Prospecting Potential Sites for Concentrating Solar Power (CSP) Plants through Embedding Similarity of the AlphaEarth Foundations Model

Eduardo Chiarani[1,2], **Marlon Marcon**[1]

[1]Universidade Tecnológica Federal do Paraná (UTFPR)

[2]Centro Universitário SENAI/SC (UniSENAI)

eduardo.chiarani@edu.sc.senai.br, marlonmarcon@utfpr.edu.br

Abstract. *Prospecting geographic sites for Concentrating Solar Power (CSP) plants is carried out by weighting and processing several geospatial layers. This work proposes an alternative, in which sites are represented by the embeddings of the AlphaEarth Foundations model. To that end, 26 reference plants are compared among themselves by cosine similarity and the Wasserstein distance. Over those measures, a feature selection based only on positive examples, Sequential Forward Floating Selection (SFFS) with Leave-One-Plant-Out (LOPO) validation, reduces the 64 dimensions to a four-band signature. With the signature fixed, a site is scored by the same measures against the 26, under a top-k criterion. The efficacy of the method was verified with 22 operational and spatially independent plants held out of the reference set, of which 21 were recovered ($95.5\,\%$). Another seventeen candidate areas were used as an application test, of which four came out adherent.*

1. Introduction

The energy transition is a global agenda reaffirmed at COP30, yet it falls short of its targets: tripling installed capacity by 2030 would require about 11.2 TW, against the 7.4 TW projected [IRENA 2025, COP30 2025]. In this scenario, Concentrating Solar Power (CSP) occupies a specific position. It concentrates direct solar irradiation to heat a fluid and stores heat, usually in molten salts [Fernández-García et al. 2010], which decouples generation from the instantaneous presence of the sun and adds dispatch value in systems with intermittent sources.

Site selection determines part of the viability of a CSP project and varies with geographic conditions such as direct irradiance, topography, water availability, infrastructure and environmental constraints, among others. Part of the literature treats it through multi-criteria decision analysis (MCDA) over GIS, in which each criterion becomes a weighted geospatial layer combined into a suitability index. In Brazil, Chiarani et al. [Chiarani et al. 2023] define the criteria and weight them explicitly with support from questionnaires evaluated by experts, followed by the Analytic Hierarchy Process (AHP); along the same lines, Demir et al. [Demir et al. 2023] combine AHP and GIS for large-scale photovoltaic plants, weighting irradiation, slope, land use and cost, and distance to roads and transmission lines, in a case study in İzmir, Türkiye. In both, AHP structures the weights and GIS combines the layers, which consolidates the AHP–GIS pairing as a reference approach for renewable-generation site selection. This approach, however, has

a cost. Defining the criteria, obtaining and interoperating each layer and arbitrating the weights demands long processes, domain knowledge and subjective judgement, redone for each new region.

Machine learning approaches already dispense with manual weighting in this task. Sun et al. [Sun et al. 2023] model the site choice of photovoltaic plants with random forests and gradient boosting over a national inventory of installations, learning the locational pattern of the plants already built instead of arbitrating weights. The method requires, however, an extensive inventory, and CSP technology has only a few dozen plants in operation, which makes training a classifier along those lines unfeasible.

Geospatial foundation models, in turn, offer a single representation of the territory, covering different criteria and enabling similarity analyses between sites without requiring a classification model. AlphaEarth Foundations, for example, synthesizes optical, radar, topographic, climatic and land-use observations into a 64-dimensional vector per position, at 10 m [Brown et al. 2025], over which similarity is measured directly, without reconstructing each criterion. Since only a few dozen operational CSP plants are known, this work proposes to measure similarity to those plants in the AlphaEarth space, rather than training a classifier.

The use of AlphaEarth embeddings in environmental tasks appears in recent studies. Alam and Simic Milas [Alam and Simic Milas 2025] estimate canopy biochemical variables and reduce the dimensionality of the embeddings to keep them compact, which indicates that dimension selection is a useful practice over this representation. Gao et al. [Gao et al. 2026], in turn, map tobacco with a phenology-guided temporal model. Both have abundant labels; the present study differs by operating with few positive examples.

The methodology is presented in five execution steps: (Section 2.1) extraction of the embeddings; (Section 2.2) definition of the measures for comparing sets of embeddings; (Section 2.3) reduction of the embeddings to a signature, by a positives-only selection that compares the 26 references among themselves by cosine and Wasserstein; (Section 2.4) validation of that signature, by LOPO, search restarts and exhaustive enumeration of optimality; and (Section 2.5) construction of the reference distribution against which a site is scored by the same two measures, summarized into an adherence index under a top-k criterion calibrated leave-self-out. The efficacy of the method is verified with operational plants held out of the reference set. Candidate areas are used for applying the method, without being part of the verification.

2. Materials and Methods

2.1. Extraction and projection of the embeddings

For each plant and each candidate, an arbitrary 500 m square window was clipped over the annual AlphaEarth collection in Google Earth Engine, index-year 2024. The reference and verification plants, which have built infrastructure, are sampled with the window shifted by at least 1 km from the point, to capture the surroundings and not the installation itself, whose embeddings would be atypical and specific to it; this is consistent with prospecting, since the areas one wishes to identify do not yet host plants, so the signature cannot depend on the constructions. The candidates, without installations, are sampled at the point of interest instead. Each window produced about 2,500 64-dimensional vectors

(Figure 1). Each dimension is considered a band, indexed from A00 to A63, by analogy with spectral bands, and selecting bands reduces dimensionality. Unlike spectral bands, however, the dimensions are learned and correlated, so a selected band represents a group of correlated dimensions, not an independent physical measurement.

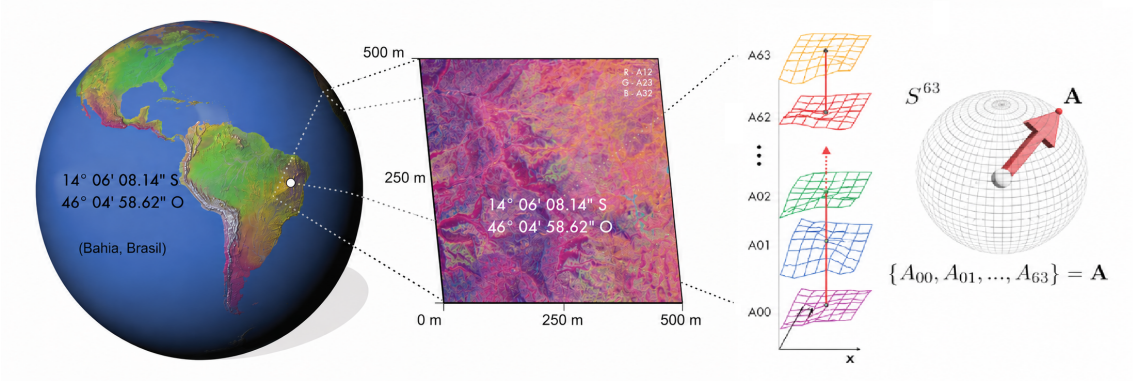


Figure 1. Extraction of regions of interest from AlphaEarth Foundations.

Source: Adapted from Brown et al. (2025).

The vectors are stored and processed in native UTM projection throughout the chain of operations.

2.2. Measures for comparing sets of embeddings

The comparison between sets of embeddings, present both in the reduction of the representation (Section 2.3) and in the subsequent evaluation of the candidates (Section 2.5), rests on two measures. Cosine similarity [Turney and Pantel 2010], a natural operation on spaces of unit vectors, between two centroids $\bar{x}$ and $\bar{y}$, given by:

$$\cos(\bar{x}, \bar{y}) = \frac{\bar{x} \cdot \bar{y}}{\|\bar{x}\| \, \|\bar{y}\|} \tag{1}$$

Where $\cos \in [-1, 1]$ measures the directional alignment between the two distributions, discarding their magnitudes.

And the Wasserstein distance [Panaretos and Zemel 2019], based on optimal transport, of order p between two distributions P_1 and P_2, defined as:

$$W_p(P_1, P_2) = \left(\inf_{\gamma \in \Gamma(P_1, P_2)} \int \|u - v\|^p \, d\gamma(u, v) \right)^{1/p} \tag{2}$$

Where $\Gamma(P_1, P_2)$ is the set of transport plans with marginals P_1 and P_2. We adopt $p = 1$ (Earth Mover's Distance), with ground cost equal to the Euclidean distance between the embedding vectors of two pixels, and not between their geographic positions, which do not enter the calculation. The transport takes place in the feature space: each pixel is a point of $\mathbb{R}^{|S|}$, the subspace of the bands in use (not the 64 original dimensions), so that cosine and Wasserstein operate at the same dimensionality. The geometry of the window is irrelevant to the measure: permuting the pixels in space does not change W_1. In the discrete case, each of the N_{pix} pixels of the window receives uniform mass $1/N_{\text{pix}}$ and

W_1 is obtained exactly by linear programming (network simplex solver), without entropic regularization or approximation. The cost of this solver grows steeply with the number of pixels, which makes the band search prohibitive, since the calculation is repeated over the $\binom{N}{2}$ reference pairs for each of the millions of subsets evaluated, motivating the pixel subsampling employed at that stage (500 per plant). Unlike the cosine, W_1 preserves the distributional structure that the centroid discards.

The cosine measures the directional alignment of the mean signal of the window, while the Wasserstein compares the shape and dispersion of the distribution, sensitive to the heterogeneity that the mean vector hides. A pair may have aligned centroids (high cosine) and distinct internal compositions (high W_1), which justifies requiring agreement on both. The cosine is computed over the mean vector (centroid) of each cloud; it compares the mean direction of the two distributions, without requiring that this mean vector be a valid embedding.

2.3. Reduction of the representation to the signature of the reference sites

Of the 64 embedding dimensions, some may be characteristics of the technology, while others may just be noise. Since only positive examples are available, with no labeled negatives to contrast against, a positives-only approach infers which is which by agreement among the 26 references: the dimensions on which they agree are taken as the signature of the reference sites, and the dimensions on which they disagree are discarded. The reduction operates in two layers.

The first layer is a univariate filter that evaluates bands in isolation and ignores interactions, serving to guide the next stage. It assigns, to each band b, a score that grows with the presence of signal and falls with the disagreement across plants (Equation 3).

$$\mathrm{consist}(b) = \frac{\mathrm{mean}_i \, |\mu_b(P_i)|}{\mathrm{sd}_i \, \mu_b(P_i) + \varepsilon} \tag{3}$$

Where $\mu_b(P_i)$ is the mean of band b in plant P_i.

The second layer is SFFS (Sequential Forward Floating Selection) [Pudil et al. 1994], a greedy wrapper that evaluates subsets as a whole. At each step, SFFS includes the band that most increases the score (Equation 4).

$$\mathrm{score}(S) = \overline{\cos}(S) \cdot \sum_{i=1}^{N} \|c_i(S)\| \cdot \frac{1}{1 + \overline{W}(S)} \tag{4}$$

Where $c_i(S)$ is the centroid of plant i restricted to the subset S (the sum runs over the N reference plants), $\overline{\cos}(S)$ is the mean cosine over all pairs of centroids and $\overline{W}(S)$ is the mean Wasserstein distance between the plants in the subspace S. SFFS maximizes $\mathrm{score}(S)$: the score grows the more the plants agree in direction ($\overline{\cos}(S) \to 1$), the more signal the centroids carry ($\sum_i \|c_i(S)\|$ high) and the less they diverge in distribution ($\overline{W}(S) \to 0$); the factor $\frac{1}{1+\overline{W}(S)}$ converts $\overline{W} \in [0, \infty)$ into a gain in $(0, 1]$, monotonic and without a scale parameter. The magnitude term $\sum_i \|c_i(S)\|$ penalizes bands with signal close to zero; its effect is isolated by ablation in Section 3.1. Being multiplicative,

the objective function vanishes when any factor fails (directional misalignment, null signal or distributional divergence), without compensation between factors and without free parameters. After each inclusion, SFFS tests the conditional removal of bands already selected, correcting early inclusions; the cardinality is not fixed a priori, and the search ends when no additional inclusion increases the score.

In the search, the term $\overline{W}(S)$ is the multivariate EMD in $\mathbb{R}^{|S|}$ over subsampled pixels, which captures the inter-band correlations. The final evaluation, whether of a candidate or a verification plant (Section 2.5), employs the exact multivariate EMD, over the same subsample of pixels.

2.4. Validation of the selection

LOPO validation measures the selection's dependence on the particular reference set. SFFS was repeated N times, removing a different plant in each round, and the frequency $f_{\mathrm{LOPO}}(b)$ with which each band was selected across the N rounds was recorded. Bands with $f_{\mathrm{LOPO}}(b) = 1.0$ are robust to the composition of the set.

A second validation measures the sensitivity to the starting point of the search, inherent to greedy algorithms. SFFS was run from distinct initial bands, including bands outside the final selection. Reconvergence to the same subset, with an identical score, indicates stability of the procedure.

2.5. Comparison of candidates and reference distribution

Evaluating a candidate area reuses Equations 1 and 2 over the selected subset of bands, now between the candidate and each of the N references, by the same two measures. From each window a random sample of 500 pixels is drawn without replacement, and the EMD is solved exactly over it. Subsampling is a cost choice. The exact EMD grows expensive quickly with the number of pixels, and using the $\sim$2,500 intact pixels of a window would make the territorial sweep prohibitive. The draw is seeded by combining a fixed seed with the area identifier, which makes it deterministic and reproducible, and uses the same 500-pixel budget as the search stage, so that selection and evaluation operate at the same sampling resolution.

By itself, the absolute value of the Wasserstein distance is not interpretable; an internal scale of the reference set is needed. A top-k criterion is adopted, with $k = 3$: the adherence index (Equation 5) is computed pairwise against each of the N references, and the k with the highest index are selected jointly; the reported statistics are the mean cosine and the mean Wasserstein distance of those same k. The candidate must thus resemble some mode of the reference space, and not the centroid of all modes, which respects the multimodality of the reference set.

These two top-k measures are then summarized into an adherence index continuous in $[0, 1]$, which serves both to rank the candidates and to classify them:

$$\mathrm{ader}(\mathrm{cand}) = \max\left(\overline{\cos}_{\text{top-}k},\ 0\right) \cdot \frac{1}{1 + \overline{W}_{\text{top-}k}} \tag{5}$$

Equation 5 repeats the same direction×distribution combination as the selection score (Equation 4), without the magnitude term, uniting the two stages of the method

under a single criterion. The magnitude does not enter the evaluation, since the bands are already fixed. Since the truncated top-k cosine is in $[0, 1]$ and $\frac{1}{1+\overline{W}_{\text{top-}k}} \in (0, 1]$, the product falls in $[0, 1]$: 1 corresponds to direction and distribution identical to the k most akin references, and negative cosines truncate to 0.

Classifying with this index is calibrated leave-self-out. For each reference, the same index is computed against the remaining $N - 1$, obtaining the distribution of the index in the set. The lower Tukey fence over that distribution ($\mathrm{Q1} - 1.5\,\mathrm{IQR}$) separates two classes: adherent and non-adherent.

To gauge the sensitivity of the method (its ability to recognize the sites of real CSP plants), all operational plants of the same technology known to the authors and not part of the reference set form the verification set: 25 plants.

Reporting on that set is stratified into three groups, since the evidential value of each is distinct:

1. Isolated (5). Geographically dispersed and with no neighboring reference. It is on them that the evidence of generalization rests, and the universe of operational parabolic-trough plants does not allow enlarging the group.
2. Regional redundancy (17). Plants from a region over-represented in the CSP universe. They neighbor no reference, but the set covers the region well, so that the positive result is expected.
3. Adjacent (3). Located a few kilometers from a reference, they sample surroundings contiguous to those that calibrate the ruler.

Of the three groups, the first two are spatially independent of the references and together make up the 22 plants over which recovery is reported; the 3 adjacent ones are excluded from that count, as their recovery is expected by construction.

Being a verification of positives, the only defined metric is recovery. Precision and its derivatives would require declared negatives, and the technique does not establish unsuitability.

3. Experiments and Results

The results described below use the reference set of 26 operational CSP plants ($N = 26$), all of the same technology (parabolic-trough collectors) and distributed across eleven countries (Table 1). An example extraction of reference is shown in Figure 2. The set was assembled in a geographically balanced way, so that no country holds a majority (six plants in Spain, six in the United States, three in South Africa, and the remaining regions with one or two each). Co-located clusters were aggregated and neighboring near-duplicates avoided. Seventeen candidate areas, randomly distributed within strategically selected regions, were evaluated: twelve in Brazil, three in the Atacama (Chile), one neighboring a cluster of plants in South Africa and one on the Bolivian Altiplano. All metrics were computed over the selected subset of bands, so that every comparison occurs at the same dimensionality.

3.1. Selected signature and its stability

SFFS converged to four bands, identified, in order of inclusion, as A12, A40, A54 and A32. LOPO validation agreed with all of them. A12, A40 and A54 were selected in

Table 1. The 26 reference plants ($N = 26$), all parabolic-trough. The names identify the sampled site; the coordinates are in the repository.

Plant	Country	Plant	Country
Shams 1	United Arab Emirates	Shagaya CSP	Kuwait
CGN Delingha (50MW Trough)	China	ISCC Ain Beni Mathar	Morocco
Royal Tech Yumen	China	Noor II	Morocco
ISCC Hassi R'mel	Algeria	Waad Al Shamal ISCC	Saudi Arabia
ISCC Kuraymat	Egypt	Genesis	United States
Andasol (1+2+3)	Spain	Martin Next Generation	United States
Enerstar	Spain	Mojave Solar	United States
Manchasol (1+2)	Spain	Nevada Solar One	United States
Solaben (1+2+3+6)	Spain	SEGS (III+IV+V+VI+VII)	United States
Solnova (1+3+4)	Spain	SEGS (VIII+IX)	United States
Valle (1+2)	Spain	Bokpoort	South Africa
Godawari Solar Project	India	Ilanga I	South Africa
Megha Solar Plant	India	Xina Solar One	South Africa



Figure 2. Example of window extracted around a CSP plant.

$26/26$ rounds ($f_{\mathrm{LOPO}} = 1.0$) and A32 in $24/26$ ($f_{\mathrm{LOPO}} = 0.92$). The next most frequent band, A33, appeared in only $1/26$ rounds ($f_{\mathrm{LOPO}} = 0.04$), and the remaining 59 bands were not selected in any round. The separation between the four bands of the signature and the rest appears in Figure 3, and all four are robust to the composition of the set.

The optimality of the selection was verified by exhaustive enumeration. All $\binom{64}{3}$ triples, all $\binom{64}{4}$ quadruples and all $\binom{64}{5}$ quintuples were evaluated. No subset exceeded the score of the signature $\{\mathrm{A12}, \mathrm{A32}, \mathrm{A40}, \mathrm{A54}\}$. For $k \leq 5$, the signature is the global optimum per cardinality according to Equation 4.

With SFFS restarted from fifty distinct starting points, including from bands outside the selection, all fifty paths reconverged to the same subset and to the same score, confirming the stability of the search procedure.

The role of the magnitude term of Equation 4 was isolated by ablation. Without it, SFFS collapses to a single band (A54) and the exhaustive enumeration finds 202 triples that exceed the signature, since bands with signal close to zero come to be taken as consistent. The magnitude is thus the factor that sustains the signature, and not a normalization detail.

The mean absolute pairwise correlation among the 64 bands, over the pixels of the references, was 0.24, and ten principal components explained about 86% of the variance, a structure compatible with the reduced cardinality of the signature, each band representing a group of correlated dimensions. Among the four bands chosen, the mean absolute correlation is low ($|r| = 0.20$), with the most correlated pair at $|r| = 0.30$, which indicates little redundancy in the selection.

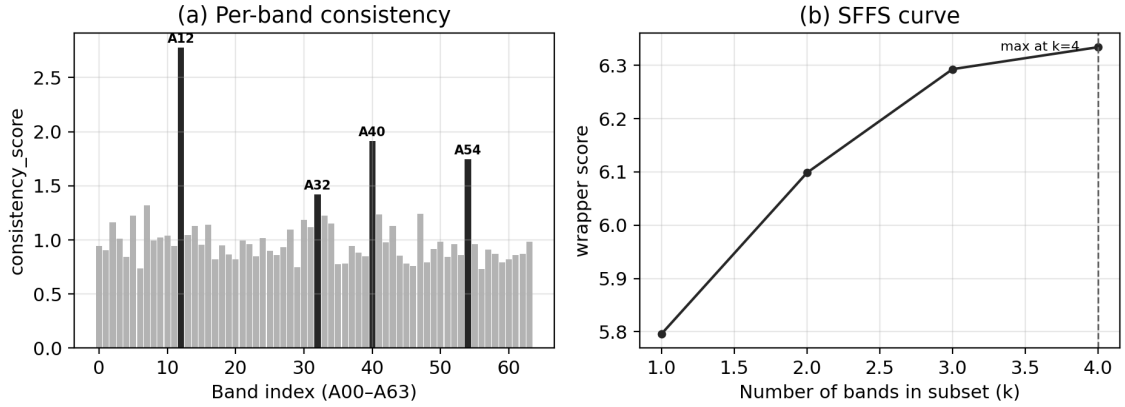


Figure 3. On the left, the consistency score (Equation 3) of each of the 64 bands, with the four selected bands highlighted. On the right, the subset score (Equation 4) along the SFFS search.

3.2. Internal coherence of the reference set

The internal distribution of the set, over the $\binom{26}{2} = 325$ reference pairs, has a median cosine of 0.875 and a median Wasserstein distance of 0.233, with a maximum Wasserstein of 0.432. The median cosine of 0.88 indicates an internally cohesive but not homogeneous set, compatible with the geographic diversity of the 26 plants and with the regional structure discussed in Section 3.3. No reference was flagged as atypical by the Tukey criterion over the inter-reference centrality. Calibrating the classification leave-self-out, the adherence index has a median of ≈ 0.89, and the lower Tukey fence ($\mathrm{Q1} - 1.5\,\mathrm{IQR}$) lies at 0.70, the threshold that separates adherent from non-adherent in the following sections.

Table 2. Recovery of the five isolated verification plants, which carry the evidence of generalization, in the selected subset of bands, ordered by decreasing adherence index (adher. and class defined in Section 2.5).

Plant	adher.	cos. top-k	Wass. top-k	Class
Solana, US	0.92	0.993	0.084	Adherent
Ashalim, IL	0.89	0.986	0.103	Adherent
Duba 1 ISCC, SA	0.89	0.976	0.103	Adherent
Gansu Akesai, CN	0.73	0.857	0.167	Adherent
Zabuye, CN	0.49	0.621	0.270	Non-adherent

3.3. Recovery of the verification plants and regional structure

Of the 22 verification plants spatially independent of the references, 21 are recovered ($95.5\,\%$): the 17 of regional redundancy and four of the five isolated.

Of those, the five isolated plants (Table 2) carry the evidence of generalization: four are adherent, from distinct regions and with no neighboring reference, which indicates that the method recognizes the sites of real, unseen CSP plants, and does not merely reproduce its own set. The single failure, Zabuye, lies below the threshold.

3.4. Adherence of the candidates

With the method verified, it is applied to seventeen candidates, as the prospecting exercise over areas whose suitability is unknown, and not as a gauge of the efficacy of the method. The adherence of each, measured by the index and classified by the Tukey range defined in Section 2.5 (Equation 5), is in Table 3 and in Figure 4.

Four candidates are adherent (index ≥ 0.70, Table 3): ZA-1, neighboring Upington, has the expected adherence of sampling the same surroundings as the references, while BO-1, BA-1 and BA-2, with no plants around them, lie just above the threshold.

In western Bahia, BA-1 and BA-7 sit a few kilometers apart yet fall at opposite ends of the index: the former is adherent, the latter anti-aligned (mean cosine -0.424); BA-6 and SC-3 reach cosine -0.40 and -0.51. The index records the relative position of each window in the AlphaEarth latent space, and relating that position to attributes of the terrain would require a covariate analysis, not carried out in this work. The three Atacama samples point the same way: CL-1, CL-2 and CL-3 resulted in weak to null adherence (mean cosine 0.410, 0.120 and -0.299), so the evaluation falls on the sampled window, and not on the region containing it. Of the remaining non-adherent candidates, five are anti-aligned, zeroing the index.

Table 3. Adherence of the seventeen candidates in the band subset $\{\mathrm{A12}, \mathrm{A32}, \mathrm{A40}, \mathrm{A54}\}$. The MCDA class column gives the suitability of the multi-criteria decision baseline [Chiarani et al. 2023].

ID	Description	adher.	Class	MCDA class
ZA-1	neighborhood of CSP plants	0.88	Adherent	n.a.
BO-1	arid plateau	0.81	Adherent	n.a.
BA-1	bare soil (west)	0.78	Adherent	0.57 (viable)
BA-2	cerrado (west)	0.74	Adherent	0.52 (viable)
CL-1	high-altitude glaciers	0.64	Non-adherent	n.a.
SC-1	coastal dune	0.59	Non-adherent	0 (not viable)
CL-2	high-altitude desert	0.46	Non-adherent	n.a.
BA-3	arid backlands (north)	0.42	Non-adherent	0.47 (low viab.)
BA-4	semi-arid backlands (north)	0.36	Non-adherent	0.47 (low viab.)
PE-1	semi-arid backlands	0.29	Non-adherent	0.41 (low viab.)
BA-5	water body (S. Francisco)	0.18	Non-adherent	0.21 (low viab.)
CL-3	Andean mountains	0.13	Non-adherent	n.a.
PR-1	urban and vegetated	0.00	Non-adherent	0 (not viable)
SC-2	coastal lagoon	0.00	Non-adherent	0 (not viable)
BA-6	dense caatinga (north)	0.00	Non-adherent	0.43 (low viab.)
BA-7	dense vegetation (west)	0.00	Non-adherent	0.45 (low viab.)
SC-3	Atlantic forest	0.00	Non-adherent	0 (not viable)

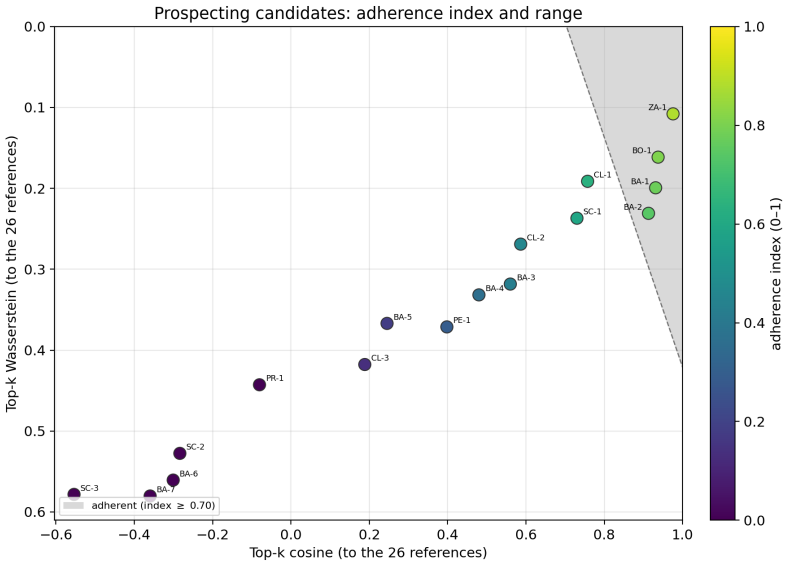


Figure 4. Top-k cosine against top-k Wasserstein distance of the seventeen prospecting candidates

4. Discussion

The results indicate that the convergence between high cosine and low Wasserstein separates the candidates coherently, without manual definition of criteria.

The gain of the top-k criterion over the mean (Section 2.5) appears in the recovery. Under the same threshold, the mean rule recovers 7 of the 22 spatially independent verification plants ($31.8\,\%$) against 21 by the top-k ($95.5\,\%$); 16 of the 22 change class between the two rules, all adherent by the top-k and non-adherent by the mean.

Of the five isolated plants, four are recovered. The site of the Gansu Akesai plant passes by a narrow margin (0.73 against the threshold at 0.70), and Zabuye, in turn, is rejected by a wide margin. It remains an open question whether this happens because their windows occupy a region of the latent space thinly populated by the references, making even their most akin matches weak.

The MCDA baseline on Table 3 provides a comparison, not a validation, since agreement between methods does not by itself confirm either against ground truth. The MCDA class column confronts the similarity screening of the twelve Brazilian candidates with the multi-criteria decision baseline [Chiarani et al. 2023]; the two methods share no inputs, since the MCDA weights explicit layers (irradiance, slope, water, infrastructure) under expert weights, while the similarity operates only over embeddings and positive examples. The two candidates adherent by the similarity are exactly the two rated viable by the MCDA, and the remaining ten agree as non-adherent ($12/12$). That the rules agree despite sharing no inputs indicates that there may be an ordering in the latent space with respect to the signature of the technology, even though these results do not say what information the selected dimensions carry.

5. Conclusion

The work proposed a methodology for prospecting potential CSP plant sites through embedding similarity of AlphaEarth, without manual definition of criteria and from positive

examples only. The representation was reduced to a four-band signature (global optimum per cardinality in the exhaustive search, stable across LOPO rounds and fifty restarts). The efficacy was verified over operational plants held out of the reference set: of the 22 spatially independent ones, 21 were recovered ($95.5\,\%$), four of them among the five isolated, which carry the evidence of generalization. Applied next to seventeen candidate areas, the method flagged four as adherent, in a screening that fully agrees ($12/12$) with the MCDA baseline [Chiarani et al. 2023] over the Brazilian ones.

The method establishes similarity in the AlphaEarth latent space, a necessary but not sufficient condition for CSP deployment. What the index measures, throughout, is the proximity between the candidate window and the windows of the reference plants in that space; the space is learned and its dimensions have no declared meaning, so the results do not identify what information sustains the proximity nor decompose it into factors. Its function is the prospecting and screening of areas of interest, to be complemented by technical, economic and social criteria. These are promising results, but broader testing with more data is needed to confirm the method's prospecting power.

From these limitations follow the next steps: (i) sweeping entire regions on a grid, producing large-scale adherence maps; (ii) reproducing the tests cited here with greater computational power and without subsampling; and (iii) a hybrid methodology that joins the similarity to the factors that AlphaEarth does not observe (transmission lines, access roads, water availability, land-tenure regime, economic indices and legal constraints).

As for reproducibility, the scripts that reproduce the results, with the identifiers of the 26 reference plants, the 25 verification plants and the 17 candidate areas (coordinates and attributes), are available in `https://github.com/eduardochiarani/pps-ae`.

References

Alam, M. M. T. and Simic Milas, A. (2025). Dimensionality optimized machine learning retrieval of canopy chlorophyll, nitrogen, and phosphorus from Google satellite embeddings.

Brown, C. F. et al. (2025). AlphaEarth Foundations: an embedding field model for accurate and efficient global mapping from sparse label data. arXiv. Disponível em: `https://arxiv.org/pdf/2507.22291`.

Chiarani, E., Antunes, A. F. B., Drago, D., Oening, A. P., and Paschoalotto, L. A. C. (2023). Optimal site selection using GIS-based multi-criteria decision analysis (MCDA): a case study to concentrated solar power plants (CSP) in Brazil. *Anuário do Instituto de Geociências*, 46.

COP30 (2025). COP30 approves Belém Package 1. Disponível em: `https://cop30.br/en/news-about-cop30/cop30-approves-belem-package1`.

Demir, A., Dinçer, A. E., and Yılmaz, K. (2023). A novel method for the site selection of large-scale PV farms by using AHP and GIS: a case study in izmir, Türkiye. *Solar Energy*, 259:235–245.

Fernández-García, A., Zarza, E., Valenzuela, L., and Pérez, M. (2010). Parabolic-trough solar collectors and their applications. *Renewable and Sustainable Energy Reviews*, 14(7):1695–1721.

Gao, H. et al. (2026). TSP-Former: a phenology-guided transformer for tobacco mapping using satellite image time series. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 19:2423–2438.

IRENA (2025). Tracking COP28 outcomes: tripling renewable power capacity by 2030. Technical report, International Renewable Energy Agency, Abu Dhabi.

Panaretos, V. M. and Zemel, Y. (2019). Statistical aspects of Wasserstein distances. *Annual Review of Statistics and Its Application*, 6:405–431.

Pudil, P., Novovičová, J., and Kittler, J. (1994). Floating search methods in feature selection. *Pattern Recognition Letters*, 15(11):1119–1125.

Sun, Y., Zhu, D., Li, Y., Wang, R., and Ma, R. (2023). Spatial modelling the location choice of large-scale solar photovoltaic power plants: application of interpretable machine learning techniques and the national inventory. *Energy Conversion and Management*, 289:117198.

Turney, P. D. and Pantel, P. (2010). From frequency to meaning: vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.