the machine learning pipeline, and wrote the paper draft. J.F. and L.B. contributed to the acquisition of training data and analysis of the dataset. S.S. contributed to the deployment of the machine learning pipeline. C.H. contributed to the analysis of the dataset and wrote the paper draft.

## Author Information

Correspondence and requests for information, materials, reprints, or permissions should be sent to the corresponding author, Lucas Kruitwagen, via email: lucas.kruitwagen@smithschool.ox.ac.uk.

## Competing Interests Statement

Authors Kyle Story and Sam Skillman are employees and shareholders of Descartes Labs Inc., the company which builds and maintains the cloud computation infrastructure used to conduct this research. Authors Johannes Friedrich and Logan Byers are employees of the World Resources Institute, a not-for-profit organisation which will host and publicly visualise a copy of our dataset.

1. BP plc. Statistical Review of World Energy 2018. Tech. Rep., London, UK (2018).

2. International Energy Agency. World Energy Outlook 2018. Tech. Rep., Paris, France (2018).

3. International Renewable Energy Agency. Renewable capacity statistics 2019. Tech. Rep., Abu Dhabi (2019).

4. Byers, L. *et al.* A Global Database of Power Plants. Tech. Rep., Washington DC, USA (2018).

5. Barbose, G. & Darghouth, N. Tracking the sun (2019). URL https://openpv.nrel.gov/.

6. Yu, J., Wang, Z., Majumdar, A. & Rajagopal, R. Deepsolar: A machine learning framework to efficiently construct a solar deployment database in the united states. *Joule* **2**, 2605 – 2617 (2018).

7. Hou, X. *et al.* Solarnet: A deep learning framework to map solar plants in china from satellite imagery. In *Climate Change AI Workshop, ICLR2020* (ICLR, 2020).

8. Platts, S. G. World electric power plant database (2018). URL https://www.spglobal.com/platts/en/products-services/electric-power/w

9. IHSMarkit. Electric plants (2020). URL https://catalogue.datalake.ihsmarkit.com/.

10. International Energy Agency. Renewables 2019. Tech. Rep., Paris, France (2019). URL https://www.iea.org/reports/renewables-2019.

11. Bolinger, M., Seel, J. & Robson, D. Utility-scale solar: Empirical trends in project technology, cost, performance, and ppa pricing in the united states–2019 edition (2019).

12. Fukushima, K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics* **36**, 193–202 (1980).

13. LeCun, Y. *et al.* Backpropagation applied to handwritten zip code recognition. *Neural computation* **1**, 541–551 (1989).

14. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, 1097–1105 (Curran Associates Inc., USA, 2012).

15. Zhang, Z., Liu, Q. & Wang, Y. Road extraction by deep residual u-net. *IEEE Geoscience and Remote Sensing Letters* **15**, 749–753 (2018).

16. Ishii, T. *et al.* Detection by classification of buildings in multispectral satellite imagery. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, 3344–3349 (IEEE, 2016).

17. Audebert, N., Le Saux, B. & Lefèvre, S. Beyond rgb: Very high resolution urban remote sensing with multimodal deep networks. *ISPRS Journal of Photogrammetry and Remote Sensing* **140**, 20–32 (2018).

18. Zuo, J., Xu, G., Fu, K., Sun, X. & Sun, H. Aircraft type recognition based on segmentation with deep convolutional neural networks. *IEEE Geoscience and Remote Sensing Letters* **15**, 282–286 (2018).

19. Bentes, C., Velotto, D. & Tings, B. Ship classification in terrasar-x images with convolutional neural networks. *IEEE Journal of Oceanic Engineering* **43**, 258–266 (2018).

20. Gray, M., Watson, L., Ljungwaldh, S. & Morris, E. Nowhere to hide: Using satellite imagery to estimate the utilisation of fossil fuel power plants. Tech. Rep., London, UK (2018). URL https://www.carbontracker.org/reports/nowhere-to-hide/.

21. Wang, A. X., Tran, C., Desai, N., Lobell, D. & Ermon, S. Deep transfer learning for crop yield prediction with remote sensing data. In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies*, 50 (ACM, 2018).

22. Wang, H. *et al.* Deep learning based ensemble approach for probabilistic wind power forecasting. *Applied Energy* **188**, 56 – 70 (2017).

23. Imamoglu, N., Kimura, M., Miyamoto, H., Fujita, A. & Nakamura, R. Solar power plant detection on multi-spectral satellite imagery using weakly-supervised cnn with feedback features and m-pcnn fusion. *arXiv preprint arXiv:1704.06410* (2017).

24. Malof, J. M., Bradbury, K., Collins, L. M. & Newell, R. G. Automatic detection of solar photovoltaic arrays in high resolution aerial imagery. *Applied Energy* **183**, 229 – 240 (2016). URL http://www.sciencedirect.com/science/article/pii/S0306261916313009.

25. Camilo, J. A., Wang, R., Collins, L. M., Bradbury, K. & Malof, J. M. Application of a semantic segmentation convolutional neural network for accurate automatic detection and mapping of solar photovoltaic arrays in aerial imagery. *CoRR* **abs/1801.04018** (2018). 1801.04018.

26. Badrinarayanan, V., Kendall, A. & Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *CoRR* **abs/1511.00561** (2015). 1511.00561.

27. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations* (2015).

28. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. *CoRR* **abs/1505.04597** (2015). 1505.04597.

29. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. *CoRR* **abs/1512.03385** (2015). 1512.03385.

30. Wu, G. C. *et al.* Power of Place: Land Conservation and Clean Energy Pathways for California. Tech. Rep. (2019).

31. Konadu, D. *et al.* Land use implications of future energy system trajectoriesthe case of the uk 2050 carbon plan. *Energy Policy* **86**, 328 – 337 (2015).

32. Turconi, R., Boldrin, A. & Astrup, T. Life cycle assessment (lca) of electricity generation technologies: Overview, comparability and limitations. *Renewable and Sustainable Energy Reviews* **28**, 555 – 565 (2013).

433 33. Hernandez, R. *et al.* Environmental impacts of utility-scale solar energy. *Renewable*
434 *and Sustainable Energy Reviews* **29**, 766 – 779 (2014).

435 34. Bukhary, S., Ahmad, S. & Batista, J. Analyzing land and water requirements for
436 solar deployment in the southwestern united states. *Renewable and Sustainable Energy*
437 *Reviews* **82**, 3288 – 3305 (2018).

438 35. Dias, L., Gouveia, J. P., Loureno, P. & Seixas, J. Interplay between the potential
439 of photovoltaic systems and agricultural land use. *Land Use Policy* **81**, pp725–735
440 (2019).

441 36. Grodsky, S. M. & Hernandez, R. R. Reduced ecosystem services of desert plants from
442 ground-mounted solar energy development. *Nature Sustainability* 1–8 (2020).

443 37. Carlisle, J. E., Kane, S. L., Solan, D., Bowman, M. & Joe, J. C. Public attitudes
444 regarding large-scale solar energy development in the u.s. *Renewable and Sustainable*
445 *Energy Reviews* **48**, 835–847 (2015).

446 38. Mulvaney, D. Identifying the roots of green civil war over utility-scale solar energy
447 projects on public lands across the american southwest. *Journal of Land Use Science*
448 **12**, 493–515 (2017).

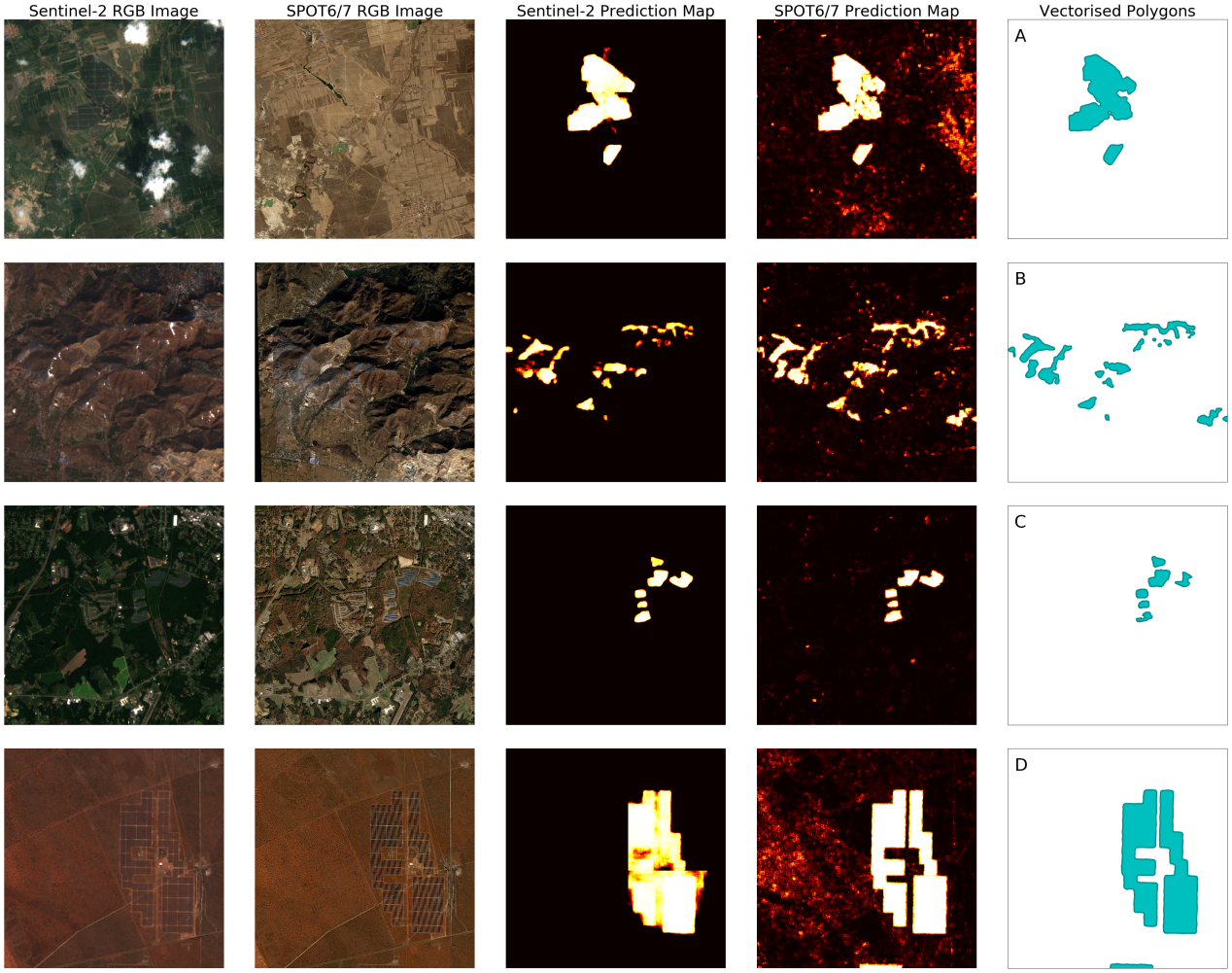
449 39. Lamarche, C. *et al.* Compilation and validation of sar and optical data products for
450 a complete and global map of inland/ocean water tailored to the climate modeling
451 community. *Remote Sensing* **9**, 36 (2017).

452 40. Cherlet, M. *et al. World atlas of desertification: Rethinking land degradation and*
453 *sustainable land management* (Publications Office of the European Union, 2018).

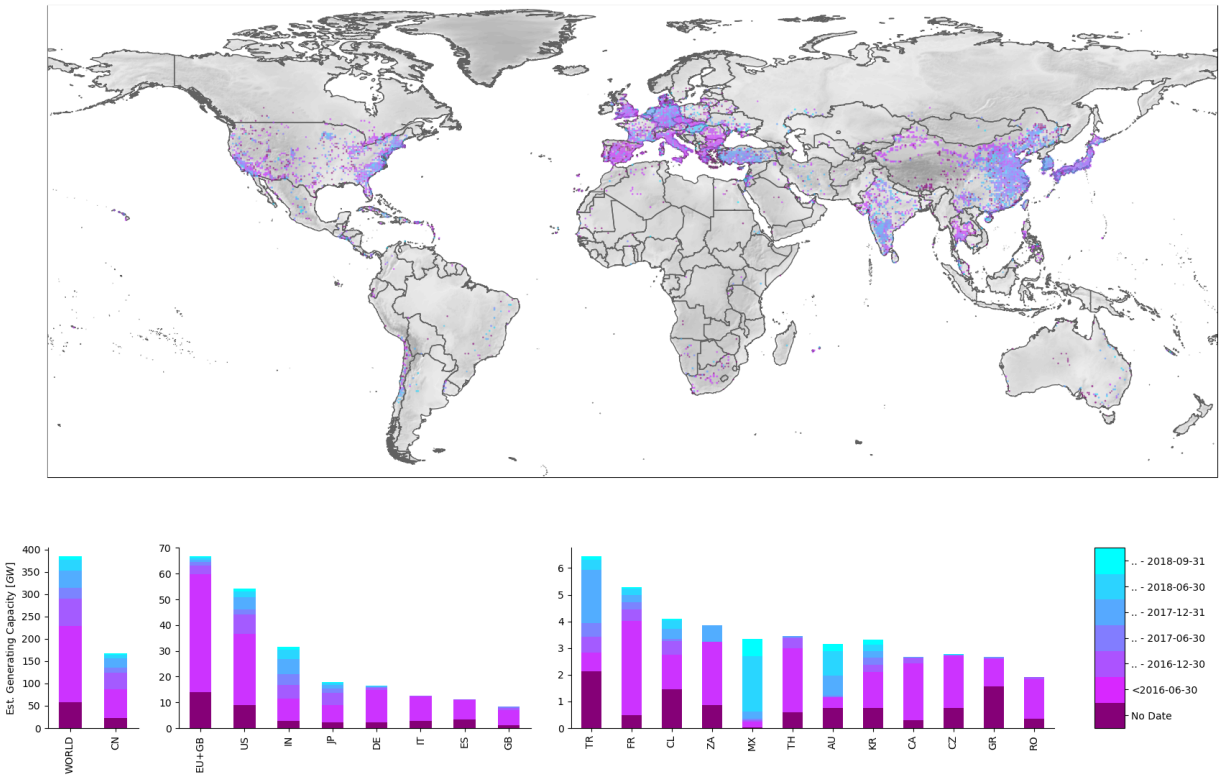
454 41. Hersbach, H. *et al.* The era5 global reanalysis. *Quarterly Journal of the Royal Meteo-*

455 *rological Society* **146**, 1999–2049 (2020).

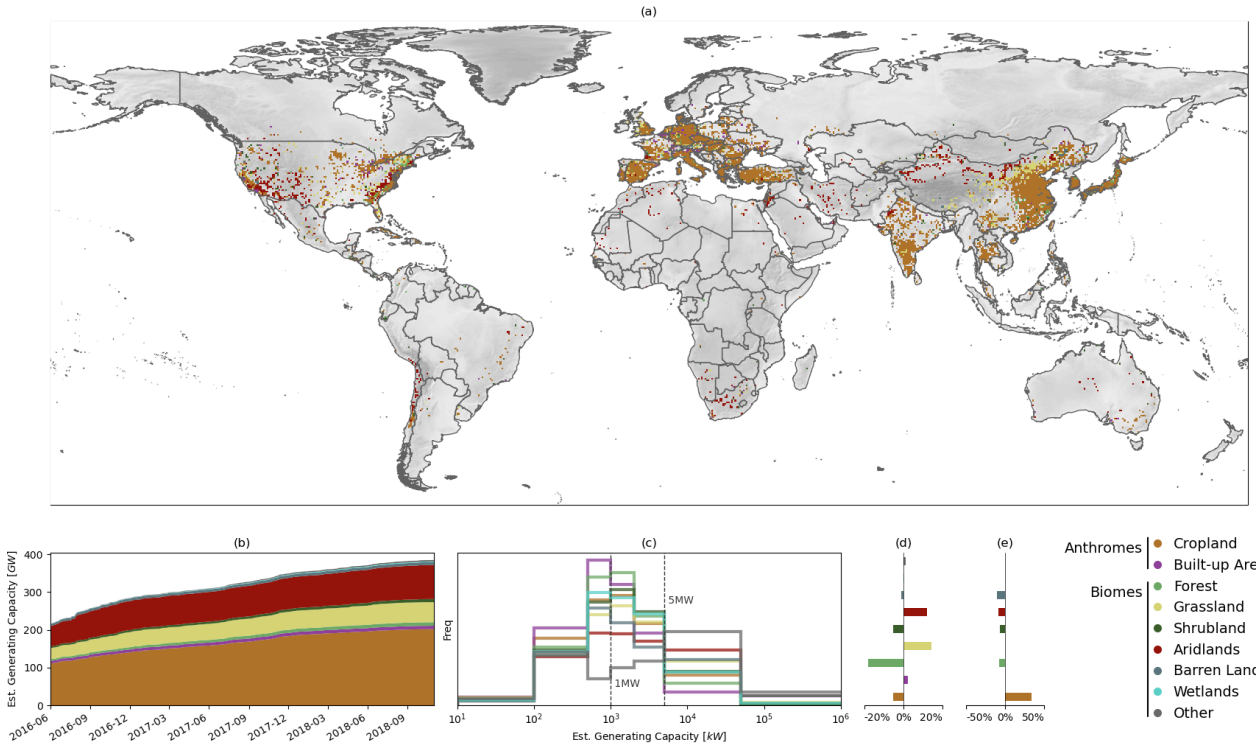
## Figure 1



## Figure 2




## Figure 3



# A global inventory of photovoltaic solar generating units

Supplementary Information

Kruitwagen, L.,[1,*] Story, K.,[2] Friedrich, J.,[3] Byers, L.,[3] Skillman, S.,[2] & Hepburn, C.[1]

[1] *Smith School of Enterprise and the Environment, University of Oxford, Oxford, United Kingdom*

[2] *Descartes Labs Inc., San Francisco, CA, United States*

[3] *World Resources Institute, Washington, DC, United States*

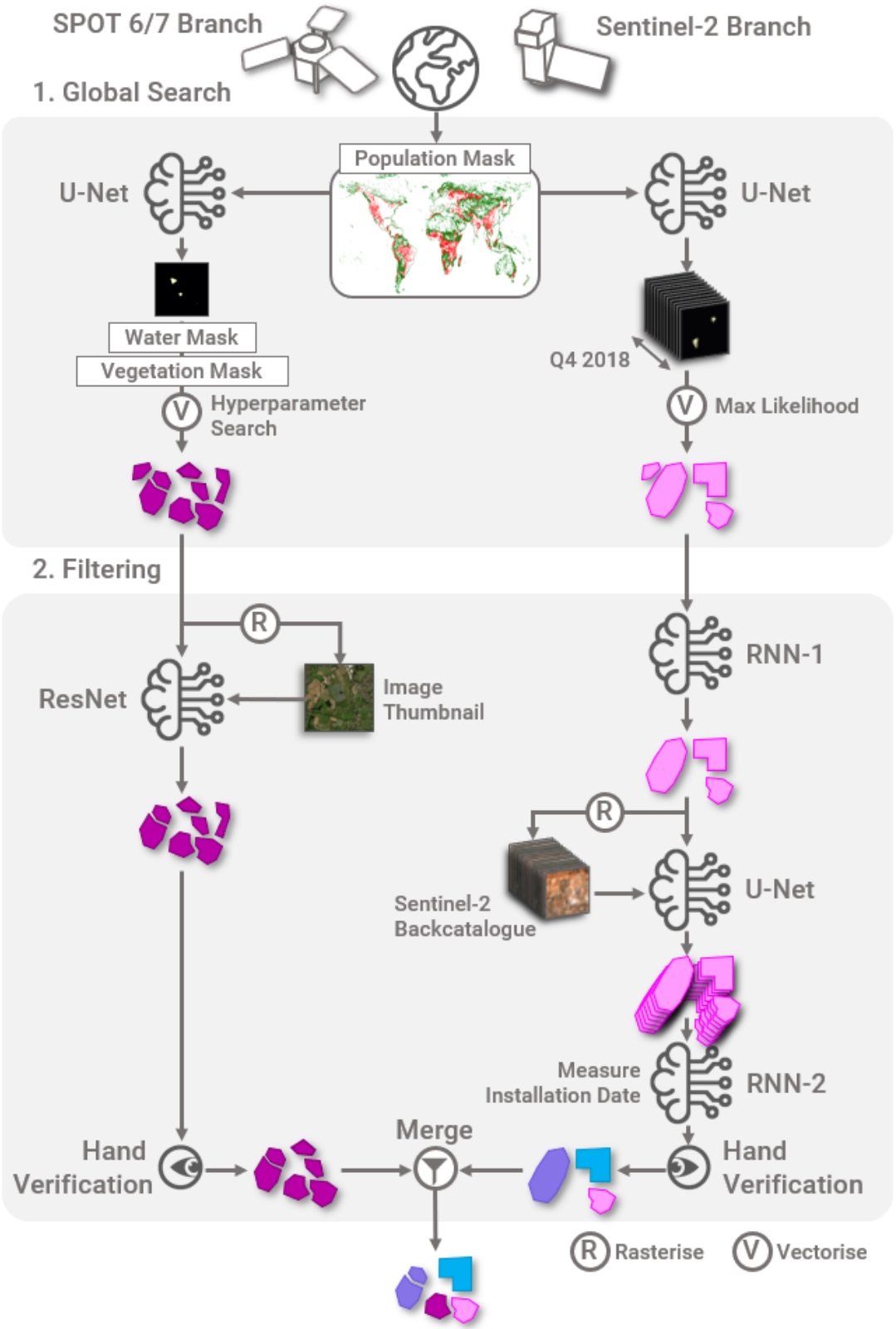
** Corresponding Author: lucas.kruitwagen@smithschool.ox.ac.uk*

## 1 Machine Learning Pipeline Description

Our pipeline for detecting solar PV installations in SPOT and S2 remote sensing imagery and estimating their installation dates uses a combination of deep-learning models, model-stacking, and human-in-the-loop techniques which are shown schematically in Supplementary Figure 1. Supplementary Table 1 summarises the machine learning models used in the pipeline and their training and validation data.

The pipeline features a shared corpus of training data, a spatial mask to limit the deployment area, and two separate branches, one for each primary sensing platform employed. The development of two separate branches allows us to employ the relative advantages and mitigate the relative downsides of both sensing platforms.

**Data Description** Solar PV installations are detected in both Sentinel-2 and SPOT-6/7 multipectral remote sensing imagery. Sentinel-2 is a multispectral Earth observation mission of the European Space Agency and European commission. The platform is designed for interoperability with other multispectral remote sensing missions, and provides imagery consisting of 12 bands ranging from ultraviolet, through the visible spectra, to short wave infrared. The maximum ground resolution of Sentinel-2 is 10m for red, green, blue, and near-infrared bands, increasing to 60m for narrow-bandwidth ultra-violet and short-wave infrared bands. Sentinel-2 data were corrected for effects of the atmosphere using the Descartes Labs Surface Reflectance (DLSR) algorithm. DLSR is a spectral processing model that corrects pixel intensity for each spectral band to remove the effects of the Earth's atmosphere for each scene.




Supplementary Figure 1: **Machine learning inference pipeline diagram.** The SPOT6/7 and S2 branches are deployed separately and merged together prior to hand verification. The global search first maximises pipeline recall, and then subsequent filtering stages eliminate false positives. Colored polygons illustrate the features of the facility geometries after each step, including their source branch at the end of the global search, the timeseries of geometries after running inference over the S2 backcatalog, and their acquisition of installation dates after RNN-2. Imagery attributions: Copernicus Sentinel data 2018; ©AIRBUS DS (2018).

The SPOT-6/7 satellites, operated by Airbus, produce a global composite of images that are selected to minimize cloud cover and maintain seasonal consistency. The four-band images (red, green, blue, and near-infrared) are pan-sharpened from 6m to 1.5m, then enhanced to adjust luminosity, contrast, and color balance[42]. Image quality and acquisition date is not consistent across the globe. The most recent available image up to December 31, 2018 was used for each location.

Training sets for both the S2 and SPOT solar PV prediction models were developed from OpenStreetMap (OSM) data. The polygon labels are precise (i.e. do not contain false positives) but are noisy due to the crowd-sourced nature of OSM data. The geometries correspond to different conventions for labelling the footprint of a PV solar energy facility. Drawing a small sample, we identify that 9% annotate the 'total area' as defined by Ong, S. et al. (2013)[43] and Hernandez, R. R. et al. (2014)[44] (i.e. the full area enclosed by the site boundary), 18% annotate the 'direct area' (i.e. the area covered by the solar arrays, land in between them, and supporting equipment), and 73% annotate the 'array area' (i.e. the area directly under the PV modules). Sets of imagery samples $4\text{km}^2$($200^2$ pixels) and $0.6\text{km}^2$ ($512^2$ pixels) for S2 and SPOT respectively were generated capturing the complete OSM polygon set. The size distribution of training polygons is shown in Supplementary Figure 2. Pixel labels, i.e. single-band class-map images, were rasterised from OSM polygons intersecting tile geometries, converted to the pixel coordinate system. For the S2 training, an additional 11,862 negative samples were randomly generated from points on the Earth's surface area not intersecting any existing tiles. A final set of 4219 'bootstrapped' tiles was added to provide hard-negatives and positives based on early model deployments over the UK, China, India, California, and North Carolina.

For the SPOT6/7 training data, a number of hand-labeled polygons were added in

China, giving a resulting dataset of 38,541 polygons of positive training examples, with 67% in Europe, 28% in the United States, and 5% in Asia. Training imagery was developed from these polygons and hard-negative samples 'bootstrapped' from early iterations of the model.

A validation dataset was developed from hand-labelled polygons seeded by latitude-longitude locations of solar PV generating stations in the World Resources Institute Global Power Plant Database[4]. The validation set locations were chosen randomly to provide a geographically-diverse sample. The most represented countries in the validation set include Japan (%28), India (%24), the United States (%23), and the United Kingdom (%10). Tiles were generated for the hand-labelled polygons. Tile imagery was then inspected for any additional solar PV installations visible in that tile, which were subsequently labeled. Both S2 imagery and very-high resolution imagery from the Google Maps basemap was inspected. The S2 imagery inspected was drawn from the final three months of the study (i.e. October through December 2018) to ensure temporal accuracy of the validation set. All PV solar energy installations were labelled in the validation set, including residential facilities, to test how small a facility could be detected in SPOT 6/7 and S2 imagery.

An out-of-sample test set was developed to evaluate machine learning pipeline and recall. The test set was generated by hand-labelling large regions-of-interest for all commercial-, industrial-, and utility-scale PV solar energy facilities. A region-of-interest test set method has been used in other machine learning studies [45,46] where the true distribution of the target for prediction is unknown. Regions were selected in countries with large aggregate amounts of non-residental PV solar energy generating capacity, and were chosen to represent the diversity of both the Earth's physical and human geography. A total of 122 rectangular regions-of-interest were selected from 40 countries, totalling 538,000 $km^2$ and

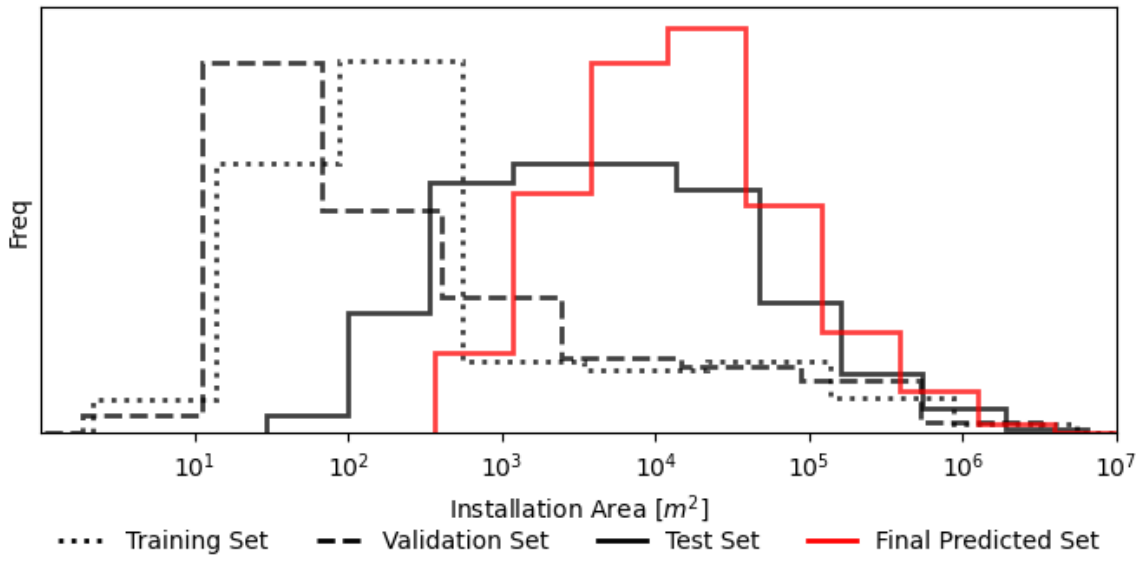
yielding 7,060 non-rsidential PV solar energy facilities. As with the validation set, polygon geometries of facility footprints were labelled in date-of-study S2 imagery and very-high basemap imagery. Country-level statistics for the test set are shown in Supplementary Table 7 and a map showing selected regions-of-interest in shown in Supplementary Figure 13.

PV solar energy facilities in the validation and test set were both labelled according to the 'direct area' land use convention. This allows the predicted facility footprints to be used for calculations of the facility nominal generating capacity, using the ground coverage ratio, inverter loading ratio, and panel efficiency. The test set can be used to determine the error between the predicted and true 'direct area' of the PV solar energy facility, allowing the areas to be corrected in the calculation of nominal generating capacity.

The size distributions of the training, validation, test, and predicted set polygons are shown in Supplementary Figure 2. The training and validation datasets are shown including the small installations less than 10kW. The training and validation datasets, drawn from OSM and the Global Power Plant Database respectively, are also geographically biased to the Global North - predominantly the United States and Europe - which also explains some of their skew to smaller installations. The test and predicted datasets are more geographically representative. The predicted dataset skews to larger facilities due to the coarse resolution of the remote sensing imagery causing weaker performance of the machine learning pipeline on smaller facilities. This aligns with our expectation that both spatial texture and shape information is required to detect PV solar energy facilities, along with spectral information. We proceed with the assumption that the validation and test sets capture the ground-truth of non-residential PV solar energy generating station 'direct area' footprints in the labelled regions at the end of Q4 2018.

Supplementary Table 1: Summary of machine learning models

| Model | Description |
|---|---|
| S2 branch U-Net | Semantic segmentation of facility footprints. Trained with 22,087 200px images from OSM training data and bootstrapping. Validated with 535 200px images generated from validation set. |
| SPOT branch U-Net | Semantic segmentation of facility footprints. Trained with 9,846 512px images from OSM training data and bootstrapping. Validated with 222 512px images generated from validation set. |
| SPOT branch ResNet | Image classifier for filtering false positives. Trained with 9,846 512px images. Validated with 222 512px images generated from validation set. |
| RNN-1 | Sequence model for filtering false positives. Trained with 4,679 bootstrapped samples. Validated with the validation set. |
| RNN-2 | Sequence model for filtering false positives. Trained with 3,057 bootstrapped samples. Validated with the validation set. |

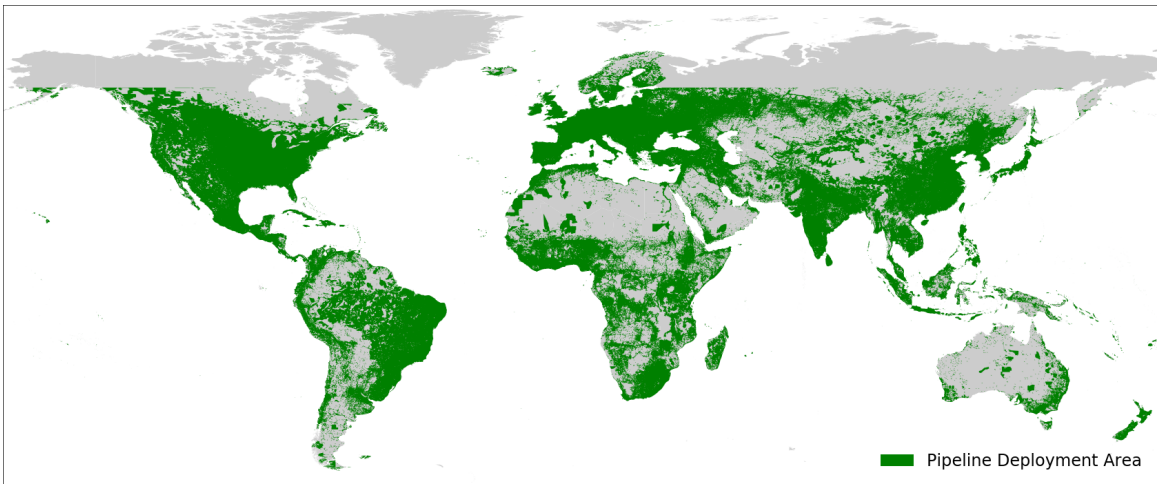



Supplementary Figure 2: **Distribution of polygon areas in training, validation, test, and predicted datasets.** The training and validation datasets are shown prior to filtering small installations. The machine learning pipeline has higher accuracy for larger PV solar energy facilities, causing a skew in the predicted dataset to larger facilities.

**Primary Semantic Segmentation Models** Two semantic segmentation models were trained as primary computer vision models to accept remote sensing imagery as inputs and to output a per-pixel class prediciton map. The U-Net architecture was chosen for its proven performance on similar problems with relatively small training corpuses, few output classes, and fixed-perspective imagery.[28] One model each was trained for the Sentinel-2 and SPOT6/7 imagery respectively. The Sentinel-2 model also incorporated the residual innovations of He et al,[29] adapted to U-Net by Zhang[47].

The Sentinel-2 Model was trained on a corpus of 22,087 200x200 images at 10m resolution. The SPOT6/7 model was trained on a corpus of 9,846 512x512px images at 1.5m resolution. Images were randomly augmented with flips and rotations, 10% additive noise, and 10% multiplicative noise. Both models were trained on NVIDIA Tesla P100 GPUs, on commercial cloud computing platforms. The SPOT6/7 model was trained for 82 epochs, taking approximately 9 hours; the Sentinel-2 model was trained for 150 epochs, taking approximately 28 hours. All machine learning models were trained with binary cross-entropy loss and Adam optimizers[48].

**Deployment Area** The computer vision models were deployed to a global corpus of remote sensing imagery. To sensibly narrow the search area, only land area in close proximity to human population was searched. This was accomplished by first making a binary raster mask of where population density exceeds 1 person per $km^2$ derived from a global high-resolution population density map taken from the EU Joint Research Council GHS Population Grid (GHSL)[49]. This mask was dilated until it included 99.9% of pixels from the validation polygons, a 7km dilation. We remove certain very large but low-density population areas from Australia and Canada, artefacts of the coarse census data for those areas used to construct the original population raster. We also clip the search area to above 60°




Supplementary Figure 3: textbfDeployment mask used to limit search area On-scope areas are indicated in green. Land area with a population density of greater than 1 person per $km^2$ was searched, plus a dialation of 7km.

North for the United States, Canada, and Russia, very large land areas with coarse census data where no PV is expected. Using the Descartes Labs tile server to tile the Earth, we search the set of image tiles that completely cover this dilated mask, an area of more than 72.1mn $km^2$, 48.4% of the land surface on Earth. We do not exclude inland water bodies and the tile generator also extents slightly beyond the coast geometry, so we also expect to include any near-land water-born 'floatovoltaic' installations. Supplementary Figure 3 shows this search area.

**Sentinel-2 Branch** The Sentinel-2 branch was designed to identify PV solar energy in sentinel-2 imagery with high confidence, and to identify the installation date of each facility using the extensive back-catalogue of imagery. The pipeline includes three key steps using three deep learning models which we now describe; see Supplementary Figure 1.

In step 1, the primary segmentation model (U-Net) was run over all S2 scenes in the fourth quarter of 2018 (at most 15 observations for any given location), see "U-Net" in the *Global Search* box of Supplementary Figure 1. These multiple looks improve detection performance by normalizing for cloud coverage and atmospheric conditions, among

other things. The 3-dimensional prediction stack is reduced to two dimensions by taking the maximum prediction value along the stack for each pixel. This maximum-prediction map is then segmented into separate installation masks using a watershed algorithm. For each separate identified installation, a scalar timeseries is produced by spatially reducing each scene of the 3-dimensional prediction stack using the maximum-prediction installation mask. Two spatial reduction functions are used: the mean of the unmasked pixel values and the count of the unmasked pixels exceeding a threshold value normalised by the total unmasked pixel count of the scene-wise stack maximum. The result is a two-feature timeseries of at most 15 observations associated with each installation detection. The original maximum-prediction installation masks are vectorised into polygons for further processing.

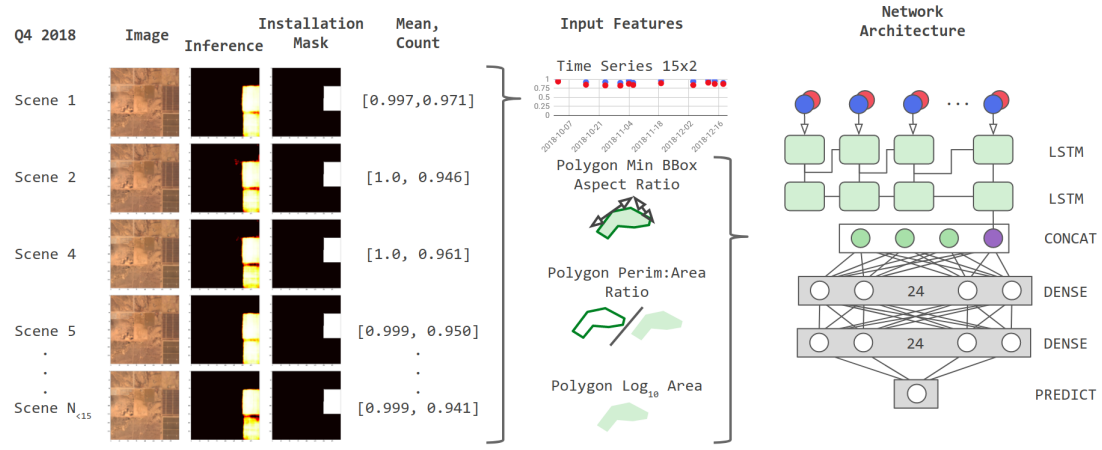
The second step uses an RNN filter model to separate true detected installations from false positives, see "RNN-1" in Supplementary Figure 1. An RNN model is required because they are capable of handling variable size inputs, like the timeseries above. RNN-1 first encodes the scalar timeseries using a Long Short-Term Memory (LSTM) neural network. We use an LSTM architecture to take advantage of network's feedback connections, recognising that signals in the timeseries are confounded by the intermittent and cyclic nature of remote sensing data availability and atmospheric conditions. The LSTM embedding is then concatenated with several hand-crafted geometric features, see Supplementary Figure 4 and Supplementary Table 2 for architecture details. The geometric features chosen are the detection's minimum rotated bounding-box aspect ratio (i.e. length-to-width ratio), the detection's perimeter-to-area ratio, and the base-10 logarithm of the detection's area in $m^2$. These features capture, respectively, the overall shape of the installation, the 'roughness' of the installation boundary, and the size of the installation. This heuristical knowledge is introduced to increase pipeline precision. The combined RNN-1 model predicts if candidates

are true installations.

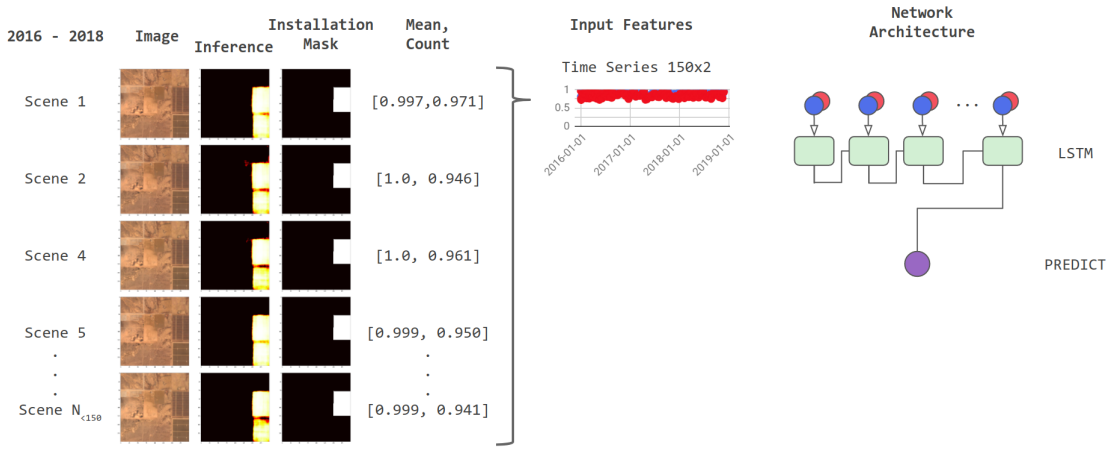
The third step uses the entire Sentinel-2 catalogue, from 2016-01-01 through to 2018-12-31, to further eliminate false positives. For each installation found in step 2, we run the primary U-Net model over all available scenes, resulting in a 3-dimensional prediction stack with up to 160 observations; see "U-Net" in the *Filtering* box of Supplementary Figure 1. The resulting timeseries of scene-wise relative coverages and means for each installation is fed into a final RNN filter (see "RNN-2" in Supplementary Figure 1 Supplementary Figure 5, and Supplementary Table 3 for architecture details) to determine whether the detection is a true or false positive. This second RNN increases pipeline precision by further mitigating remote sensing intermittency and atmospheric noise and cyclicality. The longer scope of this RNN also allows it to mitigate for seasonality. An RNN is again used for this step to manage the variable size of the input, and an LSTM is chosen for its feedback connections. The purpose of RNN-2 is again to predict if candidates are true installations. RNN-1 and -2 are separated to reduce the excessive computational requirements of running UNet inference on the entire Sentinel-2 backcatalog.

For true detections, we heuristically obtain the installation date from the final timeseries which is the input to RNN-2. We estimate the installation date from the first scene in the timeseries where the relative coverage exceeds 20%.

**SPOT6/7 Branch** The SPOT6/7 branch of the pipeline was designed to produce accurate installation geometries, taking advantage of the imagery's higher spatial resolution. The probability raster output from the U-Net model is fed into two watershed algorithms that eliminate false positives and separate individual installations. The first watershed includes lower- and upper-bound filters, as well as a gaussian filter, whose hyperparamters were




Supplementary Figure 4: **RNN-1 schematic diagram.** RNN-1 filters candidates from the primary U-Net model to remove false-positives.




Supplementary Figure 5: **RNN-2 schematic diagram.** RNN-2 is used to further eliminate false positives from candidates that pass RNN-1, now drawing on the full Sentinel-2 corpus.

Supplementary Table 2: RNN-1 Architecture

| Layer | Name | Architecture | Output Shape | Parameters |
|---|---|---|---|---|
| 0 | Timeseries_Mean_Precision | INPUT | [ : , 15 , 2 ] | 0 |
| 1 | LSTM_0 | LSTM | [ : , 15 , 8 ] | 352 |
| 2 | LSTM_1 | LSTM | [ : , 8 ] | 544 |
| 3 | Min_BBox_Aspect_Ratio | INPUT | [ : , 1 ] | 0 |
| 4 | Area_Perimeter_Ratio | INPUT | [ : , 1 ] | 0 |
| 5 | Log10_Area | INPUT | [ : , 1 ] | 0 |
| 6 | Dense_0 | DENSE | [ : , 24 ] | 288 |
| 7 | Activation_0 | RELU | [ : , 24 ] | 0 |
| 8 | Dropout_0 | DROPOUT | [ : , 24 ] | 0 |
| 9 | Dense | DENSE | [ : , 24 ] | 600 |
| 10 | Activation_1 | RELU | [ : , 24 ] | 0 |
| 11 | Dropout_1 | DROPOUT | [ : , 24 ] | 0 |
| 12 | Dense_2 | DENSE | [ : , 24 ] | 25 |
| 13 | Activation_2 | SIGMOID | [ : , 24 ] | 0 |
| | | | **Total Parameters:** | 1,809 |

Supplementary Table 3: RNN-2 Architecture

| Layer | Name | Architecture | Output Shape | Parameters |
|---|---|---|---|---|
| 0 | Timeseries_Mean_Precision | INPUT | [ : , 160 , 2 ] | 0 |
| 1 | LSTM_0 | LSTM | [ : , 12 ] | 720 |
| 2 | Dense_0 | DENSE | [ : , 1 ] | 13 |
| 3 | Activation_0 | SIGMOID | [ : , 1 ] | 0 |
| | | | **Total Parameters:** | **733** |

chosen using Bayesian search. The Bayesian search identified hyperparameters which optimised the sum of precision and recall when applied against the validation set of polygons. The second watershed algorithm identified spatial extents of individual installations using local maxima seeded from the first algorithm. Hyperparameters for the second watershed are chosen again using Bayesian search optimising for Jaccard Index (IoU) against the validation set. An edge detection algorithm was used to vectorise the watershed output, which was then transformed from the pixel coordinate space to latitude and longitude.

A single-class classifier with a ResNet-50[29] architecture was used to further eliminate false positives from the SPOT6/7 detections. The ResNet-50 was fine-tuned from ImageNet weights using sample images drawn from the initial training set.

**Dataset Completion** Sentinel-2 and SPOT pipeline branches are combined into a final vector dataset using a rules-based filter. Where Sentinel-2 and SPOT polygons intersect with a Jaccard index (Intersection-over-union) in excess of 30%, the geometry of the SPOT polygon is retained, inheriting the installation date from the Sentinel-2 detection (confidence level "A"). Detections from only the SPOT and S2 branches are retained with confidence levels "B" and "C" respectively. Where the IoU does not exceed 30%, the union of both geometries are retained, inheriting the S2 installation date (confidence "D"). The distribution of each confidence level is shown in Supplementary Table 5.

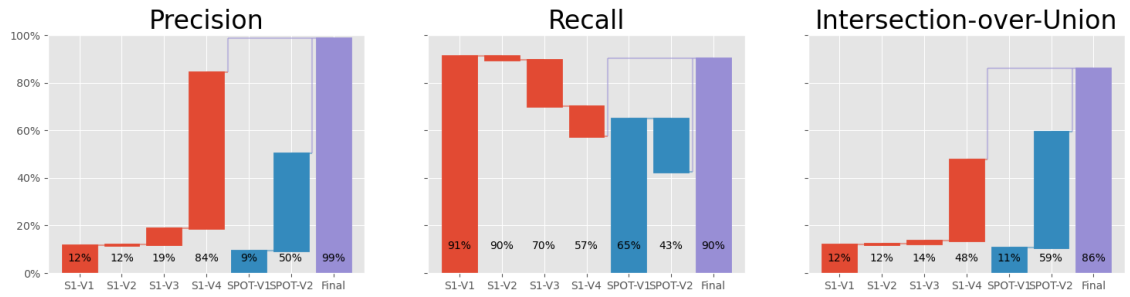
As a final check of the dataset, the merged detections were hand-verified by the authors. We provide this final step in order to present a clean, high-quality dataset for downstream applications. Detection geometries were not changed in final hand verification. The authors inspected all the samples in both very-high resolution satellite imagery and up-to-date medium resolution imagery to verify the presence of PV solar energy at the location

of the detection. Labelling tools were developed to rapidly serve sample imagery to the human annotators. The annotators could label approximately 3000 samples per hour; it took approximately one week for two annotators to label all the samples. The human labelling considerably improved the precision of the final dataset and did not impose a significant overhead in labour relative to the labour involved in the data engineering, data science, and analysis contained in this study.

The final dataset includes 68,661 detections in 131 countries with a mean detection area of approximately 70,000$m^2$. Country-level aggregates are shown in Supplementary Table 10. Any false positives remaining in the dataset are a product of human error. Final precision statistics in Supplementary Figure 6 and Supplementary Table **??** is reported again the test set. We find human error in hand labelling reduced final precision to approximately 98.6%.

## 2 Machine Learning Pipeline Performance

We design our machine learning pipeline to produce a dataset of installations suitable for a variety of applications spanning finance, policy, planning, and engineering. As such, we optimise the pipeline for precision, as false positives must be expensively hand-filtered. The first steps of both the S2 and SPOT6/7 branches optimise for recall. Then the subsequent filters eliminate false positives to improve precision. Performance benchmarks through the pipeline are shown in Supplementary Figure 6. The performance of our pipeline is comparable to other non-residential PV solar energy remote sensing computer vision studies. Yu et al. (2018)[6] achieve a precision of 93.7% and a recall of 90.5% for 'non-residential solar'. Imamoglu et al. (2017)[23] achieve a recall of 93.3% and an IoU of 56% for PV solar energy installations with generating capacities in excess of 5MW. Hou et al. (2020)[7] report a mean




Supplementary Figure 6: **Machine learning pipeline validation performance for S2, SPOT, and combined pipeline branches**. Performance shown for installations over 10,000 $m^2$, capturing 97% of installation generating capacity. Final precision is shown for the entire dataset, not validation.

IoU of 94.21% but do not provide other performance criteria nor the size of PV solar energy systems that they detect. The performance of our pipeline is comparable to these studies, and the slight impairment in key metrics might be ascribed to the considerably larger and more diverse geography we trained and deployed our model on.

In general, smaller installations were more difficult to detect, with both false negatives and false positives skewed towards smaller installations in validation. Increasing the precision of detections of smaller installations in the SPOT6/7 branch came with considerable tradeoff in SPOT6/7 recall. The combination of the SPOT6/7 and S2 branches took advantage of their comparative advantages: SPOT6/7 boosted the pipeline's precision for small installations, while the S2 branch ensured near-complete coverage of large installations.

Covariate shift between our training and validation sets, and our imagery corpus for deployment introduces uncertainty into our pipeline performance. Our training, validation, and test data is biased by geography (see Supplementary Table 7), the size of detections (see Supplementary Figure 2), and the installation date of the samples. The imagery corpuses themselves are biased - both in natural phenomena, e.g. amospheric conditions and land cover, and also measurement. While Sentinel-2 imagery is consistent, the quality

and timeliness of SPOT6/7 imagery varies across the planet, with imagery in the US being more recent and of higher quality. We observe the geographic shift in our machine learning pipeline precision by comparing the distribution of our dataset before and after hand-verification (respectively, S2-V4, SPOT-V2 and FINAL in Supplementary Table **??**), see Supplementary Figure 7. Supplementary Table 7 shows data distribution and pipeline performance by major country. Across the main countries, recall was worse for countries not represented in the validation set, suggesting the pipeline overall has been overfit to the validation data. Pre-handverification pipeline precision, however, is not as biased to the distribution of training or validation data. Outliers such as high-precision, low-representation Belgium, Greece, Turkey, and South Korea, and low-precision, high-representation United States, suggest the pipeline has been overfit to a specific type of PV solar energy facility most represented in the training data. The pipeline generalises well across human and physical geographical factors but not the distributions of size, type and geometry of PV solar energy facilities which are covarying with geography.

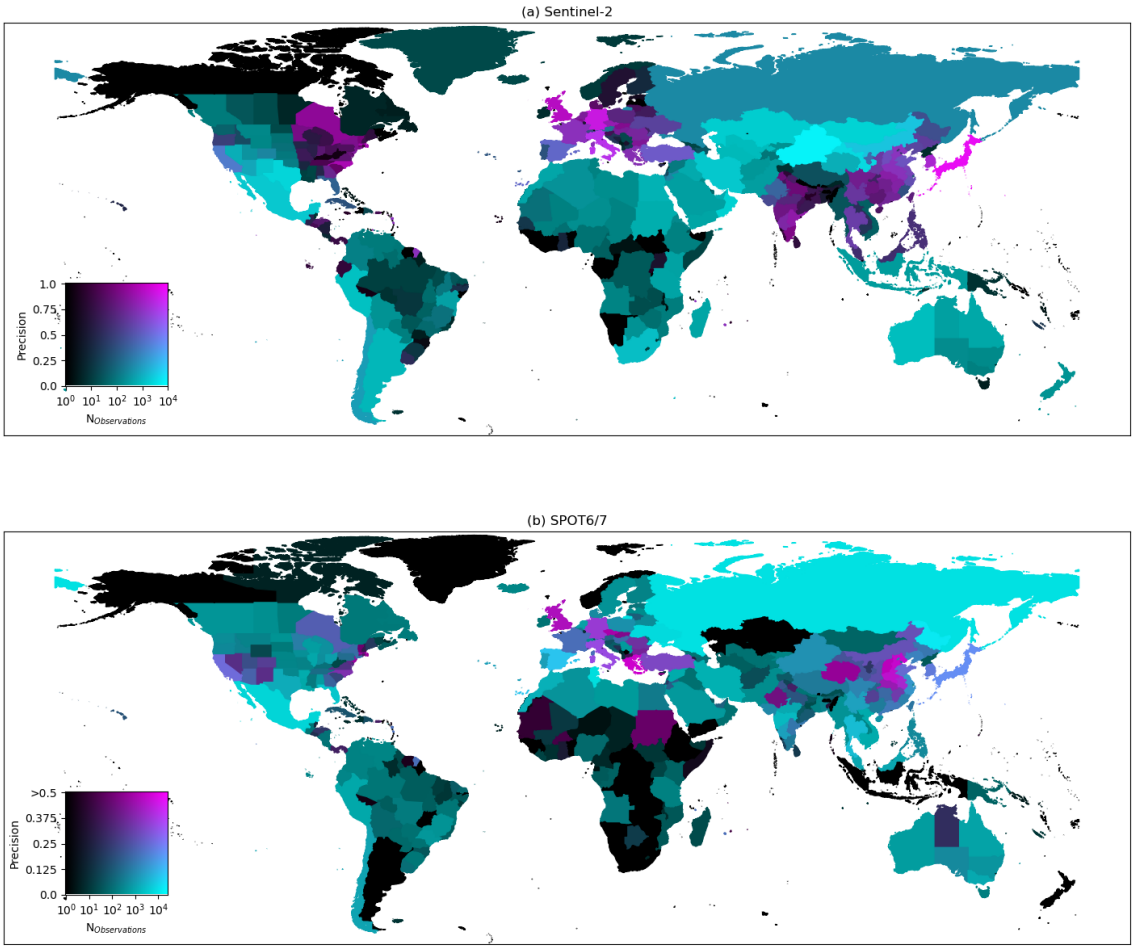
## 3 Machine Learning Model Interrogation

We interrogate our semantic segementation models to study the relative importance of each spectral band and susceptibility to noise, see Supplementary Figure 8. We test for impairment of the model‘s IoU performance against the validation set under the conditions of one-vs-all band dropout, and additive and multiplicative noise of 10%, 20% and 30%.

For the S2 ResUNet model, we find that additive noise has a similar affect on model performance when added to any band other than cirrus and the alpha channel. We hypothesise that the model‘s variance sensitivity to the cirrus band has been nullified by the heuristic atmospheric correction algorithm used to preprocess the S2 imagery. Additive

Supplementary Table 4: Pipeline Test Performance

| Area Range $[m^2]$ | Precision [%] | | | | | | Recall [%] | | | | | IoU [%] | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | S2-V1 | S2-V4 | SPOT-V1 | SPOT-V2 | PRE-HANDV. | FINAL | S2-V1 | S2-V4 | SPOT-V1 | SPOT-V2 | FINAL | S2-V1 | S2-V4 | SPOT-V1 | SPOT-V2 | FINAL |
| $10^2$-$10^3$ | 1 | 23 | 3 | 19 | 28 | **96.6** | 5 | 2 | 3 | 3 | **4** | 0 | 28 | 0 | 5 | **45** |
| $10^3$-$10^4$ | 2 | 71 | 4 | 31 | 42 | **96.0** | 41 | 16 | 35 | 34 | **43** | 6 | 49 | 2 | 25 | **65** |
| $10^4$-$10^5$ | 1 | 88 | 8 | 49 | 69 | **98.9** | 90 | 52 | 64 | 47 | **88** | 13 | 64 | 7 | 43 | **78** |
| $10^5$-$10^6$ | 18 | 85 | 20 | 58 | 85 | **98.1** | 97 | 75 | 66 | 29 | **97** | 13 | 65 | 15 | 58 | **88** |
| $>10^6$ | 29 | 71 | 20 | 79 | 93 | **97.2** | 91 | 67 | 70 | 39 | **94** | 11 | 30 | 10 | 76 | **88** |



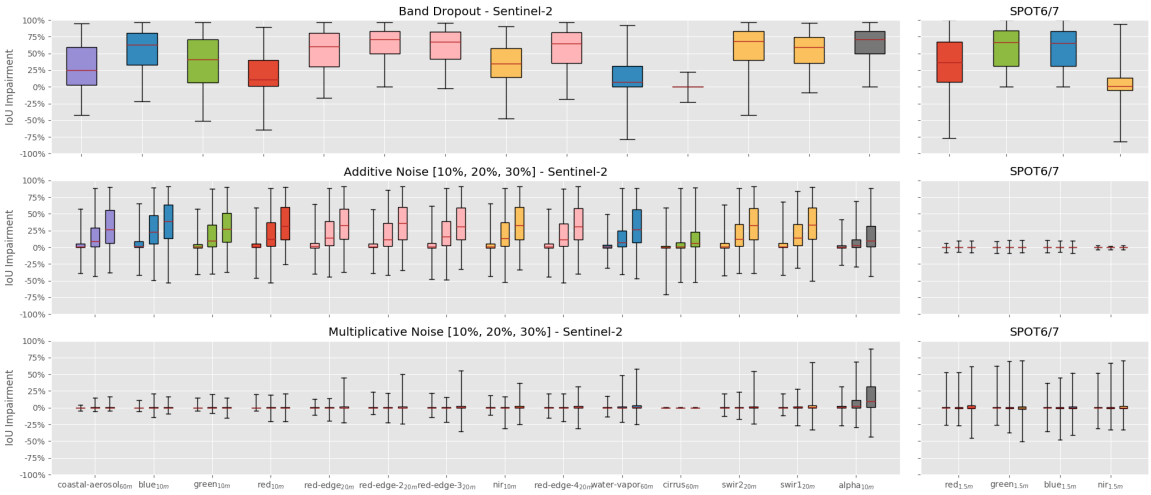

Supplementary Figure 7: **Pipeline precision for S2 (a) and SPOT (b) branches prior to hand verification.** Magenta areas indicate high model precision and many detections; cyan areas indicate low model precision and many detections; dark areas indicate few detections. S2-branch precision was high in geographies with a large amount of PV solar energy not represented in the training set (e.g. CN, IN, JP, IT), but produced excessive false positives in desert and mountainous areas (e.g. MX, ZA, IR, KZ, Texas, Western Australia) and reduced precision in areas where PV solar energy was otherwise abundant (e.g. ES, CL, TR, Xinjiang, California). SPOT precision was driven by confusion with objects and shadows. As examples, confusion with tree farms, crop patterns, chicken coups, and mountain shadows drove underperformance in Tunisia, Mexico, and Russia.

noise has a larger affect on model performance than multiplicative noise, both of which are consistent across bands. Additive noise is likely larger in absolute magnitude than multiplicative noise due to the skew of the data to the lower end of the sensor‘s operating range.

With band dropout, we find IoU performance of the S2 ResUNet is not discernibly biased to either the spectral or spatial features of the input bands. The coastal-aerosol ultraviolet band at 60m resolution shows a similar importance to model performance as the 10m near-infrared band. Blue and green 10m bands are highly important for model performance, as are narrow bandwidth red-edge bands and broad bandwidth short-wave infrared bands.

For the SPOT6/7 model, we find it is more sensitive to lower wavelength green and blue bands than red and near-infrared bands. This matches our intuition based on the blue-green appearance of PV panels in the visual spectrum. The SPOT6/7 model was not as affected by additive noise, likely due to the noise augmentations used in training.

We qualitatively examine false positive detections in both S2 and SPOT branches. In the SPOT branch, false positives were often generated by objects: greenhouses, row crops, and regularly-spaced trees, as well as mountain shadows and water, suggesting the dominance of spatial and geometric features, and color. In the S2 branch, false positives corresponded to both spatial features (chicken coups, greenhouses) and spectral features (deserts and high aerosol areas, and green sports fields and tennis courts). We conclude that the both computer vision models makes use of the whole spectrum of available bands, and robustly learn feature representations across a sophisticated spectral and spatial topography.




Supplementary Figure 8: **Impairment of IoU for S2 and SPOT semantic segmentation models under band perturbation, additive noise, and multiplicative noise.** S2 model shows sensitivity to all bands except the cirrus band, indicating the importance of both spectral and spatial features. The SPOT model shows increased sensitivity to the lower-wavelength green and blue bands, indicating an increased reliance on color.

Supplementary Table 5: Detection Confidence Levels

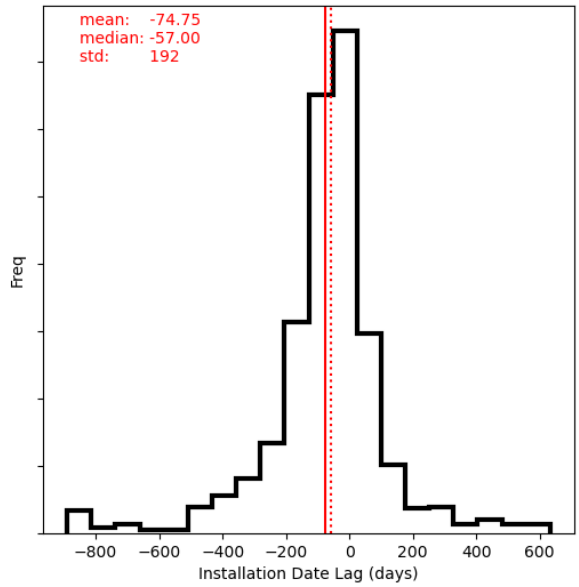
| Level | Geometry | Install Date | Description | Fraction of Polygons | Fraction of Capacity |
|---|---|---|---|---|---|
| A | SPOT | S2 | Intersecting SPOT and S2 detections where IoU>30% | 50% | 36% |
| B | SPOT | None | SPOT-only detections | 15% | 42% |
| C | S2 | S2 | S2-only detections | 17% | 14% |
| D | Both | S2 | Intersecting SPOT and S2 detections where IoU<30% | 18% | 8% |

## 4 Feature Enrichment and Analysis

We enhance our dataset with a number of metadata properties for each detection, see descriptions in Supplementary Table 8.

**Installation Date** We obtain installation dates as described in the machine learning pipeline. For positive detections determined by RNN-2, we obtain the date of the first scene in the S2 back-catalog where the pixel-wise primary inference mean is greater than 20%. To evaluate the accuracy of this heuristical rule, we compare our data to the installation dates of matched PV solar energy installations in the United States[50]. Accurate installation dates for other geographies were not available. The distribution of our estimated installation dates against the EIA sample is shown in Supplementary Figure 9. Our observed installation dates tend to lead the US EIA 'operational dates', by approximately two months. This matches our expectations of the gap between the beginning of construction and facility operation. Our uncertainty in these dates mean that our study may have captured facilities which, at the end of 2018, might not have yet been entered into official statistics. Further, while insight might be drawn at the aggregate level, this uncertainty is too high for reliable conclusions at the level of individual PV solar energy facilities. We show, however, that Earth observation might provide a valid method for more detailed studies of PV installation dates and construction progress, and might be able to provide more specific insight into the time gap between investment decisions and facility operation.

**Generating Capacity** Our study includes all installations over approximately $10kW$ in size, a cross-sectional definition encompassing the IEA definitions of 'utility-scale', 'distributed commercial/industrial', and 'off-grid' and over 89% of cumulative installed capacity in 2018 [10]. This definition captures both conventional generating stations and 'behind-




Supplementary Figure 9: **Distribution of difference in days between our estimated installation dates and US EIA[50] installation dates for matched plants in the United States.** Our observed installation date, corresponding to 20% construction completion, leads facility 'operational date' by approximately two months.

the-meter' installations that electricity system operators might otherwise have poor visibility of. We focus on these non-residential PV solar energy installations because they comprise the majority of installed PV generating capacity, suiting our target policy and engineering use cases. Our dataset does not include residential solar which is too small to detect reliably in our imagery sources and makes up 11% of total installed capacity[10]. We acknowledge that 10kW does not perfectly delineate residential and non-residential PV solar energy installations (some residential installations might be larger, some commercial or industrial installation smaller), but this definition provides consistency with the definitions of the IEA[10] for comparative purposes.

We develop a simple estimate of generating capacity for illustrative and comparative purposes, drawing on the novelty of our spatially-localised dataset.We address the non-trivial challenge of converting geospatially-localised polygons to AC generating capacities by obtaining global maps of incident irradiance at optimal tilt angle, $GTI$, and PV solar energy production intensity, $PVOUT$, from the *Global Solar Atlas*[51]. We obtain distribu-

tions of the installation ground coverage ratio (also called the packing-factor), $GCR$, and the module efficiency, $\eta$, from Ong, S. et al. (2013)[43], who obtain data for 193 sample projects in the United States. To obtain a distribution for inverter loading ratio, $ILR$, we cross-reference their survey with data obtained from the US EIA.[50] We use gamma distributions for $GCR$, $ILR$, and $\eta$ to capture the skew of the data. Values for the shape parameters $k$, the location parameters $loc$, and the scale parameter $\theta$ are shown in Supplementary Table 6. We estimate nominal peak AC generating capacity, $gencap$, for each PV solar energy facility $i$, using the array area, $area$. Using our test set, we are also able to determine the error between the true 'direct area' of the facility, and the facility's predicted area. To capture the bias of our pipeline against different sizes of facilities, we use area-binned calculations of this error, $\varepsilon$. We assume these errors are normally distributed and estimate uncertainty by sampling from their distributions shown in Supplementary Table 6.

In developing a hand-labelled test set, we observe that the ground coverage ratio for PV solar energy facilities varies significantly by geography. Using only data from Ong et al's[43] study in the the United States introduces significant epistemic uncertainty into our estimate of generating capacity. To quantify this uncertainty, we estimate that facility ground coverage ratios might vary as much as $\pm 20\%$ in different countries, and add this error, $\gamma$, to each country's ground coverage ratio distribution. We sample this interval uniformly on a per-country basis when evaluating uncertainty in our estimate. Equation 1 shows how PV solar energy facility nameplate generating capacities are estimated using the predicted area, incident irradiance, solar energy production intensity, and distributions for area error, ground coverage ratio, panel efficiency, and inverter loading ratio.

$$gencap_i\,[kW_{p,AC}] = area\left[m^2\right] \times (1+\varepsilon_{areabin,i}) \times GTI_i \left[\frac{kWh_{irr}}{m^2}\right] \times (GCR+\gamma) \left[\frac{m^2}{m^2}\right] \times \eta \left[\frac{kWh_{prod}}{kWh_{irr}}\right] \div PVOUT \left[\frac{kWh_{prod}}{kW_{DC,p}}\right] \div ILR \left[\frac{kW_{p,DC}}{kW_{p,AC}}\right] \quad (1)$$

Equation 2 shows how estimates of individual generating capacities can be aggregated to an estimate of global gross generating capacity, $gross\_gencap$, for comparison with figures from IRENA and the IEA. Facility generating capacities are summed and then multiplied by pipeline precision, $Pr$, and divided by pipeline recall $Re$. Because the area bias of the pipeline is measurable, the dataset, precision, and recall is binned by facility area. The Boolean nature of pipeline precision and recall means that it can be modelled as a Bernoulli proccess and that the distribution precision and recall is binomial. The area-binned binomial distribution parameters for pipeline precision and recall is shown in Supplementary Table 6.

$$gross\_gencap\,[kW_{p,AC}] \;=\; \sum_{areabin} \left[\sum_i gencap_{areabin,i}\right] \times\; Pr_{areabin} \;\div\; Re_{areabin} \quad (2)$$

With the area-binned precision and recall values shown in Supplementary Table **??** we are able to obtain a best-estimate of 423GW global gross installed generating capacity. Uncertainty due to the distributions of area error $\varepsilon$, panel efficiency $\eta$, inverter loading ratio $ILR$, and ground coverage ratio $GCR$ add an uncertainty of [-11GW, +12GW] to this figure with 95% confidence. Uncertainty due to the geographic bias of the ground coverage

ratio, $\gamma$, increases the uncertainty to [-75GW, +77GW]. Other sources of epistemic uncertainty require further research to be quantified (see below) and so we advise data users to carefully consider our uncertainty in any downstream analysis.

There is considerable scope to further quantify and reduce uncertainties in future work. Not included in our calculations are epistemic uncertainties around the type of PV installation, which will effect its ground coverage ratio and solar energy production intensity. The type of PV solar energy installation (e.g. fixed, single-axis, and dual-axis tilt systems) can be determined in remote sensing imagery. Estimates of the installation orientation, and even tilt angle can be obtained from high resolution imagery and used to recalculate incident irradiance. Panel counts and array areas can be measured in high resolution imagery to give a more certain distribution of ground coverage ratio. The type of cells (single- and poly-crystalline, thin film) might also be ascertained from high resolution imagery. Installation date, cell type, and other meta-data might be combined to obtain an estimate of panel efficiency. All these distributions will also vary by geography. Our dataset provides a starting point for the development of a more detailed, accurate, and certain global dataset of the world's PV solar energy facilities.

Our predicted data could be used in future studies as noisy labels, analogous to how OSM-derived training data were used in this study. The advantage our data provides for future studies is its global coverage - training data would no longer be limited to certain geographies represented in OSM. An improved machine learning pipeline could be designed which pretrains a model using our predicted data, and then fine-tunes using handlabelled data to minimise error between a facility's predicted footprint and it's true 'direct area' footprint. Our validation and global test sets can also be used to assess pipeline performance against time-benchmarked data across diverse geographies.

Supplementary Table 6: Installed capacity calculation uncertainty [43,50]

| Metric | Area Bin [$m^2$] | Symbol | $N_{samples}$ | Distribution |
|---|---|---|---|---|
| Ground Coverage Ratio | | $GCR$ | 39 | $GAMMA(k{=}19.5, loc{=}0.030, \theta{=}0.025$ |
| Inverter Loading Ratio | | $ILR$ | 61 | $GAMMA(k{=}1.8, loc{=}0.087, \theta{=}0.032)$ |
| Module Efficiency | | $\eta$ | 109 | $GAMMA(k{=}26.3, loc{=}0.68, \theta{=}0.018$ |
| $GCR$ country bias | | $\gamma$ | - | $UNIFORM(-0.2, 0.2)$ |
| Area Error | $<10^3$ | $\epsilon$ | 66 | $NORM(\mu{=}0.377, \sigma{=}0.637)$ |
| | $10^3{-}10^4$ | | 1091 | $NORM(\mu{=}0.199, \sigma{=}0.722)$ |
| | $10^4{-}10^5$ | | 1819 | $NORM(\mu{=}0.109, \sigma{=}0.513)$ |
| | $10^5{-}10^6$ | | 553 | $NORM(\mu{=}-0.013, \sigma{=}0.281)$ |
| | $>10^6$ | | 62 | $NORM(\mu{=}-0.025, \sigma{=}0.202)$ |
| Precision | $<10^3$ | $Pr$ | 90 | $BINOM(n{=}90, p{=}0.967)$ |
| | $10^3{-}10^4$ | | 973 | $BINOM(n{=}973, p{=}0.960)$ |
| | $10^4{-}10^5$ | | 1,319 | $BINOM(n{=}1319, p{=}0.989)$ |
| | $10^5{-}10^6$ | | 531 | $BINOM(n{=}531, p{=}0.981)$ |
| | $>10^6$ | | 72 | $BINOM(n{=}72, p{=}0.972)$ |
| Recall | $<10^3$ | $Re$ | 1,711 | $BINOM(n{=}1711, p{=}0.039)$ |
| | $10^3{-}10^4$ | | 2,556 | $BINOM(n{=}2556, p{=}0.427)$ |
| | $10^4{-}10^5$ | | 2,060 | $BINOM(n{=}2060, p{=}0.883)$ |
| | $10^5{-}10^6$ | | 573 | $BINOM(n{=}573, p{=}0.965)$ |
| | $>10^6$ | | 66 | $BINOM(n{=}66, p{=}0.939)$ |

Supplementary Table 7: Data Distribution and Pipeline Performance by Country

| Country | Dataset Representation | | | | Testset | Pipeline Performance [%] | |
|---|---|---|---|---|---|---|---|
| | Training | Validation | Test | Predicted | RoIs | Precision[A] | Recall[B] |
| Belgium | 0 | 0 | 264 | 156 | 1 | 100 | 64 |
| Chile | 0 | 0 | 118 | 150 | 3 | 30 | 92 |
| China | 1,444 | 0 | 804 | 18,450 | 10 | 88 | 87 |
| Germany | 24,224 | 124 | 952 | 4,702 | 16 | 90 | 91 |
| United Kingdom | 889 | 492 | 143 | 2,221 | 4 | 100 | 95 |
| Greece | 0 | 0 | 145 | 3,643 | 3 | 86 | 92 |
| India | 292 | 382 | 402 | 2,277 | 6 | 83 | 95 |
| Italy | 0 | 155 | 132 | 5,796 | 2 | 69 | 98 |
| Japan | 981 | 3,527 | 1,009 | 10,504 | 5 | 92 | 92 |
| South Korea | 0 | 0 | 436 | 2,679 | 2 | 92 | 71 |
| Netherlands | 0 | 0 | 149 | 130 | 2 | 92 | 83 |
| Thailand | 6 | 6 | 163 | 529 | 3 | 83 | 94 |
| Turkey | 0 | 0 | 109 | 1,543 | 2 | 99 | 94 |
| United States | 8,117 | 1,228 | 1,566 | 7,713 | 22 | 62 | 90 |

A: Pipeline precision prior to hand-verification.
B: Pipeline recall for facilities larger than 10,000m$^2$.

**Land Cover** Land use change for renewable energy production has climate and conventional environmental impacts and trade-offs with food production systems, community and indigenous land use, and other aspects of the human economy. We identify the pre-existing land cover type for PV solar energy installations using publically available historical land cover datasets. We sample PV solar energy sites again European Space Agency Climate Change Initiative 300m land cover map for the years 1992 through 2018.[39] We choose this land cover product for its global availability, minimising geographical bias in land cover classification. (We recognise that for certain geographies there are higher resolution land cover products available, however we choose not to introduce the bias of sampling from multiple land cover products.) For sites where the installation date has been determined, the pre-isntallation land cover is obtained from the 2012 global land cover map. Where the installation date is unknown, or is known to be prior to 2016, we choose 2007 as a representative pre-installation date. This date is sufficiently long ago to ensure the sample pre-dates the installation of the PV facility, while mitigating land cover changes which may have happened between 1992 – the first year for which data is available – and the eventual (unknown) installation date of the facility. This introduces a small uncertainty for the 8.5GW of installations built before this time period.[10]

Supplementary Figure 10 shows the diversity of land cover, installed capacity growth, and installation size distributions in the top 20 countries in our dataset. Over the study time period, installed capacity in the top four countries (China, the United States, India, and Japan) grows considerably, while EEA countries (Italy, the United Kingdom, Germany, Spain) do not grow as fast, with the exception of France. The size histograms of installations show that larger installations are installed preferentially on barren, grassland, and desert landcovers, while smaller installations are situated on forests or built-up areas.

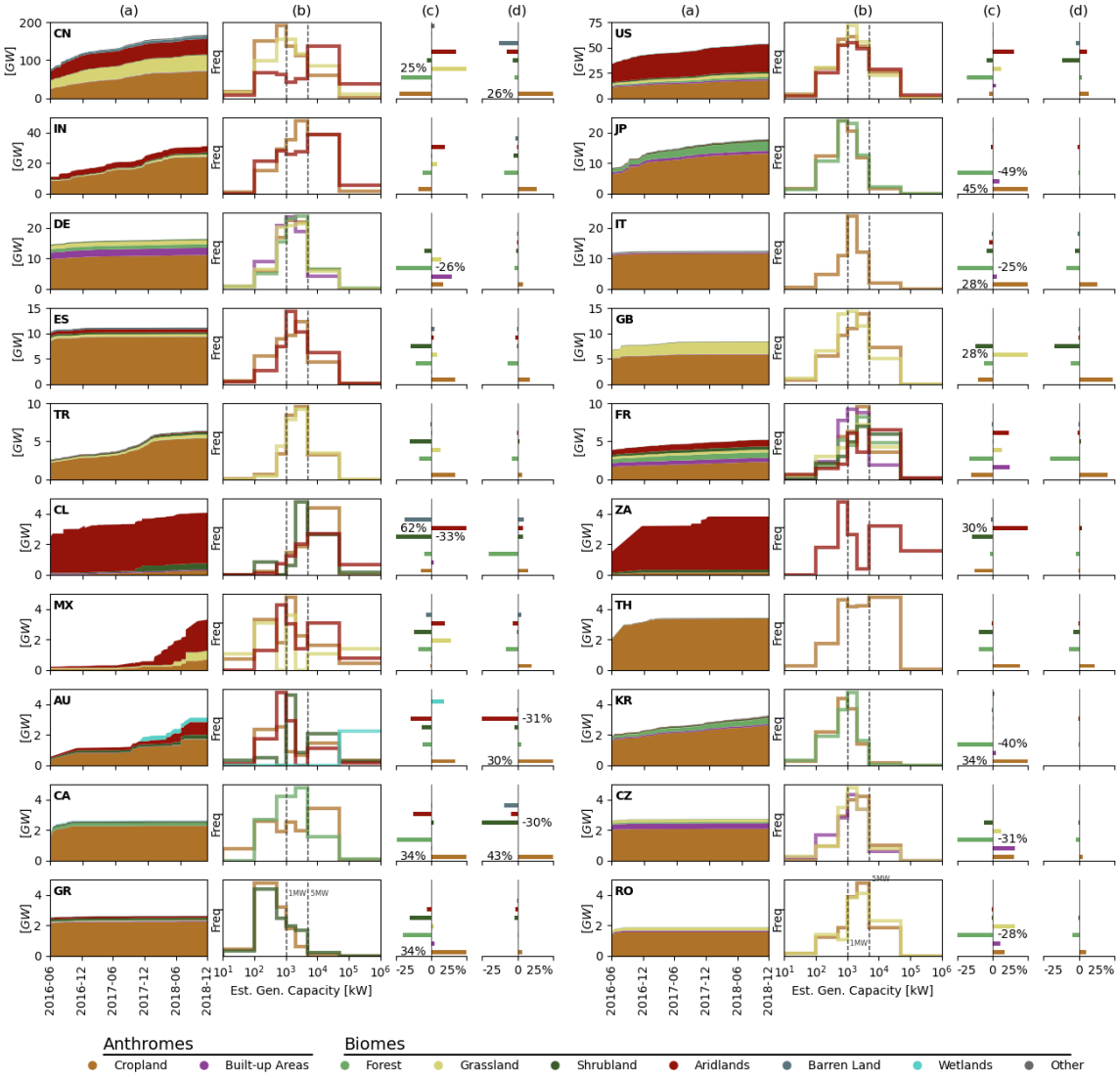
The land cover class distribution of PV solar energy sites varies significantly by country, indicating land cover is not the single driving factor for PV siting decisions. PV installations in most countries are located on cropland, but China, the United States, and France showing a larger diversity of pre-existing land covers at PV sites. Local land cover bias varies by country, suggesting that land cover considerations are secondary to other factors. We hypothesise that the bias to agricultural sites is driven by unobserved confounding whereby both PV solar energy and agriculture favour sites in close proximity to human populations and access to markets, labour, infrastructure, etc. Most countries show a large bias against forest-covered sites and a moderate bias against shurbland sites. We hypothesise that these land covers impose expensive land clearing operations prior to PV solar energy installation. Only the Czech Republic, France, and Germany show a concerted bias for the development of PV solar energy on built-up land. PV mega-projects in excess of 250MW have been built in China, the United States, Mexico, India, and South Africa. Many of these mega-projects are also built on aridlands, shown in the strong positive bias for aridlands in large facilities in excess of 5MW, see Supplementary Figure 11.

We detect a number of 'other' installations which includes installations on water. Due to the coarse 300m resolution of the land cover data, we hand-verify these to ensure that they are not mis-classified coastal installations, for example. In doing so we identify 84 floatovoltaic sites in our data.

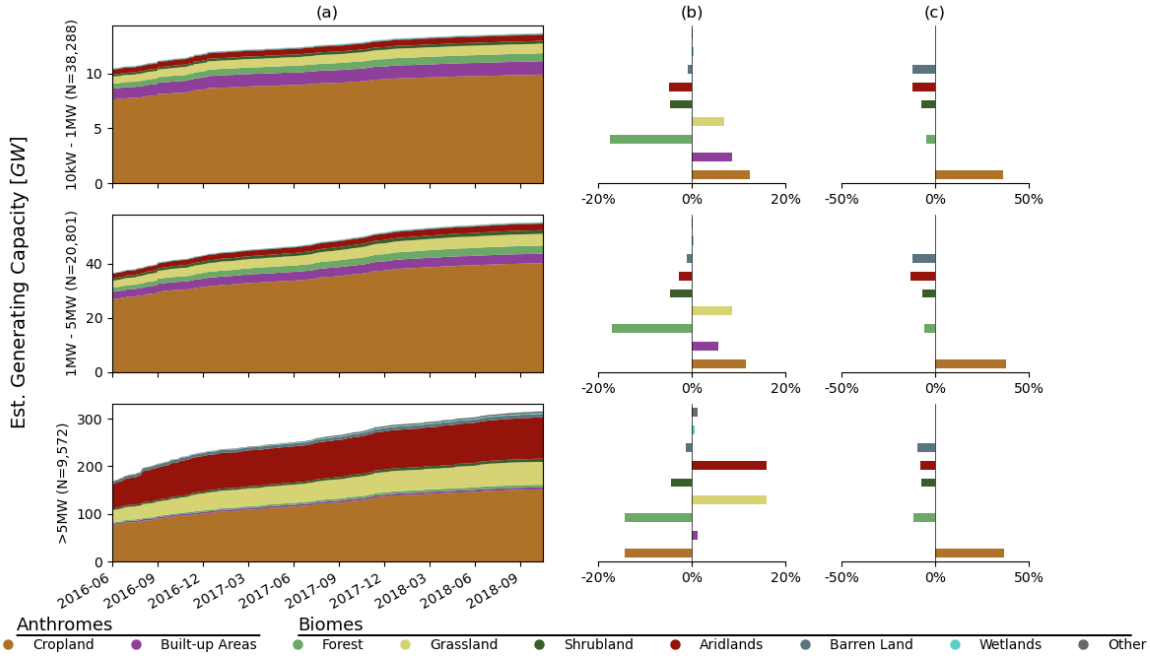
Supplementary Figure 11 shows the growth in generating capacity and land cover bias binned by generating capacity. Despite being fewer in number, large utility-scale projects make up the largest portion of gross generating capacity. Unlike installations smaller than 5MW, these large installations have a bias away from croplands and towards aridlands. Installations smaller than 5MW have a positive bias towards croplands and a stronger positive

bias towards built-up areas. These two trends suggest the fundamental drivers for PV solar energy siting decisions are dependent on the size of the installation and there is a meaningful difference between large and small PV energy installations. All size classifications show a bias away from forest and shrubland.

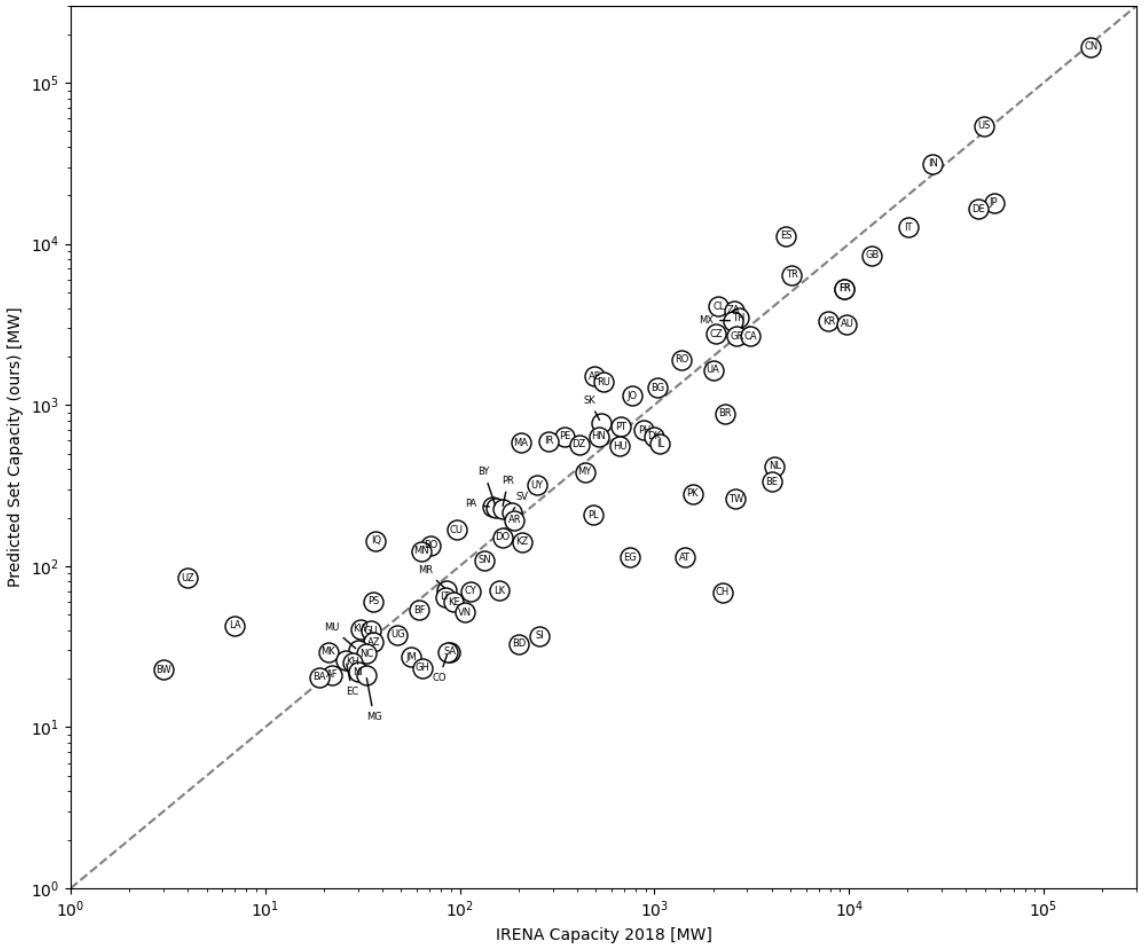
**Protected Areas and Indigenous and Community Lands** Recognising the impact that non-residential PV solar energy development can have on protected areas and community and indigenous lands (see, e.g. Hernandez, R.R. et al. (2015)[52]), we also add proximity to protected areas and community and indigenous lands as additional features of our data. For each detection, protected areas and indigenous and community lands within 10km of the site are identified. We obtain protected areas from the World Database of Protected Areas[53], and retain the distance, identification number, name, and type of protected area for each of our detections. For indigneous and community lands, we obtain proximate lands from Landmark.org[54] and note the record name. Neither of these datasets are perfect, so absence of these records should not indicate the absense of a protected, community, or indigenous area, but providing convenient access to this data alongside our detections might help researchers better identify and study impacts on these vulnerable lands.



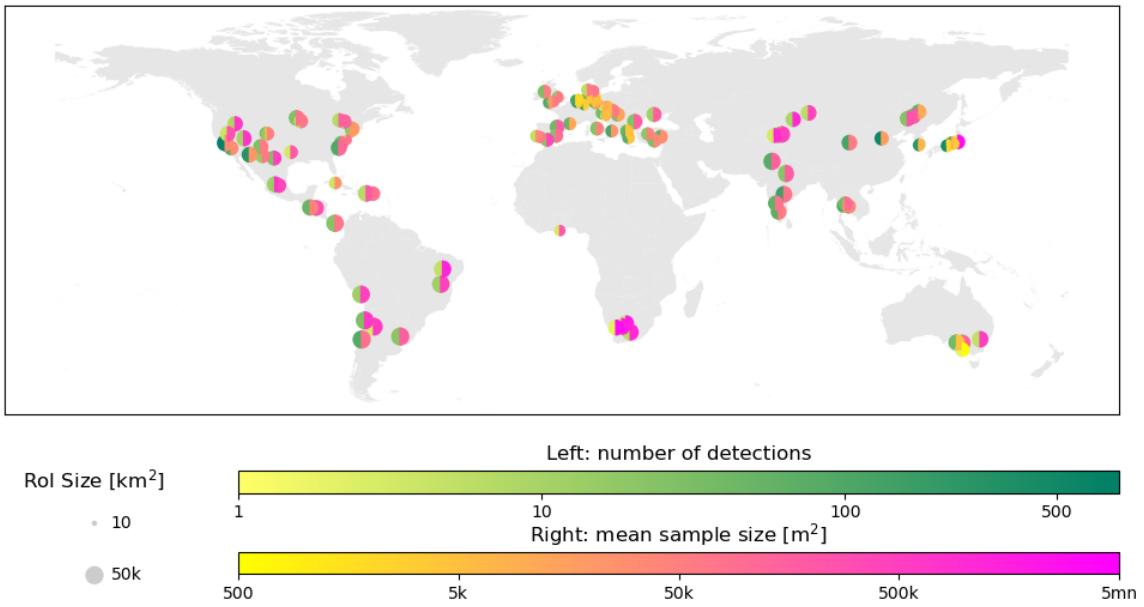

Supplementary Figure 10: **Pre-existing land cover for PV solar energy installations**. Panels show the time series of installations (a); the distribution of installation sizes by land cover (b); local bias (c) between PV land cover and the distribution of local land covers within a 0.5°x0.5° grid pixel; and country-level bias (d) between country land cover distribution and distribution of local land covers containing PV detections.




Supplementary Figure 11: **Pre-existing land cover for PV solar energy installations for three bins of nominal generating capacity: 10kW to 1MW; 1MW to 5MW; and >5MW.** Panels show the time series (a) of installations; local skew (b) between PV land cover and local land covers, where localities are described by the pixels of a 0.5°x0.5° grid containing PV detections; and global bias (c) between the global land cover distribution and locality land cover distributions for localities containing that size of PV solar energy installations.




Supplementary Figure 12: **Comparison of our data aggregated by country to country-level data from IRENA[3].** Note that IRENA data is inclusive of residential PV solar energy.




Supplementary Figure 13: **Test set regions-of-interest (RoIs).** 122 RoIs are globally distributed and cover an aggregate area of 580,000km$^2$.

Supplementary Table 8: Dataset Metadata

| Property | Description |
|---|---|
| unique_id | A unique identifier for each detection. |
| geometry | The detection geometry in WGS84 longitude/latitude coordinates. |
| confidence | Installation confidence classification, see Supplementary Table 5. |
| iso-3166-1 | Country-level political administrative ISO-A2 code. |
| iso-3166-2 | Province/state-level political administrative code. |
| area | Installation geodesic area in $m^2$. |
| capacity_mw | Estimated nominal peak AC generating capacity in MW. |
| install_date | Comma-separated estimates for installations and major expansions in 'YYYY-mm-dd' format. |
| gti | Incident solar irradiance for optimally tilted angle, in $\frac{kWh_{irr}}{m^2}$. |
| pvout | Photovoltaic power potential, in $\frac{kWh_{prod}}{kW_{p,DC}}$ . |
| wdpa_10km | Protected area records from the World Database of Protected Areas[53] within 10km, arranged as $\{$distance, WDPA_ID, name, type$\}$ |
| ind_comm_10km | Indigenous and community land records from Landmark.org[54] within 10km, arranged as $$name\_1, name\_2, etc.$$ |
| lc_20XX | Land coverage assessment for 1992 through 2018 for ESA Climate Change Initiative 300m land cover, see Supplementary Table 9.* |

Supplementary Table 9: Landcover Class Mappings - ESA Climate Change Initiative Land Cover[39]

| Class | Mapping Classes |
|---|---|
| Cropland | 10: Cropland, rainfed; 11: Cropland, rainfed, herbaceous cover; 12: Cropland, rainfed, tree or shrub cover; 20: Cropland, irrigated or post-flooding; 30: Mosaic cropland $50\%$ / natural vegetation $50\%$; 40: Mosaic natural vegatation $50\%$ / cropland $50\%$ |
| Built-up areas | 180: Urban areas; |
| Forest | 50: Tree cover, broadleaved, evergreed; 60: Tree cover, broadleaved, deciduous; 61: Tree cover, broadleaved, deciduous, closed; 62: Tree cover, broadleaved, deciduous, open; 70: Tree cover, needleleaved, evergreen; 71: Tree cover, needleleaved, evergreen, closed; 72: Tree cover, needleleaved, evergreen, open; 80: Tree cover, needleleaved, deciduous; 81: Tree cover, needleleaved, deciduous, closed; 82: Tree cover, needleleaved, deciduous, open; 90: Tree cover, mixed leaf type; |
| Grasslands | 130: Grassland; |
| Shrub, herbaceous, and sparse vegetation | 100: Mosaic tree and shrub $50\%$ / herbaceous cover $50\%$; 110: Mosaic herbaceous cover $50\%$ / tree and shrub $50\%$; 120: Shrubland; 121: Evergreen shrubland; 122: Deciduous shrubland; 140: Lichens and mosses; 150: Sparse vegetation; 151: Sparse trees; 152: Sparse shrub; 153: sparse herbaceous cover; |
| Barren areas | 200: Bare areas; 201: Consolidated bare areas; 202: Unconsolidated bare areas; |
| Aridlands | Forest, Grasslands, Shrub/herbaceous/sparse, or barren areas where the aridity index is less than 20% |
| Wetlands | 160: Tree cover, flooded, fresh or brackish water; 170: Tree cover, flooded, saline water; 180: Shrub or herbaceous cover, flooded; |
| Other | 0: No data; 210: Water bodies; 220: Permanent snow and ice; |

Supplementary Table 10: Aggregate Statistics by Country

| ISO-3166-1 | Number of Detections | Est. Total Gen. Capacity [MW] | Mean Detection Size [MW] | ISO-3166-1 | Number of Detections | Est. Total Gen. Capacity [MW] | Mean Detection Size [MW] |
|---|---|---|---|---|---|---|---|
| CN | 167,378 | 18,449 | 9.07 | PR | 226 | 36 | 6.27 |
| US | 54,144 | 7,638 | 7.09 | SV | 214 | 24 | 8.93 |
| IN | 31,444 | 2,277 | 13.81 | PL | 208 | 127 | 1.64 |
| JP | 18,005 | 10,504 | 1.71 | NA | 206 | 18 | 11.42 |
| DE | 16,483 | 4,702 | 3.51 | AR | 194 | 8 | 24.19 |
| IT | 12,649 | 5,796 | 2.18 | CU | 168 | 40 | 4.19 |
| ES | 11,238 | 1,970 | 5.70 | DO | 151 | 13 | 11.62 |
| GB | 8,470 | 2,222 | 3.81 | IQ | 144 | 6 | 23.92 |
| TR | 6,436 | 1,543 | 4.17 | KZ | 141 | 6 | 23.48 |
| FR | 5,274 | 936 | 5.63 | BO | 134 | 4 | 33.55 |
| CL | 4,097 | 150 | 27.31 | MN | 124 | 6 | 20.71 |
| ZA | 3,850 | 110 | 35.00 | EG | 113 | 24 | 4.72 |
| TH | 3,458 | 529 | 6.54 | AT | 113 | 70 | 1.61 |
| MX | 3,337 | 110 | 30.34 | SN | 108 | 9 | 12.05 |
| KR | 3,316 | 2,680 | 1.24 | UZ | 85 | 6 | 14.11 |
| AU | 3,160 | 164 | 19.27 | MR | 71 | 11 | 6.43 |
| CZ | 2,768 | 951 | 2.91 | LK | 70 | 10 | 7.01 |
| GR | 2,671 | 3,645 | 0.73 | CY | 70 | 78 | 0.90 |
| CA | 2,668 | 353 | 7.56 | CH | 68 | 48 | 1.43 |
| RO | 1,904 | 398 | 4.78 | LT | 65 | 31 | 2.08 |
| UA | 1,642 | 166 | 9.89 | PS | 60 | 18 | 3.34 |
| AE | 1,500 | 15 | 100.02 | KE | 60 | 1 | 59.77 |
| RU | 1,385 | 51 | 27.16 | BF | 53 | 4 | 13.28 |
| BG | 1,277 | 340 | 3.75 | GT | 52 | 7 | 7.40 |
| JO | 1,143 | 84 | 13.61 | VN | 51 | 5 | 10.28 |
| BR | 882 | 63 | 14.00 | LA | 43 | 4 | 10.71 |
| SK | 774 | 279 | 2.78 | KW | 41 | 5 | 8.11 |
| PT | 735 | 145 | 5.07 | GU | 40 | 2 | 19.81 |
| PH | 698 | 82 | 8.51 | UG | 37 | 3 | 12.44 |
| HN | 638 | 64 | 9.96 | SI | 37 | 22 | 1.67 |
| DK | 637 | 119 | 5.36 | AZ | 34 | 7 | 4.81 |
| PE | 635 | 14 | 45.37 | BD | 33 | 14 | 2.34 |
| IR | 595 | 47 | 12.66 | MU | 30 | 11 | 2.72 |
| MA | 581 | 44 | 13.21 | SA | 29 | 7 | 4.20 |
| IL | 574 | 66 | 8.69 | CO | 29 | 6 | 4.89 |
| DZ | 570 | 30 | 19.01 | MK | 29 | 15 | 1.95 |
| HU | 554 | 301 | 1.84 | NC | 29 | 4 | 7.13 |
| NL | 416 | 130 | 3.20 | JM | 27 | 2 | 13.60 |
| MY | 383 | 59 | 6.49 | EC | 26 | 8 | 3.24 |
| BE | 333 | 156 | 2.14 | KH | 25 | 5 | 5.07 |
| UY | 317 | 20 | 15.83 | GH | 23 | 3 | 7.80 |
| PK | 281 | 72 | 3.91 | BW | 23 | 4 | 5.69 |
| TW | 261 | 161 | 1.62 | NI | 22 | 5 | 4.41 |
| PA | 232 | 20 | 11.61 | AF | 21 | 7 | 3.01 |
| BY | 229 | 38 | 6.02 | MG | 21 | 1 | 21.08 |

**Additional References**


42. Airbus Intelligence. Spot 6 — Spot 7 Imagery. Tech. Rep., Leiden, NL (2013). URL: https://www.intelligence-airbusds.com/files/pmedia/public/r49229_9_spot_67.pdf.

43. Ong, S., Campbell, C., Denholm, P., Margolis, R. & Heath, G. Land-use requirements for solar power plants in the united states. Tech. Rep., National Renewable Energy Lab.(NREL), Golden, CO (United States) (2013).

44. Hernandez, R. R., Hoffacker, M. K. & Field, C. B. Land-use efficiency of big solar. *Environmental science & technology* **48**, 1315–1323 (2014).

45. Mateo-Garcia, G. *et al.* Towards global flood mapping onboard low cost satellites with machine learning. *Scientific reports* **11**, 1–12 (2021).

46. Meraner, A., Ebel, P., Zhu, X. X. & Schmitt, M. Cloud removal in sentinel-2 imagery using a deep residual neural network and sar-optical data fusion. *ISPRS Journal of Photogrammetry and Remote Sensing* **166**, 333–346 (2020).

47. Zhang, Z., Liu, Q. & Wang, Y. Road extraction by deep residual u-net. *CoRR* **abs/1711.10684** (2017). URL: http://arxiv.org/abs/1711.10684.

48. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

49. European Commission, C. f. I. E. S. I. N. C., Joint Research Centre; Columbia University. Ghs population grid, derived from gpw4, multitemporal (1975, 1990, 2000, 2015) (2015). URL: http://data.europa.eu/89h/jrc-ghsl-ghs_pop_gpw4_globe_r2015a.

50. US Energy Information Administration. EIA-860. Tech. Rep., Washington DC, USA (2020). URL: https://www.eia.gov/maps/layer_info-m.php.

51. SolarGIS. Global solar atlas (2016). URL: https://globalsolaratlas.info/downloads/world.

52. Hernandez, R. R., Hoffacker, M. K., Murphy-Mariscal, M. L., Wu, G. C. & Allen, M. F. Solar energy development impacts on land cover change and protected areas. *Proceedings of the National Academy of Sciences* **112**, 13579–13584 (2015).

53. UNEP-WCMC, I. World database on protected areas (2016).

54. WRI, L. Landmark: Global platform of indigenous and community lands (2017). URL: http://www.landmarkmap.org/.