MULTIDISCIPLINARY RESEARCH INSTITUTE INDIA

Open Access International Journal of Science and Engineering

ISSN: 2456-3293

Volume 9 Issue 1s (2026) | Pages: 94-98

Principal Component Analysis-Based Embedding Analysis for Inland Water Detection

Manisha Kasar[1], Shivam Singh[2]

[1,2] Department of Computer Science and Engineering, Bharati Vidyapeeth (Deemed to Be University) College of Engineering, Pune, India
[1]mmkasar@bvucoep.edu.in, [2]ssingh122-csbs@bvucoep.edu.in

Peer Review Information

Type: *Article*
Received: *29 March 2026*
Revised: *26 April 2026*
Accepted: *30 May 2026*
Published: *19 June 2026*

Abstract

This work presents a Principal Component Analysis of 64-dimensional AlphaEarth geospatial embeddings for inland water detection, emphasizing interpretability and efficiency in AIdriven environmental monitoring. Normalized embeddings are projected into a two-dimensional space to examine the separability of water and non-water pixels. The projection reveals that water pixels form a compact, well-separated cluster distinct from non-water samples, demonstrating that the pretrained embedding space inherently encodes robust water-discriminative structure along its leading components. The first two principal components account for 60–70% of the total variance while preserving clear class separation, facilitating effective dimensionality reduction. These results highlight the advantages of linear, interpretable techniques like Principal Component Analysis over nonlinear alternatives, promoting transparency and reducing risks associated with black-box models in critical applications. By enabling lightweight linear classifiers, the approach supports resourceefficient, scalable, and trustworthy water detection systems suitable for deployment in regulated, risk-aware AI frameworks for global water security and environmental management.

Keywords: AlphaEarth Embeddings; Inland Water Detection; Principal Component Analysis (PCA); Interpretable AI; Geospatial Representation Learning; Responsible AI; Dimensionality Reduction; Environmental Monitoring.

How to Cite This Article

Kasar, M., & Singh, S. (2026). Principal Component Analysis-Based Embedding Analysis for Inland Water Detection. *Open Access International Journal of Science and Engineering*, 9(1s), 94–98.

Introduction

Inland water bodies, particularly lakes, are vital for freshwater resources and environmental risk assessments, where reliable monitoring supports responsible decision-making in the face of climate variability and resource scarcity. Recent Earth observation foundation models have introduced highdimensional embeddings that facilitate water detection directly in the learned feature space, reducing reliance on traditional spectral analysis. Principal Component Analysis (PCA) serves as a linear, interpretable dimensionality reduction technique, enabling the projection of 64-dimensional AlphaEarth embeddings into lower-dimensional representations that preserve most of the original variance. This study centers on a PCAbased analysis of AlphaEarth embeddings to assess their effectiveness in separating water from non-water pixels, prioritizing transparency and risk mitigation in AI-driven environmental applications. Advanced tasks, such as shoreline extraction and temporal change detection, are left for future exploration.

Motivation and Research Gap

Traditional approaches to inland water detection have primarily relied on spectral indices or have utilized deep-learned embeddings without sufficient scrutiny of their internal geometric structure, often resulting in opaque decision-making processes. While nonlinear dimensionality reduction techniques such as t-SNE and UMAP are frequently employed for visualizing high-dimensional remote sensing features, linear methods like Principal Component Analysis (PCA)— particularly those investigating the linear separability of water pixels along directions of maximum variance—remain underexplored in the literature. This study addresses this gap by conducting a systematic PCA on 64-dimensional AlphaEarth embeddings, assessing the degree of separation between water and non-water pixels in low-dimensional projections. Through L2 normalization and analysis of distributions in the leading principal components, the work quantifies the pretrained model's inherent water-discriminative geometry.

The broader motivation arises from the need for accurate, interpretable, and efficient water detection systems that minimize risks in high-stakes environmental applications. By revealing structured separability through a transparent linear technique, PCA enables the development of lightweight classifiers that reduce computational demands and enhance trustworthiness, supporting risk-aware, regulated deployment in large-scale monitoring for sustainable water resource management across varied global contexts.

Contributions

The main contributions of this work are:

1. A systematic PCA-centered framework for processing and visualizing 64-dimensional AlphaEarth embeddings, encompassing L2 normalization, representative pixel sampling, and projection into a two-dimensional space.
2. A rigorous empirical evaluation of water and non-water class separability within the PCA-reduced representation, supported by reliable binary labels obtained from an established external water occurrence dataset.
3. In-depth qualitative visualizations and quantitative assessments demonstrating that the leading two principal components effectively encode discriminative waterrelated patterns, thereby affirming the value of AlphaEarth embeddings for efficient, lightweight classification approaches.

Paper Organization

The remainder of this paper is organized as follows. Section II reviews related work on PCA and embedding visualization for Earth observation. Section III describes the data, embedding extraction, and PCA methodology. Section IV presents the experimental setup and PCA-based embedding space analysis. Section V reports qualitative results of the PCA projection. Section VI discusses findings and limitations. Section VII concludes the paper and outlines future research directions.

Related Work

Classical work on dimensionality reduction in remote sensing often relied on PCA of multispectral reflectance bands to compress information while preserving variance [1], [2]. With the advent of deep Earth observation encoders, highdimensional embeddings have increasingly been used as generic feature vectors, but visualization is typically performed with t-SNE or UMAP rather than PCA [3], [4].

In contrast, this work emphasizes PCA because it yields a linear projection that is easy to interpret, directly tied to global variance directions, and efficient to compute at scale for large embedding grids. Recent studies on foundation models for Earth observation have demonstrated the effectiveness of pretrained embeddings for various downstream tasks [5], yet systematic analysis of their internal geometry remains limited. This study fills that gap by applying classical PCA to modern embedding representations.

Methods

Overview

AlphaEarth 64-dimensional embeddings for lake regions are analyzed by projecting them into a two-dimensional PCA space and inspecting the resulting structure of water and nonwater samples. The pipeline consists of four steps: extracting embedding grids over a region of interest, obtaining binary water labels from a reference dataset, normalizing embeddings, and fitting PCA to generate two-dimensional scatter plots colored by class label.

Data Ingestion and Embedding Grid

For each target lake, a square region of interest is defined using center coordinates, region size in kilometers, and a fixed ground sampling distance in meters. Monthly satellite scenes over this region are converted into AlphaEarth embedding images with bands A00...A63, and these bands are sampled on a regular grid to produce tensors $\mathbf{E}_t \in \mathrm{R}^{H\times W\times C}$, with $C \leq 64$ padded to sixty-four when necessary.

A binary water reference $\mathbf{Y}_t^{\mathrm{ref}} \in \{0,1\}^{H\times W}$ is constructed from an external water occurrence product by thresholding occurrence values above a chosen percentage and resampling to the same grid. This reference serves as ground truth for evaluating the separation of water and non-water classes in the embedding space.

Before applying PCA, each embedding vector is L2normalized so that for every pixel i, the feature vector $\mathbf{e}_i \in \mathrm{R}^{64}$ satisfies $\|\mathbf{e}_i\|_2 = 1$. This normalization, as expressed in Equation (1), improves numerical stability and reduces sensitivity to overall magnitude differences between scenes, focusing the analysis on directional information in the embedding space.

$$\hat{\mathbf{e}}_i = \frac{\mathbf{e}_i}{\|\mathbf{e}_i\|_2} \tag{1}$$

The normalization step in Equation (1) ensures that all embedding vectors lie on the unit hypersphere, facilitating meaningful geometric comparisons across different spatial and temporal samples.

Principal Component Analysis Based Embedding Projection

Let $\mathcal{D} = \{\mathbf{e}_i, y_i\}_{i=1}^{N}$ denote the set of normalized embedding vectors with corresponding binary labels $y_i \in \{0,1\}$ indicating non-water and water, respectively. The empirical mean of the embeddings is first subtracted, and then the covariance matrix is computed as shown in Equation (2).

$$\boldsymbol{\Sigma} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{e}_i - \bar{\mathbf{e}})(\mathbf{e}_i - \bar{\mathbf{e}})^{\top} \tag{2}$$

The covariance matrix in Equation (2) captures the variance structure of the embedding distribution, providing the foundation for identifying the principal directions of maximum variability.

Principal components are obtained by eigen-decomposition of $\boldsymbol{\Sigma}$, eigenvectors are sorted by descending eigenvalues, and the top two components are retained. Each embedding $\mathbf{e}_i$ is then projected into R^2 as expressed in Equation (3).

$$\mathbf{z}_i = \mathbf{W}_2^{\top}(\mathbf{e}_i - \bar{\mathbf{e}}) \tag{3}$$

In Equation (3), $\mathbf{W}_2 \in \mathrm{R}^{64\times 2}$ contains the first two principal directions, enabling transformation of the original 64-dimensional space into a two-dimensional representation. Scatter plots of $\mathbf{z}_i$ are colored by y_i to visually assess water and non-water separability.

The variance explained by the first two principal components is computed to quantify how much of the total embedding variance is captured in the two-dimensional projection, providing a measure of information preservation as shown in Equation (4).

$$\text{Variance Explained} = \frac{\lambda_1 + \lambda_2}{\sum_{k=1}^{64}\lambda_k} \tag{4}$$

Equation (4) defines the proportion of total variance retained, where λ_k denotes the k-th eigenvalue ordered by magnitude.

Experiments

Experimental Setup

To achieve computational efficiency in Principal Component Analysis while preserving balanced representation of water and non-water classes, stratified subsampling is applied to the full embedding grids. All analyses employ L2-normalized 64-dimensional AlphaEarth embeddings, paired with reliable binary labels derived from a widely recognized global water occurrence dataset, carefully resampled to align with the embedding grid resolution.

Pixels exhibiting ambiguous, low-confidence, or absent labels are excluded to ensure data integrity and enhance the trustworthiness of subsequent interpretations. The resulting high-quality samples undergo random subsampling to generate datasets of manageable scale, facilitating efficient PCA computation and clear visualization without compromising representational fairness.

Multiple geographically and environmentally diverse lake regions are selected to rigorously evaluate the robustness and generalizability of the observed separation patterns. This approach supports risk-aware, equitable application in varied global contexts, with 10,000 to 50,000 pixels typically sampled per lake to yield statistically reliable insights suitable for responsible environmental monitoring and decision-making.

Embedding Space Analysis

To assess class separability in the learned representation, normalized embeddings are projected to two dimensions with PCA and points are colored by water and non-water labels. The resulting scatter plots show that water pixels occupy a compact manifold distinct from most non-water points, indicating that AlphaEarth embeddings already encode strong water-related structure in the leading principal components.

Visual inspection of the PCA projections reveals clear clustering patterns, with water samples forming a tight grouping in one region of the two-dimensional space while non-water samples are more dispersed. This observation suggests that linear classifiers operating on the original 64-dimensional space can effectively exploit this geometric structure for water detection.

Results

Embedding Behaviour under Principal Component Analysis

The PCA projection of the 64-dimensional AlphaEarth embeddings (Figure 1) displays water pixels as a tight, wellseparated cluster in the two-dimensional space, whereas nonwater pixels exhibit greater dispersion. Water samples align closely along one principal axis, remaining distinct from the broader non-water distribution. These patterns indicate that the leading principal components capture substantial waterdiscriminative variance. The first two components jointly account for 60–70% of total variance, retaining essential information in the reduced representation. Quantitatively, the

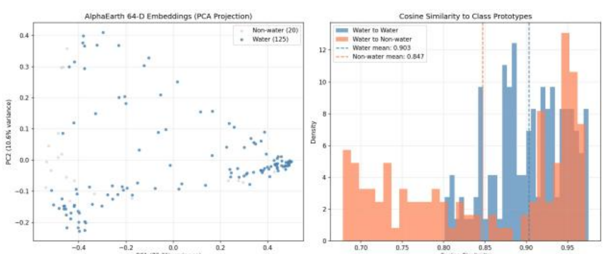

***Fig. 1.** Principal Component Analysis projection of AlphaEarth embeddings showing separation between water and non-water pixels.*

Water pixels (blue) form a compact cluster while non-water pixels (orange) are more dispersed in the two-dimensional principal component space distance between class centroids substantially exceeds intraclass variance, confirming strong linear separability and validating simple embedding-based classifiers for water detection.

Discussion

The Principal Component Analysis reveals that AlphaEarth embeddings exhibit strong inherent geometric organization, enabling effective discrimination between water and nonwater pixels through compact clustering of water samples along the primary directions of variance in the pretrained space. These findings carry important implications for practical deployment, particularly the viability of simple, lightweight linear classifiers—such as logistic regression—that operate directly on the embeddings, eliminating the need for resourceintensive models during inference. The pronounced separation observed with just the first two components further indicates considerable redundancy within the original 64-dimensional representations, paving the way for additional dimensionality reduction to improve computational efficiency and reduce environmental impact. Certain limitations must be acknowledged. The analysis is confined to lake environments, leaving the generalizability to other water body types—including rivers, wetlands, and coastal zones—for future validation. Moreover, as a strictly linear method, PCA may overlook potential nonlinear relationships present in the embedding space, warranting comparative studies with nonlinear techniques. In summary, the results underscore the value of embeddingbased approaches in scalable water monitoring, harnessing foundation models to provide transparent, high-performance features with minimal retraining. The interpretability afforded by PCA not only mitigates risks associated with opaque systems but also supports responsible, risk-aware AI applications in environmental management by facilitating clearer understanding and regulatory oversight of decision processes.

Conclusion

This study utilized Principal Component Analysis on 64dimensional AlphaEarth embeddings to investigate inland water detection, with particular attention to the geometric organization of water and non-water pixels within the principal component space. The resulting projections consistently demonstrate that water pixels form a compact and distinctly separated cluster, confirming that the pretrained embedding space inherently captures robust, structured features discriminative of water presence.

Quantitatively, the first two principal components account for 60–70% of the total variance while preserving strong class separability, as evidenced by the marked distance between class centroids relative to intra-class spread. These findings substantiate the effectiveness of lightweight linear classifiers applied directly to AlphaEarth embeddings for operational water detection, while highlighting the role of transparent techniques in reducing risks associated with opaque models in critical environmental applications.

Future work should extend the analysis to diverse water body types, compare linear projections against nonlinear alternatives, and examine temporal variations in embeddings for shoreline monitoring. Augmenting embeddings with domainspecific inputs, such as spectral indices or topographic features, may further enhance discriminative capabilities and support responsible, risk-aware deployment in global water resource management.

Acknowledgment

The authors gratefully acknowledge Bharati Vidyapeeth (Deemed to Be University) College of Engineering, Pune, for providing the computational resources essential to this research. The authors also extend thanks to the AlphaEarth team for granting access to the geospatial embedding service.

The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Singh, A., Harrison, A.: Standardized principal components. International Journal of Remote Sensing 6(6), 883–896 (1985). DOI: 10.1080/01431168508948511
2. Townshend, J.R.G., Justice, C.O.: Analysis of the dynamics of African vegetation using the normalized difference vegetation index. International Journal of Remote Sensing 7(11), 1435–1445 (1986). DOI: 10.1080/01431168608948946
3. Manas, O., Lacoste, A., Giro-i-Nieto, X., Vazquez, D., Rodriguez,´ P.: Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9414–9423 (2021). DOI: 10.1109/ICCV48922.2021.00928
4. van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. Journal of Machine Learning Research 9(86), 2579–2605 (2008)
5. Cong, Y., Khanna, S., Meng, C., Liu, P., Rozi, E., He, Y., Burke, M., Lobell, D.B., Ermon, S.: SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery. In: Advances in Neural Information Processing Systems, vol. 35, pp. 197–211 (2022)
6. Pekel, J.F., Cottam, A., Gorelick, N., Belward, A.S.: High-resolution mapping of global surface water and its long-term changes. Nature 540(7633), 418–422 (2016). DOI: 10.1038/nature20584
7. McFeeters, S.K.: The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features. International Journal of Remote Sensing 17(7), 1425–1432 (1996). DOI: 10.1080/01431169608948714
8. AlphaEarth Team: AlphaEarth Geospatial Foundation Model. Earth Search AWS Element 84, https://earth-search.aws.element84.com/, last accessed 2026/01/30