Solar Asset Mapper: A continuously-updated global inventory of solar energy facilities built with satellite data and machine learning

Mason Phillpott[1], Joe O'Connor[1,*], André Ferreira[2], Max Santos[1], Lucas Kruitwagen[1], and Michael Guzzardi[1]

[1]TransitionZero
[2]Loka, previously TransitionZero
[*]Corresponding author: Joe O'Connor (joe@transitionzero.org)

ABSTRACT

TransitionZero's Solar Asset Mapper (TZ-SAM) is a global, satellite-derived dataset of utility-scale solar energy facilities (facilities with an excess of 500kW nominal generating capacity) generated with a combination of machine learning and human annotation. Our Q1 2024 dataset contains the location and geometry of 63,616 assets, along with estimated nominal generating capacities. We estimate the construction date for over 80 % of these assets. The dataset contains 19,121 square kilometres of solar energy facilities across 183 countries, with a total estimated nominal generating capacity of 711 GW. We make this and future releases of this dataset publicly available for non-commercial use.

1 Background & Summary

Solar photovoltaic (PV) is the fastest growing power generation technology in history. In 2023, the world added almost 400 GW of solar generating capacity, a ten-fold increase on the 40GW capacity installed in 2013 a decade earlier[1, 2]. The International Energy Agency (IEA)'s net-zero scenario projects a substantial increase solar generation capacity, increasing from 1,200 GW in 2023 to an estimated 4,800 GW by 2030[3].

Accurate and up-to-date facility-level data are crucial for managing intermittency, planning the grid, and identifying trade-offs with biodiversity, conservation, and land protection priorities due to the land-use and land-cover changes required for ongoing solar deployment. Currently-available datasets of solar generating capacity do not fully meet these needs. Widely used aggregated statistics, for example those from the International Renewable Energy Agency (IRENA)[2], are country-level, and don't provide the facility-level resolution required for policy, conservation, and engineering applications. Further, the year-lagged latency and cadence of this data cannot keep up with the needs of planners and operators that are changing with the speed of PV deployment.

The most complete openly-available facility-level inventories are the Global Energy Monitor (GEM)'s Global Solar Power Tracker (GSPT)[4] and the facility annotation in OpenStreetMap (OSM)[5]. The GSPT is a worldwide dataset of utility-scale PV and solar thermal facilities. It covers solar facility phases with capacities of 20 MW or more - with partial coverage of phases between 1 MW and 20 MW. In GSPT's December 2023 release[4], GEM featured a total of 620 GW of operating solar capacity – a considerable difference from the global aggregate total of 1.1 TW at the time (as reported by IRENA[2]) – and about 41 % of these records were missing precise latitude and longitude coordinates. OSM is a free and open mapping platform. Data is crowd-sourced from amateur annotators, resulting in inconsistent conventions for facility footprint geometries, and sparse availability of additional features like generating capacities. Certain geographies have much denser coverage of annotations - see Figure 1 for a map showing the distribution.

Both OSM and GSPT's large-scale (> 20 MW) records[6] are curated by hand. This labour-intensive process results in compromises being made in tracked facility sizes and release schedules, and their coverage is limited by the availability of reliable sources. In addition to the considerable challenge of keeping track of new developments manually, accurately matching announced solar facility projects with their on-the-ground facilities can be difficult - leading to unreliable location data critical for various user applications. This absence of precise facility location also hinders the ability to verify whether a project was indeed completed as anticipated, given instances where announced initiatives have been unexpectedly shelved.

With TZ-SAM we continue the work of using satellite imagery to build large-scale inventories of utility solar PV facilities. Kruitwagen et al.[7] published the first global inventory of this type, using the medium-resolution Copernicus Sentinel-2 and high-resolution Airbus SPOT satellites. They also enriched their data with estimates of generating capacity and installation date, which we also provide with our data. Their work, which we now improve upon and update, was built on the contributions of others, including Malof et al.[8] and Camilo et al.[9] who first used emerging Convolutional Neural Network (CNN) approaches to locate solar facilities in aerial imagery; Imamoglu et al.[10] who applied similar methods with medium-resolution satellite data; and particularly Yu et al.[11] who mapped solar facilities in the contiguous United States. Other recent work has focused on deployments for specific geographies, including, for example, Ortiz et al.[12] who map India or Xia et al.[13] who map China.

We consider the task of inferring nominal generating capacity of a solar facility as separate to the task of locating the facility in satellite imagery. Capacity estimation models are well established, but make heavy use of parameters that must be estimated. Ong et al.[14], for example, measured the relationship between land-use intensity (i.e. the land use per unit of solar capacity) and the stated factors of: PV Module Efficiency, Array Configuration, and Tracking type[14]. They focus solely on the US up to 2012, and rely on an array of sources such as official/developer documents and third party reports. In this work, we build an extensive training and validation set for capacity estimation, and train an estimator with fixed-effects by country.

In this work we develop a machine learning and human-validation pipeline similar to Kruitwagen et al. and deploy it on a global corpus of Sentinel-2 imagery as recent as 1st March 2024. We prepare a validated, enriched dataset by grouping polygon detections, estimating installation dates, and nominal generating capacities. Methodological enhancements relative to Kruitwagen et al. include quarterly compositing for improved pipeline performance and installation date estimation; an extensive training and test set for estimating nominal generating capacity; new validation tooling for distributed, parallel hand-validation; and new deployment tooling to greatly reduce the cost of a global survey. The resulting dataset has 144,621 polygons detections, which we group into 63,616 assets with a gross estimated capacity of 711 GW. Installation dates have been inferred for over 80 % of these assets. Our dataset is made freely available for non-commercial use[1], and is being integrated into future releases of the GSPT[6]. We also intend to release periodic updates of this dataset, with this dataset being the first in a series.

Datasets such as ours are vital for meeting the dual challenges of the 21st century - ensuring sufficient energy is available to meet development and welfare needs for all peoples, while transitioning to a net-zero energy system quickly enough to constrain anthropogenic climate change. The ongoing provision of this dataset will allow us to track progress towards these goals in near-real-time. The open-access nature of our data makes it available for all of society's stakeholders including planning and policy-making, engineering, and investment applications.

2 Methods

Our dataset is developed using a machine learning algorithm and satellite data to identify solar energy facilities and estimation their generating capacities and construction dates for facilities built after 2017. This pipeline can be deployed globally, enabling an exhaustive view of the geographical distribution and power generation potential of utility-scale solar facilties worldwide. By building a new, custom dataset of nominal facility generating capacities, we can better estimate generating capacities from country to country, facilitating detailed research for how the energy transition is unfolding around the globe. Our methodology for producing this dataset can be summarised as follows:

1. Solar Facility Detection:
   (a) Construct a training set of known solar facilties and satellite imagery.
   (b) Train a deep semantic segmentation model to predict the location and shape of a solar facility from a composite Sentinel-2 image.
   (c) Deploy this model on imagery covering the land surface of the Earth and process the results into candidate solar facility polygons.
   (d) Manually prune False Positive (FP)s from the proposed detections.

[1] https://zenodo.org/records/11368204 or https://solar.transitionzero.org

2. Solar Facility Construction Dates:
   (a) Run semantic segmentation inference through the historical imagery back-catalog for each facility.
   (b) Estimate the earliest date in which each plant is detected.
3. Solar Facility Capacities:
   (a) Construct a training set of solar facility polygons with known capacities.
   (b) Build a model to estimate the capacity of a solar facility from shape and country information.
   (c) Apply this to our validated solar facility detections.

2.1 External Datasets

We describe 3 distinct models in this section: Solar Facility Detection and Segmentation model, Solar Construction Date Estimation model, and the Solar Capacity Estimation model. The development or validation of which require the following external datasets:

Satellite Data We utilised the "Sentinel-2" dataset from the European Space Agency (ESA) Copernicus Sentinel mission for satellite images. This dataset includes images from Sentinel-2A and Sentinel-2B satellites, offering a resolution of 10 metres and a 5-day revisit period at the equator. The data is accessible free of charge on the Copernicus Open Access Hub; we access it via the Google Cloud Platform public cloud storage bucket. We process quarterly sets of Sentinel-2 images into composites for both our training and inference datasets. We filter for cloud coverage and atmospheric conditions prior to compositing, allowing us to avoid the overhead of multiple rounds of model inference per location. We start by selecting Sentinel-2 image tiles based on cloud coverage and date. We then filter these images to select the least cloudy and most recent images and generate a simple median composite from up to 5 images. The time span for creating composite images can be adjusted to suit specific tasks. Larger time spans may improve image quality but might lose recent information. Since dataset recency is one of the aims of this tool, we default to quarterly (i.e. 3-month) composites unless otherwise specified.

Solar Polygon Data We used OSM, a free and open crowd-sourced mapping tool, as our primary training set for solar plant geometries. This platform allows users to map solar facilities or even individual solar panels, providing detailed data. The quality and completeness of this data varies based on local user activity. We scrape over 2 million solar installation geometries from OSM. Of these, we retain only those larger than 1,000 m^2. This results in a training set of around 122 k polygons. We further collect a globally distributed set of 20,000 'hard negative' images that do not contain solar plants to reduce the number of FPs generated by the model at inference time. A summary illustration of the data collected and its global distribution can be found in Figure 1. For evaluating model performance we use the test-set developed by Kruitwagen et al. This high-quality timestamped polygon dataset is formed by exhaustive manual inspection of large areas of interest in satellite imagery. This dataset covers approximately half a million km^2 of globally-diverse areas-of-interest and identifies 7,263 solar projects. Testing against this dataset allows us to understand our false negative rate: how many solar facilities exist that we are unable to detect. Figure 2 shows a sample of the areas selected for manual inspection across Europe.

Asset Level Solar Capacity Data Our primary source for asset level solar capacity data used in our modelling is from OSM. Capacity values, attached to either solar 'nodes' or 'ways' under tags *capacity*, *plant:output:electricity*, or *generator:output:electricity*, require a degree of processing prior to use. We apply a range of checks to extract outliers and ensure consistent units and formatting. We partnered with GEM, to cross-check OSM values with those of GSPT-listed plants[4] validating the process. This yielded several thousand solar facilities from OSM with both a listed capacity and defined boundary polygon.

Aggregate Level Solar Capacity Data Aggregate level, either country or global, data is used in this study for benchmarking and validation purposes. There are three sources used: GEM's GSPT, Standard & Poor's (S&P)'s Global Commodity Insights (GCI) and IRENA. As previously discussed, GSPT's December 2023 release features an asset-level dataset with global coverage of solar facility phases exceeding a capacity of 20 MW or more and partial coverage of assets between 1 MW and 20 MW. Assets are tracked via government data, company statements, media reports and other non-governmental organisations[6]. New data releases are produced on a bi-annual update schedule. The GCI is an information provider for the energy and commodities markets. It provides project level capacity estimates and installation forecasts with information sourced from market surveys and industry reports. Lastly, IRENA provides national level statistics based on official government data either sourced from national reports, surveys or via informed estimates based on analysis and is updated on an annual basis with a year delay in reported data[15].

2.2 Solar Facility Detection and Segmentation

To develop our utility solar PV asset database we utilised a CNN based approach. CNNs are a class of machine learning models that are typically applied to image data. This is combined with a UNet architecture which is capable of utilising these learned features to generate a segmentation mask. This will, for example, define the likelihood of any given pixel belonging to the solar PV class. In this section we will outline the approach utilised in developing the solar PV segmentation model and the results of this process.

Model Training Our training process uses CNN for image processing tasks. We utilise two libraries: segmentation_models.pytorch and TorchGeo. We train our models using a Sentinel-2 image and solar mask derived from OSM datasets. Our best model uses the UNet++ encoder-decoder architecture with ResNet-50 encoder. This encoder is pretrained on Sentinel-2-derived SSL4EO-S12 dataset using a Momentum Contrast task. These weights are published under the CC-BY-4.0 licence by Wang et al[16]. We split the data into non-overlapping subsets, allowing us to train and predict on distinct geographical regions. Our segmentation model serves two purposes: finding solar facilities (detection) and drawing boundaries around them (segmentation). We measure its performance on both tasks across multiple size bins, noting that performance is likely to be a strong function of plant size. We evaluate our models on the detection task using plant-level recall, and on the segmentation task using the Intersection Over Union (IOU) and pixel-level precision. We want to generate the most complete (i.e., highest recall) dataset possible. This is a trade-off against precision. The lower the precision, the more time and resources will be spent on manual verification work. Following experimentation, we opt to binarise our predictions at a threshold of 0.95. This yields high precision for the 1-100+ MW ranges while maintaining a relatively high plant recall. From this we expect to find around 70 % of plants between 1 and 20 MW and 90 % of plants above 20 MW.

Model Inference For a global inference run we collect satellite images for the entire land surface of the earth between +70 and -60 latitude. We process around 3 million image chips, each covering a 2.5 by 2.5 km square, with our best performing segmentation model and apply a threshold to produce a binary mask. We process this mask with an erode-dilate step to smooth borders and remove very small predictions. We convert each contiguous detection into a geo-referenced polygon and save it to our database for further processing. This pipeline processes around 100 TB of Sentinel-2 data for a global deployment. We run inference in 16 hours on 2,000 CPUs at an approximate cost of £600 per run.

Manual Pruning Each global inference run produces several hundred thousand polygons. We expect our model to generate FPs. For large polygons (100+ MW) we expect around 10 % of detections to be false. For small polygons (<1 MW) this rises to around 90 %. To maximise the utility of our dataset, every polygon we publish is reviewed either manually or by reference with existing polygon datasets. Due to the scale of the task we built an in-house labelling tool which allowed for user processing speeds of up to 3,000 images per hour. The outputs of this validation step are a human verified label for a given polygon - True, False or Unknown. An illustration of the tool as presented to the labellers is displayed in Figure 3. Instructions and logs are provided in the terminal while the user is presented with 2 images for each polygon: a close up fit to the polygon size, and a wider shot to provide contextual information. The first release of this dataset required approximately 400,000 manual validations, at around 10 full days worth of labelling work. While this is significant, it is far less than the manual work required to construct a traditional asset-level dataset of a similar size.

2.3 Solar Construction Date Estimation

Much of the tooling developed for the Solar Facility Detection and Segmentation task was of use in the estimation of solar construction dates. By analysing the confidence of our solar detections over time, across a set of defined periods, we were able to infer plant construction. The value of this attribute is to allow for interpretations of changes in global, national and region-level solar capacity over time. It also has downstream applications for estimating the efficiency and expected retirement date of a facility. An inherent limitation of this approach is that construction date estimates are only available from 2017 onwards, owing to the relatively recent deployment dates for the Sentinel-2 satellites.

Model Training For model training we opted to develop our own training set due to the increased confidence this provided us in precise construction dates. To achieve this we made use of a modified version of the quicklabel tool (see Figure 3) to present the user with a series of Sentinel-2 based images for known solar plants. For this task 1,000 composite images were sampled from our validated correct solar detections back through time until 2017. We generate annual composites for each year from 2017 to 2022 inclusive, and quarterly composites for 2023 and beyond. The user is then required to annotate the first instance/period that a completed solar facility is present. Labelled data is then split into a training set that was used to develop the model and a testing set which was used to evaluate model performance. Model selection was an experimental process and best results were produced by monitoring segmentation overlap of the historical mask with the most recent prediction. When this overlap increases above 10 %, we mark the plant as constructed.

Model Inference At inference time we submit a list of positive assets as confirmed by the capacity model and already pruned via our manual verification process. This selective approach substantially reduces the cost compared to applying a global inference run at each historical time point. For each facility we estimate the construction date by determining an upper and lower bound. The upper bound is the date of the image in which the plant was first seen in a constructed state, and the lower bound is the date of the image in which the plant was last seen in an unconstructed state.

2.4 Solar Capacity Estimation

We train an additional model to estimate the capacities of the facilities we detect. The Alternating Current (AC) capacity of a plant is calculated using the following formula:

$$C_{AC} = A \times I \times \eta \times GCR \times ILR \tag{1}$$

where C_{AC} is the AC capacity of the plant, A is the Plant Footprint (m^2), the total area occupied by the solar plant. I is the Nominal Solar Irradiance (applied at 1 $\frac{kW}{m^2}$), the amount of solar power received per unit area. η is the Panel Efficiency (10-20 %), the efficiency with which the solar panels convert solar irradiance into electrical power. GCR is the Ground Coverage Ratio (GCR), the ratio of the total panel area to the total plant footprint, typically ranging from 20-80 %. ILR is the Inverter Loading Ratio (ILR), the ratio of the AC capacity to the Direct Current (DC) capacity of the plant. Previous work in the area tended to use global assumptions for the panel efficiency, GCR and ILR. We conducted an analysis into these assumption finding that GCR varies substantially country-to-country and for different plant sizes (see Figure 4). As a direct result we applied improvements upon this previous approach by using a model that accounts for these factors.

Model Training To achieve these improvements we generate a dataset of over 7,000 solar facility polygons linked to capacities from OSM. Around a third of these were contributed by labelling organised by GEM. OSM is known to have occasional data reliability issues. We clean the dataset by first deriving the approximate GCR of each plant and removing any plants that fall outside of the range 5-95 %. We also exclude plants below 1,000 m^2. Finally, we manually inspect any remaining outliers and remove any that are clear annotation mistakes. The resulting dataset allows us to study solar plant ground coverage ratios in detail. We use 5-fold cross-validation to estimate the expected performance of our model on unseen data, with each fold containing a similar distribution of solar facility sizes and country locations. We optimise model performance against the Root Mean Squared Error (RMSE) validation metric.

Model Inference Given a geo-referenced solar asset in the form of a polygon or multipolygon we are able to make an inference on its associated capacity. This is designed such that it can work efficiently with the output of our Solar Facility Detection and Segmentation model or any other polygon based dataset.

3 Data Records

This section provides details on the data records associated with this work. It includes a description of each data file, its format, and its location in the repository. Each external data record is cited numerically within the text and referenced in the main reference list. Additionally, data citations are placed in the Methods section, specifying the data-collection or analytical procedures used. Our analysis-level dataset provides a comprehensive view of global asset-level solar installations, incorporating our detections and known solar facility geometries from external datasets. The analysis-level datasets mask underlying complexities, which we expose in the `raw_polygons` and `sources` files. These files capture overlapping and clustered geometries, essential for tracking raw detections and providing detailed sourcing information. Our clustering process combines overlapping and nearby geometries from various sources, including large solar facilities from OSM and validated geometries from Kruitwagen et al.[7]. Each cluster in the analysis-level dataset corresponds to a single row. To facilitate traceability and sourcing, we provide the raw polygons and a source file detailing the contents of each analysis-level polygon. Our files are located in the following file formats and online repositories, specific contents can be found in listed tables:

`analysis_polygons.gpkg`
Description: Our primary "analysis-ready" dataset with geometries, capacities, and construction dates.
Format: GeoPackage (.gpkg)
Location:
`https://zenodo.org/records/11368204/files/2024Q1_final_analysis_polygons.gpkg`
Table: 1

`analysis_polygons.csv`
Description: A `.csv` version of `analysis_polygons.gpkg`, facilitating parsing without geospatial software.
Format: Comma-Separated Values (.csv)
Location:

https://zenodo.org/records/11368204/files/2024Q1_final_analysis_polygons.csv
Table: 2

sources.csv
Description: A table mapping the IDs of the analysis-ready dataset to the source specific IDs comprising them.
Format: Comma-Separated Values (.csv)
Location:
https://zenodo.org/records/11368204/files/2024Q1_final_sources.csv
Table: 3

raw_polygons.gpkg
Description: A table mapping the source IDs to the raw geometries comprising them.
Format: GeoPackage (.gpkg)
Location:
https://zenodo.org/records/11368204/files/2024Q1_final_raw_polygons.gpkg
Table: 4

4 Technical Validation

The integration of segmentation, capacity and construction date modelling have been used to create an exhaustive and rich dataset of the world's solar energy facilities. By training the capacity model on a diverse range of solar installations we have created an accurate and scalable approach to deliver global capacity estimates. Here we discuss the validation of our approach and the resulting dataset and present considerations for improvement and any avenues that may be taken in further work.

Segmentation Model Despite the low-resolution limitations of satellite images, which impede the detection of smaller installations, the CNN segmentation model demonstrates a promising capacity for distinguishing solar arrays within varied landscapes. Additional considerations such as atmospheric conditions like cloud cover at times impeded image clarity and may have subsequently biased the model towards predictions of larger installations. Image compositing was introduced to overcome the limitations of atmospheric conditions and low-resolution imagery. Composite images substantially improved modelling performance but brought introduces a temporal lag whereby sufficient imagery must accumulate before a sufficiently clear composite can be made. In some cases a composite may consist of several images which can result in recent changes - such as the development of a solar facility - being lost. The effectiveness of this model to capture very recent solar developments, say in the last few months prior to deployment, is still being explored.

Model performance is assessed against a range of metrics and for varying plant capacity values against the test dataset developed by Kruitwagen et al. This gives us a strong indication of what to expect from our model in practice. We were able to optimise our model to maximise recall while minimising precision loss based on these results. We show the results in Figure 5 where the IOU, precision and plant-level recall are broken into multiple bins according to plant size. The majority of global capacity is covered by the 20-100 MW and 100 MW+ bins, while the 1-10 MW and 10-20 MW bins contain large numbers of plants not published elsewhere. The total number of detection's are dominated by these smaller bins. After model inference our dataset underwent a manual pruning effort to remove FP from the dataset - however some challenges remain due to the difficulty of manually validating detections in 10 m satellite imagery. To estimate final FP prevalence throughout the data a subset of approximately 2,000 detections were selected at random from our positively labelled solar assets. Each of these were validated through a higher degree of scrutiny utilising high-resolution imagery - yielding a rate of FPs at around 1 %.

Capacity Model The capacity modelling framework offers a detailed methodology for estimating the power output potential of identified solar facilities and offers an approach that goes above scaling of a bounded polygon area. The DC capacity of a solar panel is the product of its size, local solar irradiance, and its efficiency. For a solar facility made up of multiple arrays, the total surface area of the arrays is often expressed as the ground area of the plant multiplied by its GCR — the ratio of array area to ground area. For utility-scale solar that is connected to the grid we are often concerned with its AC capacity, which is additionally dictated by the size of its inverter - standard practice in this case being to size the inverter 10-30 % smaller than the DC capacity of the plant. Our model expands upon previous efforts to estimate capacity by recognition of the influence country and plant size have on the GCR (see Figure 4). To evaluate the performance of our model we apply the RMSE metric which is a measure of the difference between the predicted and actual values. It is calculated by taking the square root of the average of the squared differences between the predicted and actual values. We select the model with the best RMSE on plants between 0.01–0.1 km^2 (around 1–10 MW) since this is the region where we expect our pipeline to be most useful. Secondly, we select for models with good performance on larger plants and less complexity. Table 5 shows our model performance according to the RMSE metric (average of 5-fold cross-validation) for all samples across our testing set. We compare our model performance to that of the constant GCR model for plants across three bins: $\leq$0.01 km^2, 0.01–0.1 km^2, and >0.1 km^2. These bins correspond to plants with capacities approximately of $\leq$1 MW, 1-10 MW, and >10 MW respectively. We see that our model outperforms the constant GCR model across the larger two bins while there is little to no difference in the smallest bin. This is a substantial improvement on the constant GCR model particularly in the larger bins which are of greatest importance as they correspond to a larger share of overall global solar PVcapacity. There are still limitations in this approach however as it relies only on geo-referenced polygons as a basis for capacity estimates. This can introduce complexities in cases with unusual or nuanced solar facility layouts - plants with unusually high or low ground coverage ratios will not have accurate capacity estimates.

Additionally, the model has no way to distinguish between plant technology types, e.g. dual-axis-tracking or fixed, plant in the same country which will impact solar capacity over a given footprint even with a known GCR (an example illustration in Figure 6).

Construction Date Model Understanding the construction dates of solar facilities is crucial for various analyses, including assessing changes in solar capacity over time and estimating the potential lifetime of existing plants. We develop a model pipeline to predict the construction dates of solar plants using quarterly time series predictions spanning from 2017 to 2024. This approach provides valuable insights into the change of solar infrastructure. We estimate the construction date of a plant based on estimating an upper and lower bound where the range signifies the possible period in which the solar facility was constructed. For plants that were constructed before the launch date of Sentinel-2 in 2017, we produce only an upper bound. This process allows us to measure construction dates to the nearest quarter for more recent plants. The results of this work are assessed by way of an in-house developed validation set of 1,000 solar facilities. Figure 7 shows the output of these results. The model predicts the exact period (year) 92 % of the time and the one-off-error (within ± 1 year) of 97.8 %. There are some limitations and caveats associated with our methodology that should be considered. For example, the predictions are based on quarterly composite images, which may be taken from any period within each quarter. This introduces uncertainty regarding the precise timing of plant construction within the detected quarter. Consequently, our predicted date represents the earliest quarter in which the plant was identified by the model, rather than the exact construction date. It is plausible that the plant could have been built earlier, potentially in the preceding quarter. Additionally, since our dataset begins in 2017, we cannot determine the construction dates of plants built prior to this year. This temporal constraint restricts the applicability of our model to historical solar infrastructure. If a solar plant is detected in every time frame considered, it is likely to have been built prior to 2017, and we are therefore unable to predict its construction date. Lastly, we provide an estimated error range or confidence interval associated with our predictions not a specific date. While we aim to provide construction dates to the nearest year or quarter, there may be inherent uncertainties in the model outputs, leading to a margin of error.

TZ-SAM Dataset Overall our top-level datasets contain 63,616 assets, with a total area of 19,121 km^2 and a total estimated capacity of 711 GW. Each of these assets have been validated either by reference with an existing dataset, or by manual inspection. For each solar asset we provide a capacity estimate which ranges from 0.4 MW for the smallest asset and 5,044 MW for the largest. In total 87 % of these are provided with a construction date range. In order to validate our work we make efforts to compare our dataset to other asset and country level datasets that are available. Firstly, we analyse our data compared to GEM's GSPT asset level dataset according to capacity ranges (see Figure 8). Solar PV facilities are split into the following groups based on their capacity sizes: 0-1, 1-10, 10-20, 20-100, 100+ MW. For the GSPT an asset can either be 'exact' where the location of the plant is precisely known and can be located on a map, or 'approximate' where the location of a plant is typically given as the centroid of its listed country or region. All of our own assets are exactly geo-located and we therefore compare to both stated values for clarity. In total we find 63,616 solar facilities at an estimated capacity of 711 MW compared to GEM's 23,899 operating facilities with stated capacity of 620 MW. Assets with "approximate" locations comprise 41 % of GEM's total listed assets and 59 % of their listed capacity. Our overall estimated capacity is higher than GSPT, however the breakdown shows that this is attributed to greater capacity in the 0-20 MW ranges while our capacity for larger sized plants is lower. It can generally be observed that TransitionZero (TZ) capacity estimates are lower when adjusting for facility count within a given capacity group. Our predicted capacity per facility is on average between 76 % and 89 % that of GSPT depending on the group in question. Secondly we make efforts to compare to country level aggregate datasets. This is both in the form of direct comparisons, such as in Table 6 and Table 7 in addition to an analysis of solar PV development over time as in Figure 7. For direct comparisons against GSPT, GCI and IRENA we see that ours is close to that of GSPT in all of Global, USA, China and EU27+GB regions. As initially implied in Figure 8, this breakdown reinforces that we find more capacity when looking at <5 MW range but shows consistency across all regions. Our data shows markedly greater capacity relative to the GCI dataset in almost all categories except the USA in which results are comparable. IRENA serves as an approximate upper bound for capacity comparisons. Values are sourced from government or privately reported statistics or surveys and then validated based on expert opinion and trends. Despite this it demonstrates our relative model performance in each region, indicating that we perform relatively best in the USA, capturing 75 % of the IRENA stated capacity. In Table 7 compare our values against IRENA for the top 20 countries according to total capacity. We can also utilise our construction date estimates to form a regional trend analysis as illustrated in Figure 7. This highlights that we are capturing the relative development of solar in each region in line with IRENA's published figures while operating at an improved recency with our dataset release from an annual to quarterly lag.

References

1. International Energy Agency. Renewables 2023. Tech. Rep., IEA (2024).

2. International Renewable Energy Agency. Renewable capacity statistics 2023. Tech. Rep., IRENA (2024).

3. International Energy Agency. World Energy Outlook 2023. Tech. Rep., Paris, France (2023). https://doi.org/10.1787/20725302.

4. Global Energy Montor. Global Solar Power Tracker retrieved from. https://globalenergymonitor.org/projects/global-solar-power-tracker/ (2023).

5. OpenStreetMap contributors. Planet dump retrieved from https://planet.osm.org . https://www.openstreetmap.org (2024).

6. Global Energy Monitor. Methodology retrieved from. https://globalenergymonitor.org/projects/global-solar-power-tracker/methodology/ (2024).

7. Kruitwagen, L. *et al.* A global inventory of photovoltaic solar energy generating units. *Nature* **598**, 604–610, 10.1038/s41586-021-03957-7 (2021).

8. Malof, J. M., Bradbury, K., Collins, L. M. & Newell, R. G. Automatic detection of solar photovoltaic arrays in high resolution aerial imagery. *Appl. Energy* **183**, 229 – 240, https://doi.org/10.1016/j.apenergy.2016.08.191 (2016).

9. 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.

10. 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).

11. 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, https://doi.org/10.1016/j.joule.2018.11.021 (2018).

12. Ortiz, A. *et al.* An artificial intelligence dataset for solar energy locations in india. *Sci. Data* **9**, 497 (2022).

13. Xia, Z. *et al.* Mapping the rapid development of photovoltaic power stations in northwestern china using remote sensing. *Energy Reports* **8**, 4117–4127 (2022).

14. Ong, S., Campbell, C., Denholm, P., Margolis, R. & Heath, G. Land-Use Requirements for Solar Power Plants in the United States. Tech. Rep. NREL/TP-6A20-56290, 1086349 (2013). 10.2172/1086349.

15. International Renewable Energy Agency. Data Methodology. Tech. Rep., IRENA (2015).

16. Wang, Y. *et al.* Ssl4eo-s12: A large-scale multimodal, multitemporal dataset for self-supervised learning in earth observation [software and data sets]. *IEEE Geosci. Remote. Sens. Mag.* **11**, 98–106 (2023).

17. S&P Global. Global Clean Energy Technology (2024).

Author contributions statement

M. P. Writing - Draft, Software, Investigation, Analysis, Data Curation, Visualisation

J. O'C. Conceptualisation, Software Lead, Investigation, Analysis, Data Curation, Writing - Review and Editing

A. F. Software, Investigation, Analysis, Data Curation

M. S. Software, Investigation, Analysis, Data Curation, Writing - Editing

L. K. Conceptualisation, Funding Acquisition, Writing - Draft, Review and Editing, Analysis, Software Troubleshooting

M. G. Analysis, Writing - Review

Figures & Tables

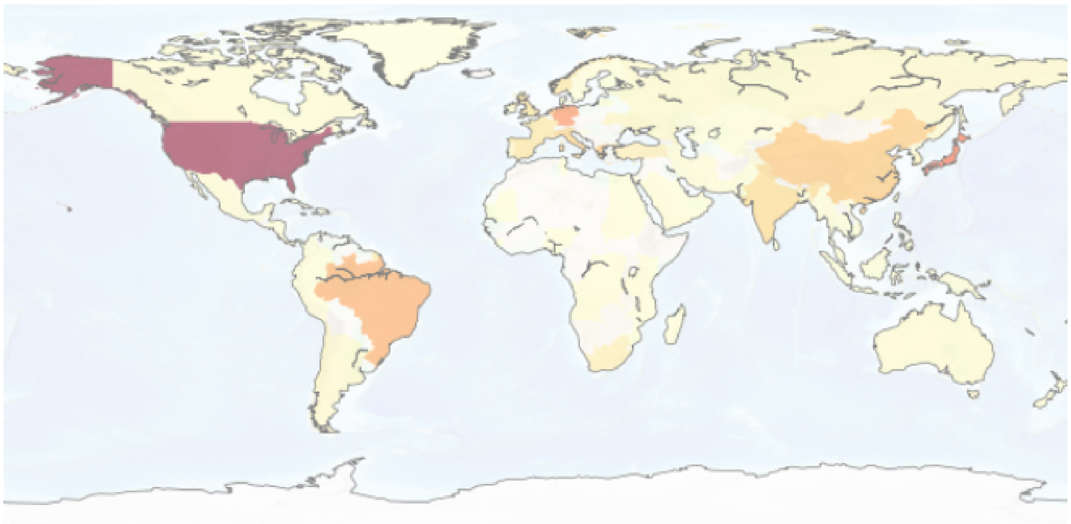

Figure 1. A map visualising the distribution of our OSM sample set. This numbers 122 k in total. The top 3 countries by sample count are USA (19 k), Japan (12 k), and Germany (10 k). In contrast (not displayed) by aggregate solar PV area the largest country is China (2,500 km^2) followed by USA (1,500 km^2) and India (800 km^2).

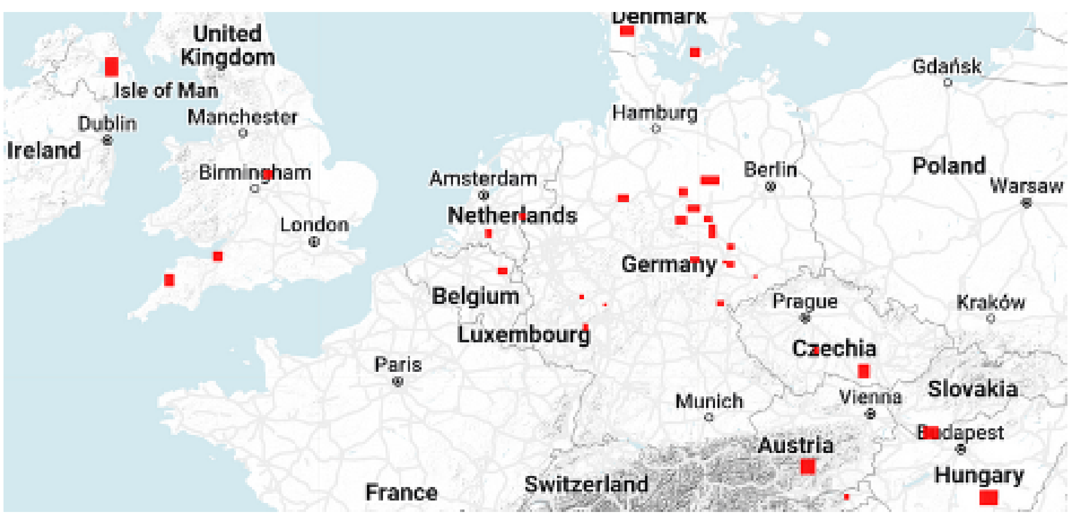

Figure 2. A sample of geometries covered in Kruitwagen et al.'s hand-labelled test set[**?**]. Within each red geometry an exhaustive search for utility solar facilities was performed which produced a high quality and high confidence data set for testing purposes.

Figure 3. An overview of the TZ Quicklabel tool. Developed in-house, it allowed for a substantial degree of customisation which was required to complete the labelling task efficiently. **Top left**: the initial terminal display. **Bottom right**: a sample image provided to the labellers for a given solar PV predicted polygon (note the polygon outlined in red).

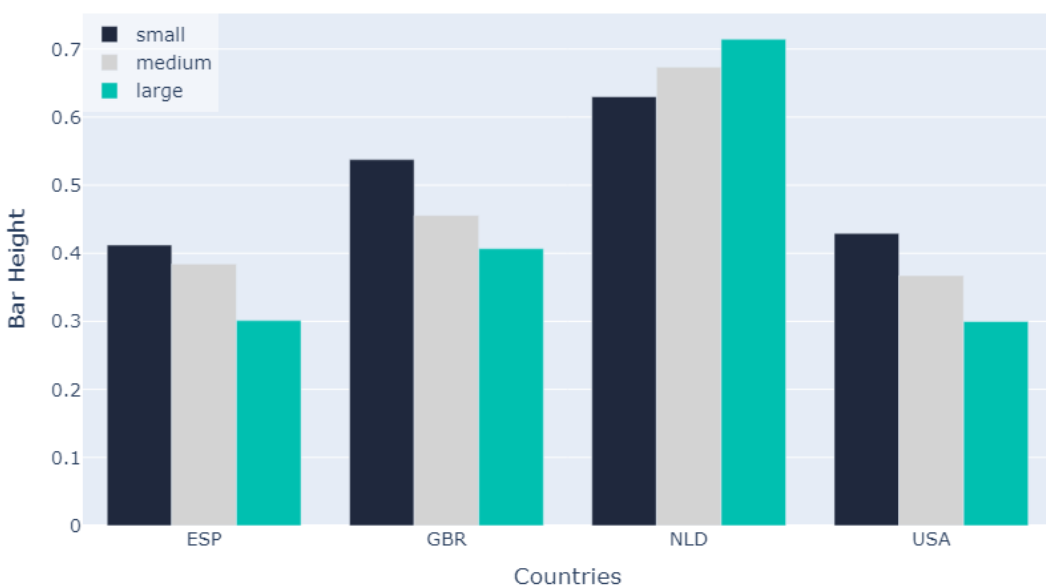

Figure 4. A comparison of four countries and their associated mean GCR for solar PV facilities of different sizes.

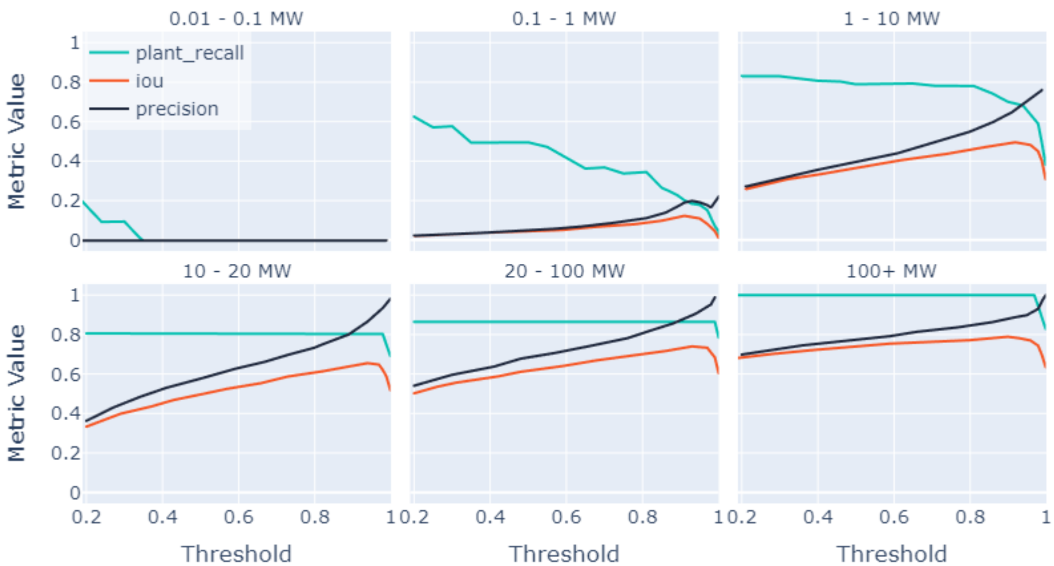

Figure 5. Performance of model by plant capacity and threshold.

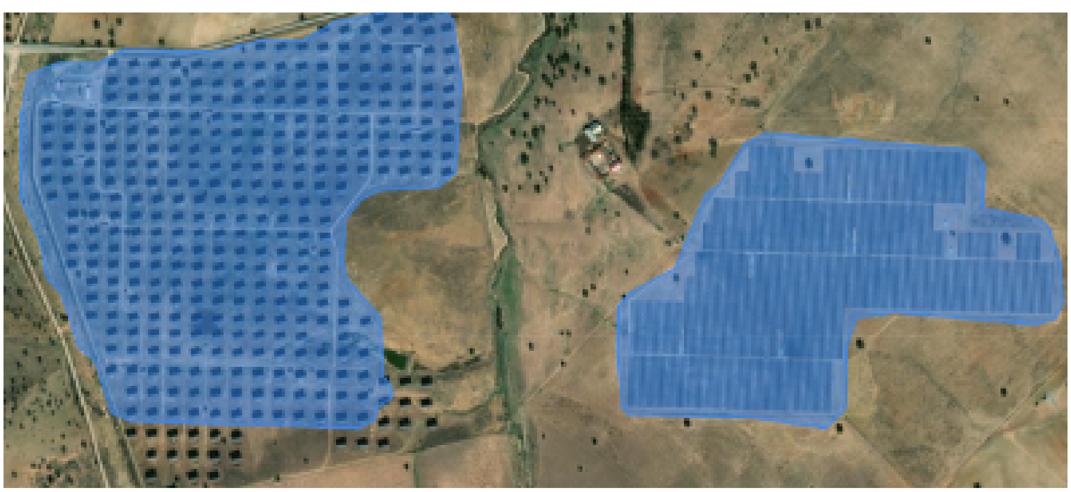

Figure 6. **Left**: a dual axis facility. **Right**: a static facility. The static facility has a notably higher GCR and therefore greater capacity. This is not directly captured however. GCR - and by extension capacity - estimates for both of these facilities are based on the size of the facility and country of origin.

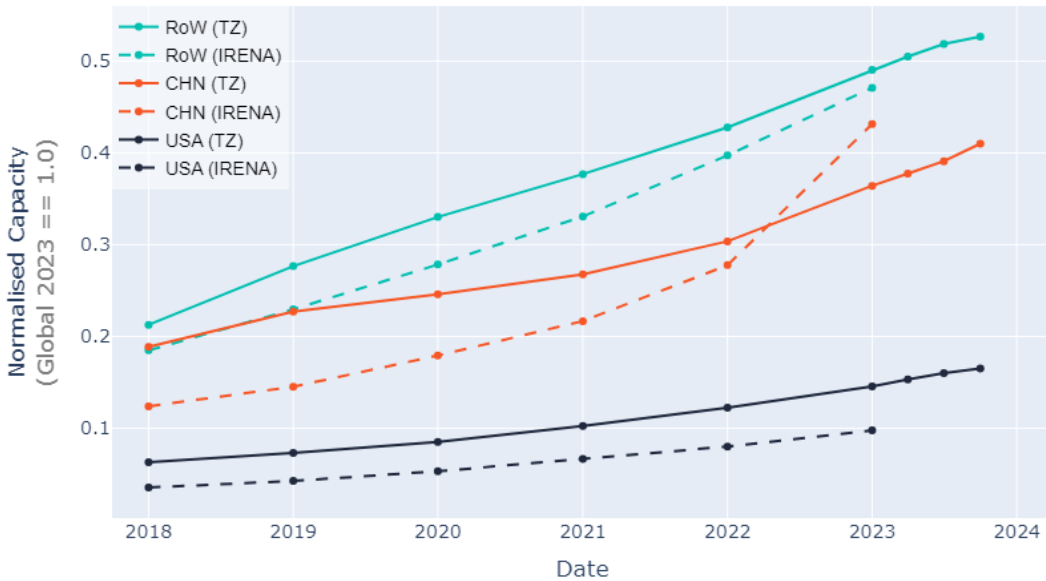

Figure 7. A comparison of solar PV capacity expansion for the United States of America (USA), China (CHN) and Rest of World (RoW). Expansion rates are compared between TZ and IRENA with values relative to start of 2023. Prior to 2023 TZ construction date model runs were conducted at annual cadence, Since 2023 model runs are conducted at quarterly cadence.

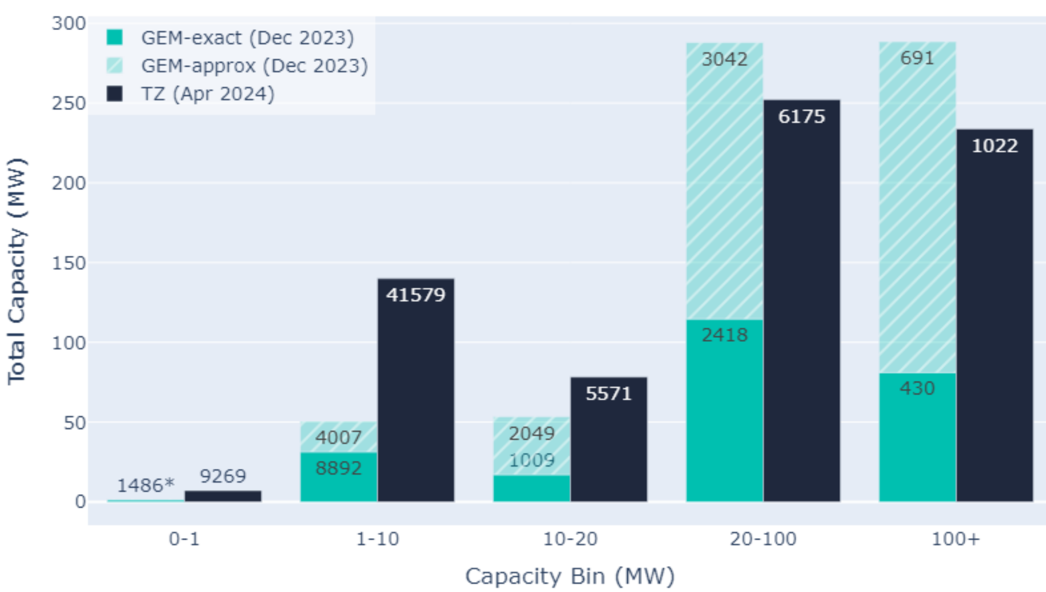

Figure 8. A comparison of total identified solar PV capacity between TZ and GEM's GSPT (filtered for operating plants only) for different capacity ranges. The GEM dataset is split into facilities with "exact" and "approximate" location accuracy. These two categories are combined in the 0-1 MW group for visibility.

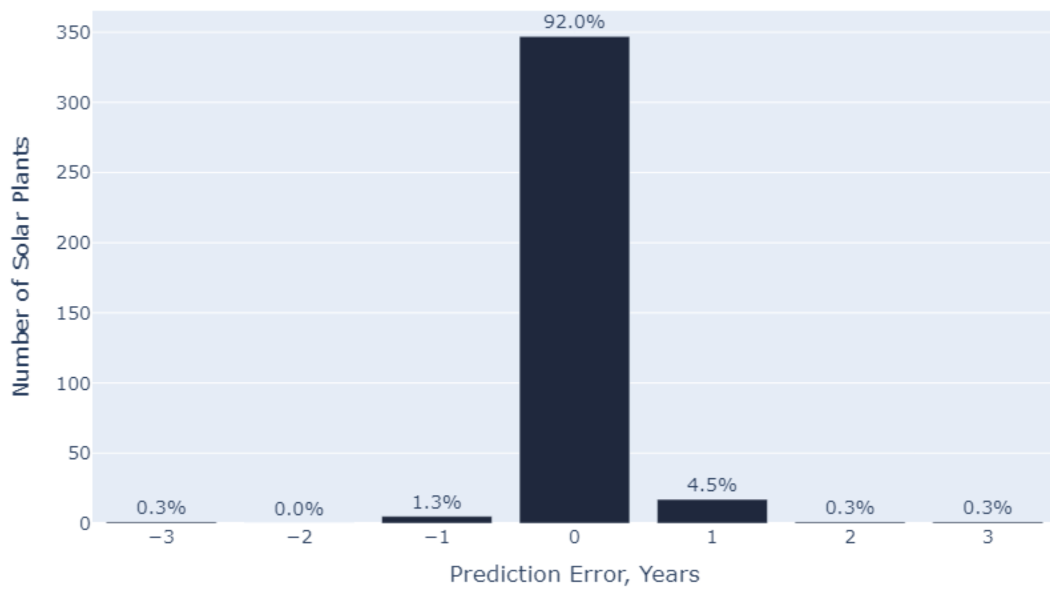

Figure 9. A demonstration of the construction date model performance on a set of 377 known utility scale solar PV facilities. It achieves an accuracy of 92 % to the correct time period.

Table 1. Fields in `analysis_polygons.gpkg`.

Field	Type	Description
`id`	INTEGER	Unique ID for the asset.
`geometry`	GEOMETRY	Polygon or MultiPolygon defining the asset.
`capacity_mw`	FLOAT	Estimated capacity of the asset in megawatts.
`constructed_before`	DATE	Upper bound for construction date.
`constructed_after`	DATE	Lower bound for construction date.

Table 2. Fields in `analysis_polygons.csv`.

Field	Type	Description
`id`	INTEGER	Unique ID for the asset.
`latitude`	FLOAT	Latitude of the centroid of the asset.
`longitude`	FLOAT	Longitude of the centroid of the asset.
`country`	TEXT	Administrative country name.
`capacity_mw`	FLOAT	Estimated capacity of the asset in megawatts.
`constructed_before`	DATE	Upper bound for construction date.
`constructed_after`	DATE	Lower bound for construction date.

Table 3. Fields in `sources.csv`.

Field	Type	Description
`cluster_id`	INTEGER	Corresponding ID from `analysis_polygons.*`.
`source_id`	INTEGER	Source specific ID of the raw polygon.
`source`	TEXT	Original source of the raw polygon.
`acquisition_date`	DATE	Detection/acquisition date of the asset.

Table 4. Fields in `raw_polygons.gpkg`.

Field	Type	Description
`id`	INTEGER	Source specific ID of the raw polygon.
`geometry`	GEOMETRY	Polygon or MultiPolygon defining the asset.
`source`	TEXT	Original source of the raw polygon.
`acquisition_date`	DATE	Detection/acquisition date of the asset.

Table 5. Solar capacity model cross-validation performance by plant size. Here we compare two models: TZ capacity model and the Constant GCR model for different plant size ranges. Capacity range is given as a guideline based on the Area range. Capacity varies according to geographical location in addition to area and is therefore not directly proportional.

		Range		
Plant size	Area, km^2 Approximate Capacity, MW	$\leq$0.01 $\leq$1	0.01–0.1 1–10	>0.1 >10
TZ capacity model	RMSE, MW	0.15	1.05	20.40
Constant GCR capacity model	RMSE, MW	0.15	1.36	46.49

Table 6. Comparison of Solar Generating Capacity Datasets

	Global		USA		China		EU27+GB	
	<5MW	>=5MW	<5MW	>=5MW	<5MW	>=5MW	<5MW	>=5MW
TZ-SAM	85 GW (41,979)	626 GW (21,637)	11 GW (5,649)	93 GW (2,447)	1 GW (5,440)	246 GW (6,706)	32 GW (16,646)	87 GW (5,566)
GEM[1]	22 GW (9,695)	666 GW (12,612)	7 GW (3,120)	72 GW (1,831)	0.004 GW (2)	294 GW (4,003)	14 GW (6,086)	119 GW (2,964)
S&P Global[2]	16 GW (7,594)	308 GW (7,129)	8 GW (4,160)	95 GW (2,293)	0.1 GW (57)	39 GW (510)	5 GW (2,025)	46 GW (1,957)
IRENA[3]	1,419 GW		139 GW		610 GW		272 GW	

[A] Figures in parentheses indicate the number of facilities in the dataset.
[1] *Global Energy Monitor (2023), Global Solar Tracker*[4]. Figures presented in this table include all 'certain' and 'uncertain' facilities. Data from Luxembourg, Malta, and Slovenia are not present. TransitionZero is currently working with GEM to include TZ-SAM in future GEM Solar Tracker releases.
[2] *S&P Global Commodity Insights (2024), Global Clean Energy Technology*[17]. 6,491 solar assets do not have a capacity estimation and are excluded from analysis.
[3] *IRENA (2024), Renewable Capacity Statistics*[2]. IRENA provides aggregate country-level capacities only.

Table 7. A comparison of total identified solar in IRENA (2024) and TZ (April 2024) datasets. Top 20 largest countries by TZ capacity are shown.

Country Code	IRENA (2024)	TZ (Apr 2024)
CHN	609,350	255,089
USA	137,725	104,012
IND	72,766	54,472
JPN	87,068	38,315
ESP	28,712	26,561
DEU	81,737	24,967
AUS	33,680	15,069
ITA	29,789	13,436
BRA	37,449	13,037
TUR	11,291	11,574
GBR	15,656	11,540
MEX	10,893	10,796
FRA	20,542	9,935
CHL	8,366	9,604
VNM	17,077	9,358
KOR	27,046	7,130
UKR	8,062	5,366
ZAF	5,664	5,083
NLD	23,904	5,039
POL	15,809	4,622