Point-by-point response to Reviewer 1

#C1.1

Spelling and Grammar Issues
ZZ 45-46
Original: "The codes and dataset … is available."
Correct: "The code and dataset … are available."
ZZ 83
Original: "there is a geospatial wind turbine dataset for 2020 is introduced."
Correct: "A geospatial wind turbine dataset for 2020 was introduced."

Response:

Thank you very much for your comments. We have corrected the spelling and grammar errors as you suggested (lines 51-52, line 79), and conducted a full-text review to address similar issues (line 118, line 438-439).

"The codes and dataset of the global onshore wind turbines are available at …" (lines 51-52)

"At the global scale, a geospatial wind turbine dataset for 2020 is introduced " (line 79)

"…, combining with validation through…" (line 118)

"Compared to the current datasets of available global onshore wind turbines, our dataset provides…" (line 438-439)

#C1.2

Unclear Description: OSM Query
The query string in line 129 (["generator: source"="wind"]) appears syntactically incorrect.
OpenStreetMap commonly uses:
generator: source=wind
or power=generator combined with generator: source=wind
The manuscript should clearly state the exact Overpass query used.

Response:

Thanks very much for pointing this out. We revised the description of the OSM query operation accordingly (lines 128-129).

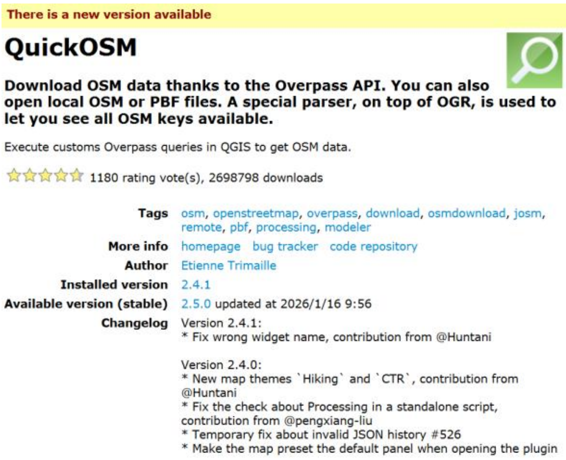
" …through the QuickOSM plugin (based on the Overpass API) in QGIS software with the query parameter: generator: source=wind." (lines 128-129)

#C1.3

Major Comment: Missing Methodological Detail on OSM Extraction
The manuscript does not specify whether nodes, ways, or relations were extracted.
It is unclear whether both power=generator and generator: source=wind were used.

Response:

Thank you very much for your valuable comment. As mentioned in #C1.2, we used generator: source=wind for OSM data extraction. The query is executed using the QuickOSM plugin in QGIS (as shown in the figure below), which leverages the Overpass API. Accordingly, the data were downloaded via the query interface within the QGIS software. We further clarify the data extraction process by adding a description of this plugin in the manuscript (line 129).



" …through the QuickOSM plugin (based on the Overpass API) in QGIS software with the query parameter: generator: source=wind." (line 129)

The handling of offshore turbines is not described.

Response:

Thanks very much for pointing this out. As this study focuses on onshore wind turbines, the initial data extraction was spatially constrained using OSM land polygons. Consequently, offshore wind turbines were excluded from the dataset (lines 132-134).

"Since we focus on onshore wind turbines, OSM land polygon derived from https://osmdata.openstreetmap.de/data/land-polygons.html is used to define the study extent and refine the dataset." (lines 132-134)

No link to the exact Overpass script is provided.

Response:

Thanks very much for your comment. As clarified in #C1.3, the data were queried and downloaded using the QuickOSM plugin within QGIS software, which provides a pre-packaged interface for querying.

There is no information on whether wind farms, meteorological masts, or power-line

infrastructure were filtered out.
These omissions significantly limit reproducibility and must be clarified.

Response:

Thanks very much for raising this issue. Given our focus on the individual wind turbine level, we utilized the Overpass query generator: source=wind filters for nodes representing wind turbines as point features. We further clarify it in the Methods section (lines 128-132).

"...through the QuickOSM plugin (based on the Overpass API) in QGIS software with the query parameter: generator: source=wind. Given our focus on the individual wind turbine level, we utilized this query filter for nodes representing wind turbines in the format of point features." (lines 128-132)

#C1.4

Major Comment: Unexplained OSM Error Rate
The manuscript states a "10% error rate in OSM's global wind turbine dataset" but provides no methodological explanation.
Missing information:
How was this error rate calculated? What was the validation procedure?

Response:

Thanks very much for pointing this out. We apologize for the omission of this explanation. We have added a detailed description of the error rate calculation for the OSM data in the Methods section (lines 257-266). This includes a thorough explanation of the computational methodology and the validation workflow. Besides, we further incorporated an analytical discussion of these error rates based on our updated data in the Results section (lines 287-291).

"Based on our updated wind turbine dataset, we evaluated the data accuracy and errors within the OSM wind turbine records. We calculated omission and commission errors using a spatial proximity analysis in ArcGIS Pro with a 30-meter tolerance buffer. We applied a 30-m buffer to our generated points and performed a spatial selection on the OSM reference points to calculate the omission error. And the OSM points not captured within these buffers were classified as omissions. Conversely, to calculate the commission error, we buffered the OSM points and identified our generated points that fell outside these zones. The respective error rates were derived by dividing the count of omitted or committed points by the total number of OSM turbines. Finally, these two rates were summed to provide a total error rate." (lines 257-266)

"The calculated discrepancy is yielded from the omission and commission error rates of 14.4% and 4.1%, respectively. It is worth noting that this error rate represents global averages, significant regional variations could exist as the OSM data fluctuate across different countries due to varying mapping efforts." (lines 287-291)

Was the error rate spatially or regionally variable?

Response:

Thank you very much for raising this issue. The error metrics reported were calculated based on the total number of wind turbine locations globally. We acknowledge that spatial heterogeneity exists, as the accuracy and completeness of OSM data inevitably vary across regions.

To clarify the specific boundaries and limitations of this error assessment, we have added a detailed explanation in the Results section (lines 289-291).

"It is worth noting that this error rate represents global averages, significant regional variations could exist as the OSM data fluctuate across different countries due to varying mapping efforts." (lines 289-291)

Were commission and omission errors distinguished?

Response:

Thank you very much for your comment. As stated above, we conducted a quantitative assessment of the OSM data by calculating both commission and omission errors to derive a composite error rate. Accordingly, we have expanded the Methods section to include a detailed description of the analytical procedures and the computational workflow used to determine these error metrics (lines 259-266).

"We applied a 30-m buffer to our generated points and performed a spatial selection on the OSM reference points to calculate the omission error. And the OSM points not captured within these buffers were classified as omissions. Conversely, to calculate the commission error, we buffered the OSM points and identified our generated points that fell outside these zones. The respective error rates were derived by dividing the count of omitted or committed points by the total number of OSM turbines." (lines 259-266)

Was the result compared to existing studies with likewise approaches, e.g. https://www.mdpi.com/2220-9964/14/6/232

Response:

Thanks very much for your valuable comment. While we initially provided comparisons with several existing studies and public datasets, we acknowledge they may not have been sufficiently comprehensive.

We have downloaded the data from the provided links and benchmarked our results against this dataset as you suggested. These updates, including expanded discussions and comparative analyses, have been integrated into the Introduction (lines 88-90), the Results (lines 302-304, lines 308-310), and Table 1.

Besides, we further conduct comparisons with current global-scale wind turbine datasets, including Dunnett et al. (2020) and Global Renewables Watch (lines 320-334).

"...there are geospatial datasets for the United States (Rand et al., 2020), Germany

(Manske et al., 2022), Italy (Smeraldo et al., 2020), and South Africa (Kleebauer et al., 2025)." (lines 88-90)

"...along with official and research-based turbine inventories from the United States (Rand et al., 2020), Italy (Smeraldo et al., 2020), Germany (Manske et al., 2022), and South Africa (Kleebauer et al., 2025)." (lines 302-304)

"The consistency between our estimates and official records for temporally comparable years is high, with discrepancies of less than 1.8% in the United States and less than 3.4% in South Africa." (lines 308-310)

Table 1. *Comparison of open-source datasets of onshore wind turbines with our results.*

Scope	***Time***	***Number***	***Ours (2024)***
Dunnett et al. (2020)	*2020*	*33,514*	*416,532*
Global Renewables Watch (Robinson et al., 2025)	*2024(Quarter 2)*	*375,197*	*416,532*
United States (Rand et al., 2020)	*2024*	*75,781*	*74,052*
Germany (Manske et al., 2022)	*2021*	*28,156*	*29,971*
Italy (Smeraldo et al., 2020)	*2020*	*8,729*	*10,591*
South Africa (Kleebauer et al., 2025)	*2025*	*1,487*	*1,483*

"We further benchmark our dataset against the current global-scale wind turbine datasets, including Dunnett et al. (2020) and Global Renewables Watch (Table 2).. Results show that our dataset contains the largest number of identified onshore turbines while maintaining nation-level coverage and land type classification compared to Dunnett et al. (2020). In terms of data records, the Global Renewables Watch is updated to the second quarter of 2024 with 375,197 wind turbines and includes a limited number of offshore wind turbines that are not comprehensively. Our dataset focuses on onshore turbines and incorporates additional updates by the end of 2024. Methodologically, the Global Renewables Watch requires massive training datasets and substantial computational resource budget exceeding 650 V100 GPU hours to process around 14 terapixels of satellite imagery (Robinson et al., 2025). In contrast, our hybrid framework utilizes medium-to-high resolution imagery to enable global-scale updates with significantly lower computational demands. By leveraging publicly available platforms, this framework lowers the barrier to entry through a cost-effective and resource-efficient alternative." (lines 320-334)

#C1.5

Major Comment: Insufficient Documentation of Random Forest Sampling
Missing details include:
From which land cover classes were negative samples drawn?

Response:

Thank you very much for your valuable comment. We have added a detailed description of the land-use categories for the negative samples in the Methods section (lines 185-188).

"The negative samples are obtained via globally uniform random sampling to ensure spatial objectivity. The resulting dataset encompasses diverse land-cover categories including grasslands, bare land, cropland, and forests." (lines 185-188)

Was spatial autocorrelation considered?

Response:

Thanks very much for raising this issue. We employed a global-scale uniform random sampling to mitigate spatial autocorrelation. This ensured sufficient spatial separation and geographic diversity among samples, minimizing spatial dependency and maintaining sample independence. We have added this explanation in the Methods section (lines 178-180).

We further provide a visualization of the positive and negative samples to demonstrate these spatial distribution characteristics (Figure 4).

"We employ a global-scale uniform random sampling strategy. This ensured sufficient spatial separation and geographic diversity among samples, minimizing spatial dependency and maintaining sample independence." (lines 178-180)

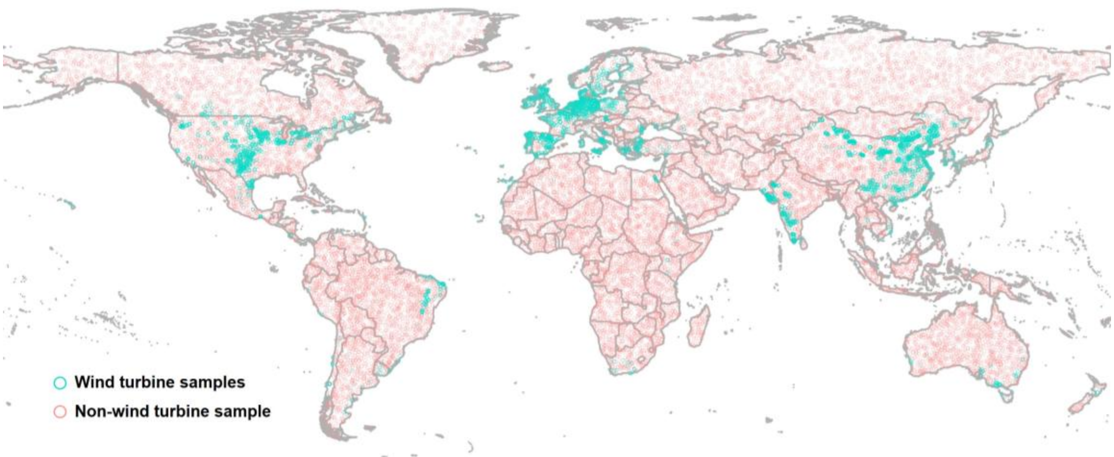


Figure 4. *Spatial distribution of wind turbine (positive) and non-turbine (negative) training samples for machine learning.*

How was it ensured that negative samples were not within 30 m of existing turbines?

Response:

Thanks very much for your comment. To implement spatial constraints, we established exclusion buffers around all existing turbines. During the sampling process, these buffered areas were masked out to ensure that no negative samples were generated within a 30-meter proximity of a positive sample. We have added this detailed description to the Methods section (lines 180-183).

"Besides, we apply a 30-meter buffer around all existing wind turbine locations (positive samples). These buffered areas are then masked out from the global sampling pool to ensure that no negative samples are drawn within this exclusion zone." (lines 180-183)

Was the global distribution of positive and negative samples balanced?
The Random Forest sampling workflow requires a clear methodological description.

Response:

Thank you very much for your comment. We utilized a 1:2 ratio of positive to negative samples for machine learning model training, which were generated through uniform random sampling. To further demonstrate their spatial balance on a global scale, we have visualized the geographic distribution of both sample sets in Figure 4 as shown above.

#C1.6

Minor Comment: Sentinel-1/2 Features and Missing GEE Scripts
The manuscript lists processing steps, but details are missing, included:
Exact satellite collections used.
Time span and temporal compositing strategy.
Cloud masking method (e.g., QA60 for Sentinel-2).
Preprocessing steps such as resampling, mosaicking, and normalisation.
Referencing Zenodo alone is insufficient; the core processing steps must appear in the manuscript.

Response:

Thanks very much for your valuable comment. We added a paragraph in the Methods section detailing the Sentinel-1/2 features and providing the GEE scripts. Specifically, we further added a detailed description of the satellite data collections used (lines 198-203), the cloud-masking method (lines 203-205), the temporal coverage and compositing strategy, and the detailed preprocessing steps (lines 205-209).

"We constructed a comprehensive feature set for machine learning based on Sentinel-1 and Sentinel-2 satellite imagery integrated via the GEE platform. We utilized the Sentinel-1 Ground Range Detected (GRD) dataset (COPERNICUS/S1_GRD), extracting the VV and VH polarization bands. We employed the Sentinel-2 Surface Reflectance collection (COPERNICUS/S2_SR_HARMONIZED), which includes the visible and near-infrared (NIR) bands. To ensure data quality, we applied the QA60 band for cloud masking in Sentinel-2 images. Both datasets were processed using a median reducer across the entire year of 2024 to generate cloud-free, representative composites. All spectral bands and backscatter coefficients were then normalized to a range of [0, 1] to mitigate effects from illumination conditions and sensor characteristics. Finally, these processed layers were stacked into a unified feature collection to serve as input for the machine learning models." (lines 198-209)