scientific reports

Check for updates

OPEN

Optimized spectral indices for global vegetation and water mapping using Sentinel-2

Charalambos Chrysostomou✉, Stelios P. Neophytides, Michalis Mavrovouniotis & Diofantos G. Hadjimitsis

Reliable mapping of vegetation and surface water from satellite imagery remains challenging, as common spectral indices can saturate at high biomass, show limited sensitivity across ecosystems, and confuse targets with soil, shadows, or built-up surfaces. We present two indices, the Symbolic Regression Vegetation Index (SRVI) and the Symbolic Regression Water Index (SRWI), discovered with a data-driven symbolic regression framework applied to Sentinel-2 Level-2A reflectance and guided by ESA WorldCover labels. Expressions were evolved from physically interpretable building blocks using non-linear combinations of visible, NIR, and SWIR bands. Indices were derived on a spectrally complex Mediterranean site and evaluated on eleven independent regions spanning diverse biomes. Performance was assessed with the Jeffries–Matusita distance, averaged across months to account for phenology, and compared against established vegetation indices (NDVI, EVI, SAVI, MSAVI2, NDRE) and water indices (NDWI, MNDWI, AWEI, TCW, WI2015). SRVI improves separability between vegetation and non-vegetation and shows higher discrimination among vegetation types relative to all benchmarks. SRWI yields more consistent water delineation with reduced confusion with built-up and shadowed surfaces, outperforming standard alternatives on the same datasets. Results indicate that symbolic regression can produce compact, interpretable indices that generalise across regions and seasons, offering practical gains for global vegetation and water mapping.

The continuous and accurate mapping of terrestrial and aquatic ecosystems is fundamental to modern environmental science, providing the empirical basis for understanding climate change impacts, managing natural resources, and ensuring global food and water security[1,2]. Satellite-based remote sensing offers an unparalleled capacity for this task, delivering consistent observations of the Earth's surface across vast spatial and temporal scales unattainable through ground-based methods alone[3]. A primary method for extracting thematic information from multispectral satellite data is the use of spectral indices. These are mathematical combinations of reflectance values from different spectral bands that enhance the signal of a specific surface feature, such as healthy vegetation or open water, while suppressing noise from confounding factors such as soil background and atmospheric effects[4].

For decades, vegetation mapping has been dominated by the Normalized Difference Vegetation Index (NDVI), a simple yet effective ratio leveraging the unique spectral signature of chlorophyll[5]. However, the limitations of NDVI are well-documented as it is known to lose sensitivity and saturate over dense, multi-layered canopies and is highly susceptible to soil brightness in sparsely vegetated regions, which can confound the vegetation signal[6]. These drawbacks spurred the development of more sophisticated indices. The Soil-Adjusted Vegetation Index (SAVI) and its successor, the Modified Soil-Adjusted Vegetation Index (MSAVI2), were introduced to explicitly minimise soil-induced variations[6]. Concurrently, the Enhanced Vegetation Index (EVI) was developed to correct for both soil and atmospheric influences by incorporating the blue band[7], while the Normalized Difference Red Edge Index (NDRE) utilises the red-edge portion of the spectrum to improve sensitivity to chlorophyll content in high-biomass conditions where NDVI typically fails. Despite this proliferation, a robust, general-purpose index that performs reliably across the full spectrum of global vegetation conditions remains challenging.

Similarly, the accurate delineation of surface water bodies presents its own set of challenges. The Normalized Difference Water Index (NDWI), which uses the green and near-infrared bands, was the foundational index for this task[8]. However, its susceptibility to confusion with built-up areas led to the development of the Modified Normalized Difference Water Index (MNDWI), which substituted the near-infrared band with a short-wave infrared (SWIR) band to improve separation from urban features[9]. Nevertheless, even MNDWI can struggle with features like topographic shadows and dark, impervious surfaces[10]. This spurred the creation of more

ERATOSTHENES Centre of Excellence, Limassol, Cyprus. ✉email: charalambos.chrysostomou@eratosthenes.org.cy

advanced indices like the Automated Water Extraction Index (AWEI)[10]. While other techniques such as the Tasseled Cap Wetness (TCW) transformation or the Water Index 2015 (WI2015) also exist, developing more reliable water indices remains an active area of research.

This paper addresses these gaps by replacing traditional, human-crafted index design with a data-driven methodology based on symbolic regression (SR)[11–13]. SR is a form of genetic programming that autonomously discovers mathematical expressions from data without a predefined functional form[14]. Unlike standard regression, which fits parameters to a known equation (e.g., a linear model), SR searches the space of mathematical operators (e.g., addition, division, square root) and variables to evolve the functional form itself. This approach allows for the discovery of non-intuitive relationships and has been successfully applied to discover physical laws from experimental data[14] and model complex material properties[15] in other scientific domains[16]. This capability enables the discovery of novel, potentially non-linear relations among spectral bands that are tailored to a specific objective. Using Sentinel-2 imagery and ESA WorldCover as the reference standard, we develop two indices: the Symbolic Regression Vegetation Index (SRVI) and the Symbolic Regression Water Index (SRWI).

The aim of this study is twofold. First, we develop SRVI and SRWI via SR on a single, spectrally diverse Mediterranean training site. Second, we rigorously evaluate their performance against comprehensive suites of established vegetation and water indices using independent, unseen datasets from eleven globally distributed regions. We hypothesise that indices discovered through direct optimisation on diverse spectra will yield statistically and practically significant gains under out-of-sample validation, demonstrating robust generalisation across biomes and environmental conditions.

Methods and materials

The methodological framework of this study was systematically organized into two principal phases: **(1) data-driven index development** and **(2) comprehensive global validation**. In the initial phase, the novel SRVI and SRWI indices were discovered by applying a SR algorithm to a focused training dataset. This training was performed exclusively using data from a single, carefully selected site in Limassol, Cyprus, a location chosen specifically for its complex and well-documented land cover mosaic, providing a challenging environment to derive robust and discriminating index formulations. In the subsequent phase, the performance and generalisability of these newly derived indices were rigorously assessed. This validation was conducted using an entirely separate and independent dataset comprised of eleven globally distributed sites. The selection of these diverse locations was designed to test the robustness and applicability of SRVI and SRWI across a broad spectrum of biomes and environmental conditions, ensuring that their utility extends beyond the specific characteristics of the training environment.

Study areas and dataset curation

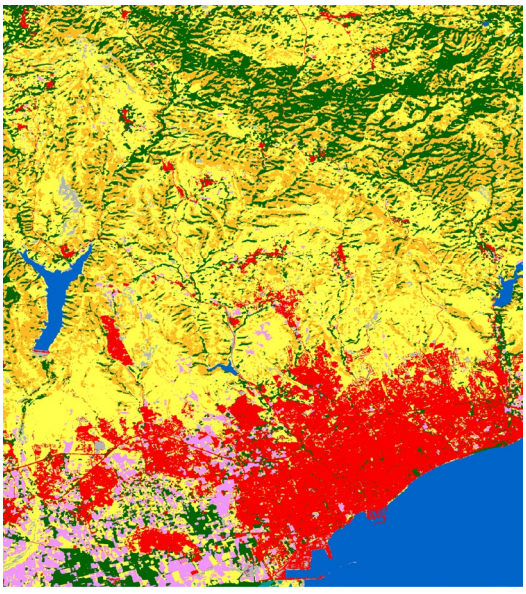
The primary satellite data were Copernicus Sentinel-2 Level-2A surface reflectance products provided by the European Space Agency (ESA)[17]. Sentinel-2 comprises twin polar-orbiting satellites with a multispectral instrument sampling 13 bands, of which 12 are available in Level-2A surface reflectance: B1 coastal aerosol (443 nm, 60 m), B2 blue (490 nm, 10 m), B3 green (560 nm, 10 m), B4 red (665 nm, 10 m), B5 red-edge 1 (705 nm, 20 m), B6 red-edge 2 (740 nm, 20 m), B7 red-edge 3 (783 nm, 20 m), B8 near-infrared (842 nm, 10 m), B8A narrow near-infrared (865 nm, 20 m), B9 water vapour (945 nm, 60 m), B11 shortwave infrared 1 (1610 nm, 20 m), and B12 shortwave infrared 2 (2190 nm, 20 m). The cirrus band B10 (1375 nm) is not included in Level-2A surface reflectance. These bands span the visible, near-infrared, and shortwave infrared regions and are well suited to mapping vegetation, soil, and water. ESA WorldCover 2021[18], a global 10 m land cover map with 11 classes derived from Sentinel-1 and Sentinel-2, served as the reference standard for training and validation. Index development and training were conducted exclusively over the Limassol district, Cyprus (406 km^2; bounding box [32.90, 33.09], [34.64, 34.85]), a heterogeneous Mediterranean site comprising dense urban fabric, irrigated cropland, natural shrublands, seasonal grasslands, bare soils, and open water, as illustrated in Fig. 1a and detailed in Fig. 1b and c. While limited to a single geographic footprint, the Limassol district was selected as a "spectral microcosm" representing a high diversity of land cover types within a compact area. The site contains a heterogeneous mosaic of dense urban fabric, industrial zones, irrigated cropland, rain-fed agriculture, natural shrublands (maquis), seasonal grasslands, bare soils (calcareous and sedimentary), and open water (sea and reservoirs). This diversity ensures that the training data captures the spectral variance required to learn robust features that generalise to other biomes, mitigating the risk of overfitting to a simpler landscape.

To evaluate generalisation on unseen data, we selected eleven additional regions that span diverse biomes and land cover compositions: Sydney (Australia), São Paulo (Brazil), Nile Delta (Egypt), Aquitaine (France), Po Valley (Italy), Atlas Mountains (Morocco), Punjab (Pakistan), Constanța (Romania), Cape Town (South Africa), Central Valley (USA), and the Mekong Delta (Vietnam). Bounding boxes and areas are listed in Table 1; across these validation regions the total mapped area is 13,488 km^2. For each study area we selected twelve single-date Sentinel-2 Level-2A acquisitions from 2021 corresponding to the least-clouded scene in each month where available. Scenes were filtered and ranked using Sentinel-2 quality information (QA60 and the Scene Classification Layer), visually screened, and cloud and cloud-shadow pixels were masked before analysis. Twenty-metre and sixty-metre bands were resampled to a common 10 m grid to match WorldCover. This procedure yielded 12 images per site without temporal compositing, resulting in 144 images overall (12 images for each of 12 study areas), with training over 406 km^2 in Limassol and independent validation over 13,488 km^2 worldwide.

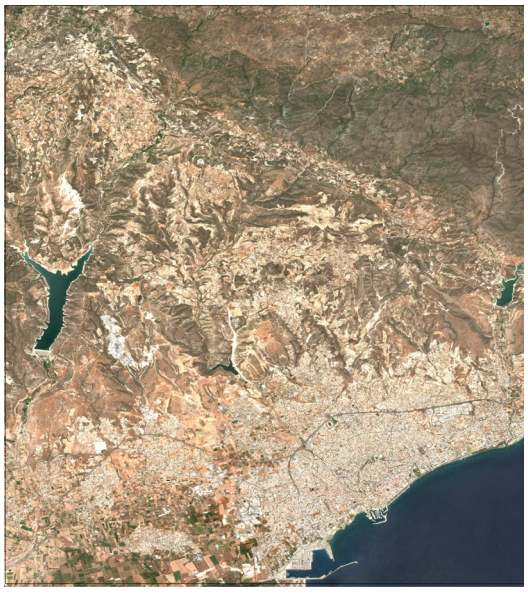
Index development using symbolic regression

We derived two compact indices with SR using only the Limassol training site (twelve least-clouded Sentinel-2 Level-2A acquisitions from 2021 per area, selected via QA60 and the Scene Classification Layer with cloud and shadow masking). Sentinel-2 surface reflectances were harmonised to a common 10 m grid to match ESA

(a) A wide-view Sentinel-2 satellite image of Cyprus with the training area outlined.



(b) ESA WorldCover classification of the training area.



(c) Detailed true-color Sentinel-2 image of the training area.

Fig. 1. The training site in the Limassol district, Cyprus. (**a**) A wide view with the training area highlighted. (**b**) The ESA WorldCover classification map for this area (e.g., red: built-up, yellow: cropland, dark green: tree cover, blue: open water). (**c**) The corresponding true-color Sentinel-2 image. All maps were generated using Google Earth Engine (https://earthengine.google.com/)[19].

WorldCover. The candidate predictor set comprised B2 (blue), B3 (green), B4 (red), B5–B7 (red-edge 1–3), B8 (NIR), B8A (narrow NIR), and B11–B12 (SWIR1–2). Bands B1 and B9 were excluded owing to resolution and water-vapour sensitivity; B10 is not provided in Level-2A reflectance. WorldCover 2021 labels were sampled on the 10 m grid for training only. To limit spatial autocorrelation and preserve class balance, we applied stratified block sampling on 1 km tiles across months and classes. Reflectances were clipped to $[0, 1]$ and per-band normalised using robust statistics (median and interquartile range) computed on the training set.

Region (CC)	Coordinates [Lon, Lat]	Climate and dominant cover	Area (km²)
Sydney (Australia)	[150.80, – 34.10]	Temperate (Cfa); Urban, Forest	2053
São Paulo (Brazil)	[– 46.80, – 23.80]	Tropical (Cwa); Urban, Dense Veg.	2266
Nile Delta (Egypt)	[31.10, 30.60]	Arid (BWh); Irrigated cropland	425
Aquitaine (France)	[– 0.50, 44.60]	Oceanic (Cfb); Agriculture, Forest	527
Po Valley (Italy)	[10.80, 45.00]	Humid Subtropical (Cfa); Cropland	349
Atlas Mtns (Morocco)	[– 6.50, 32.30]	Semi-arid (BSk); Shrubland, Bare	418
Punjab (Pakistan)	[74.00, 31.30]	Semi-arid (BSh); Intensive Crops	2108
Constanța (Romania)	[28.42, 44.08]	Humid continental (Dfa); Crops	543
Cape Town (S. Africa)	[18.20, – 34.20]	Mediterranean (Csb); Shrubland	2460
Central Valley (USA)	[– 121.06, 37.75]	Mediterranean (Csa); Agriculture	391
		Total validation area	**13,488**

Table 1. Selected regions for validation, including climatic and land cover characteristics to demonstrate biogeographical diversity.

Inputs: bands $\mathbf{x}$, targets $y \in \{-1, +1\}$, operators $\{+, -, \times, \mathrm{div}_\epsilon, \min, \max, \mathrm{abs}, \sqrt{\cdot}_\epsilon\}$, max depth 6, node cap 40, population N_p=1000, generations $G_{\max}$=800, penalties (λ_C, λ_P)
Output: expression $f^\star$ (SRVI or SRWI)
Initialise population with random trees under structural limits
for $g = 1$ to $G_{\max}$ **do**
  for each f in population **do**
    Sample a class-balanced mini-batch mixing months and 1 km blocks
    Compute $\mathrm{JM}(f)$, $\mathrm{Comp}(f)$, and $\mathrm{Phen}(f)$; set $\mathcal{F}(f) = \mathrm{JM}(f) - \lambda_C \mathrm{Comp}(f) - \lambda_P \mathrm{Phen}(f)$
  end for
  Select parents (tournament size 7); apply crossover (0.5) and mutation (0.1) with protected operators and limits
  Update best validation fitness on disjoint Limassol blocks; if no improvement for 50 generations, **break**
end for
Return $f^\star$ from the Pareto set by simplicity and fitness

Algorithm 1. Symbolic regression workflow for index discovery

Targets were defined at the pixel level and used directly in fitness evaluation (no auxiliary classifier). For the vegetation index (SRVI), we assigned $+1$ to vegetation classes (Tree cover, Shrubland, Grassland, Cropland) and -1 to non-vegetation (Built-up, Barren, Open water), with Open water explicitly -1 to enforce negative responses over water. For the water index (SRWI), Open water was labelled $+1$ and all other classes -1. Candidate expressions $f(\mathrm{x})$ were tree structures over $\mathrm{x} = \{blue, green, red, red-edge1, red-edge2, red-edge3, NIR, narrowNIR, SWIR1, SWIR2\}$ built from a deliberately conservative operator set to promote interpretability and numerical stability: binary $\{+, -, \times, \mathrm{div}_\epsilon, \min, \max\}$ with protected division div_ϵ, and unary $\{\mathrm{abs}, \sqrt{\cdot}_\epsilon\}$ with protected square root; exponential, logarithmic, trigonometric, and arbitrary power operators were excluded. Expression complexity was bounded by a maximum tree depth of 6 and a node cap of 40. Real constants were permitted as terminals and optionally refined by local least-squares on mini-batches after structural edits.

The optimisation objective prioritised statistical separability as measured by the Jeffries–Matusita (JM) distance on class-balanced mini-batches, with problem-specific definitions. For SRVI we combined the JM between vegetation and non-vegetation with the mean pairwise JM among vegetation subclasses to promote within-vegetation discrimination; for SRWI we used the JM between water and non-water. A scalar objective balanced separability against two regularisers,

$$\mathcal{F}(f) = \mathrm{JM}(f) - \lambda_C \,\mathrm{Comp}(f) - \lambda_P \,\mathrm{Phen}(f),$$

where $\mathrm{Comp}(f)$ penalises size (nodes and depth) and $\mathrm{Phen}(f)$ softly constrains monthly behaviour by comparing f's monthly means to reference phenological trends from established indices (NDVI, EVI, SAVI for vegetation; NDWI, MNDWI for water) averaged over the training site. Hyperparameters λ_C and λ_P were selected via grid search on training mini-batches to balance parsimony and separability without over-regularising seasonal dynamics.

We employed `PySR`[13] with an initial population of 1000 random trees. Each generation used tournament selection (size 7), crossover (probability 0.5), and mutation (probability 0.1) restricted to protected operators

and structural limits. After symbolic edits, numeric constants could be fine-tuned by least-squares on the current mini-batch. Mini-batches mixed pixels across months and classes to reduce temporal and class bias. The search proceeded for up to 800 generations with early stopping after 50 generations without improvement in the best validation fitness, evaluated on a hold-out of Limassol blocks disjoint from those used for structural edits and constant tuning. No data from the eleven validation regions entered model design, selection, or tuning at any point.

Model selection proceeded from the Pareto frontier defined by $(-\mathcal{F}(f), \mathrm{Comp}(f))$. From this set we chose final candidates by interpretability and expected physical behaviour–monotonic increase with NIR relative to Red for vegetation, decrease with SWIR for water, boundedness where ratio forms apply, and absence of degenerate or redundant sub-terms. Expressions were algebraically simplified (cancelling common factors, removing identity multiplications, merging duplicates) while preserving semantics. The selected SRVI and SRWI were the simplest expressions achieving near-maximal fitness under these constraints. All preprocessing and fitness computations used `NumPy`/`SciPy`, SR was run with `PySR`, protected operators used small ϵ guards, and random seeds were fixed for population initialisation and sampling. Code paths were constrained to Limassol training data to maintain a clean separation between development and the eleven global validation regions.

Standard indices for comparison

The performance of SRVI and SRWI was evaluated against a comprehensive suite of established spectral indices. These benchmarks were selected to represent the most common standards, historical advancements, and different formulation strategies for vegetation and water remote sensing. The performance evaluation for all indices was conducted on the same eleven globally distributed validation sites to ensure a direct and fair comparison.

Vegetation indices

For vegetation analysis, the selected indices represent a range of benchmarks, from the most common baseline (NDVI) to more advanced formulations designed to correct for specific environmental and atmospheric effects (SAVI, MSAVI2, EVI, NDRE).

- **NDVI:** The Normalized Difference Vegetation Index is the most widely used vegetation index and serves as the primary baseline for comparison. Its simple formulation is effective but has well-known limitations, such as signal saturation in dense vegetation[5].
- **SAVI and MSAVI2:** The Soil-Adjusted Vegetation Index and its successor, the Modified Soil-Adjusted Vegetation Index, were chosen because they represent a major advancement designed to minimize the effect of soil brightness in areas with sparse vegetation, a key weakness of NDVI[6,20].
- **EVI:** The Enhanced Vegetation Index was selected as a high-performance benchmark. It was designed to optimize the vegetation signal by correcting for both atmospheric influences and soil background noise, making it one of the most robust standard indices[7].
- **NDRE:** The Normalized Difference Red Edge Index was included to test performance specifically in high-biomass conditions. By using the red-edge band instead of the red band, NDRE is less susceptible to the signal saturation that affects NDVI, making it a crucial benchmark for dense canopy analysis[21].

Water indices

The water indices were chosen to provide a thorough comparison against the current operational standard (NDWI) and more recent indices developed to address specific challenges in water body delineation (MNDWI, AWEI, WI2015, TCW).

- **NDWI:** The Normalized Difference Water Index is the original, widely-used index for delineating open water and serves as the primary baseline. It is known to have limitations in distinguishing water from built-up features[8].
- **MNDWI:** The Modified Normalized Difference Water Index improves upon NDWI by using a SWIR band to more effectively suppress signals from built-up land, making it a crucial benchmark for comparison[9].
- **AWEI:** The Automated Water Extraction Index was selected because it was specifically formulated to improve classification accuracy in complex environments, such as urban areas with building shadows, where other indices can struggle[10].
- **WI2015:** The Water Index 2015 was chosen as it represents a more recent formulation designed to improve the separability of water from other land cover types[22].
- **TCW:** The Tasseled Cap Wetness component is not a simple ratio index but is derived from a standardized transformation. It is a well-established indicator of both soil moisture and open water, providing a different methodological benchmark[23].

The mathematical formulas for these standard indices are defined as follows:

$$\mathrm{NDVI} = \frac{(\mathrm{NIR} - \mathrm{Red})}{(\mathrm{NIR} + \mathrm{Red})} \tag{1}$$

$$\mathrm{EVI} = G \times \frac{(\mathrm{NIR} - \mathrm{Red})}{(\mathrm{NIR} + C_1 \times \mathrm{Red} - C_2 \times \mathrm{Blue} + L)} \tag{2}$$

$$\mathrm{SAVI} = \frac{(\mathrm{NIR} - \mathrm{Red})}{(\mathrm{NIR} + \mathrm{Red} + L_{savi})} \times (1 + L_{savi}) \tag{3}$$

$$\text{MSAVI2} = \frac{2 \times \text{NIR} + 1 - \sqrt{(2 \times \text{NIR} + 1)^2 - 8 \times (\text{NIR} - \text{Red})}}{2} \tag{4}$$

$$\text{NDRE} = \frac{(\text{NIR} - \text{RedEdge})}{(\text{NIR} + \text{RedEdge})} \tag{5}$$

$$\text{NDWI} = \frac{(\text{Green} - \text{NIR})}{(\text{Green} + \text{NIR})} \tag{6}$$

$$\text{MNDWI} = \frac{(\text{Green} - \text{SWIR1})}{(\text{Green} + \text{SWIR1})} \tag{7}$$

$$\text{AWEI} = 4 \times (\text{Green} - \text{SWIR1}) - (0.25 \times \text{NIR} + 2.75 \times \text{SWIR2}) \tag{8}$$

$$\text{WI2015} = 1.7204 + 171 \times \text{Green} + 3 \times \text{Red} - 70 \times \text{NIR} - 45 \times \text{RedEdge} - 71 \times \text{SWIR2} \tag{9}$$

$$\begin{aligned} \text{TCW} = c_B \times \text{Blue} + c_G \times \text{Green} + c_R \times \text{Red} \quad + c_{NIR} \times \text{NIR} + c_{SWIR1} \\ \times \text{SWIR1} + c_{SWIR2} \times \text{SWIR2} \end{aligned} \tag{10}$$

where for EVI, $G = 2.5$, $C_1 = 6.0$, $C_2 = 7.5$, and $L = 1.0$; for SAVI, the soil brightness correction factor $L_{savi} = 0.5$; and for TCW, the coefficients $(c_B, c_G, \ldots)$ are specific to the Sentinel-2 MSI sensor as defined in the literature[23].

Performance evaluation and statistical analysis

The performance of each index was quantified using the Jeffries-Matusita (JM) distance, a widely used metric for measuring the statistical separability between class distributions. The JM distance, which ranges from 0 (indistinguishable) to 2 (completely separable), was chosen because it considers both the distance between class means and the variance within each class, providing a more reliable measure of separability than simple differences in means. The analysis was conducted for three distinct tasks:

1. **Vegetation versus non-vegetation separation:** Assessing the ability to distinguish between a composite vegetation class (Tree cover, Shrubland, Grassland, Cropland) and a composite non-vegetation class (Built-up, Barren, Open water).
2. **Vegetation type separation:** Assessing the ability to distinguish between the four different vegetation types.
3. **Water body detection:** Assessing the ability to distinguish open water from all other land cover classes.

To ensure a robust comparison that accounts for phenological changes, monthly JM distance values were calculated for all unique class pairs within each task across all study regions. These monthly values were then averaged to produce a single, annual mean JM distance for each class pair. A comprehensive statistical framework was then employed on these annually averaged values:

- **Paired t-test and Wilcoxon signed-rank test:** Used to determine if the mean JM distance of a new index was statistically significantly higher than that of a standard index. A p-value < 0.05 was considered significant.
- **Cohen's d:** Calculated to measure the effect size (practical significance) of the difference. A value > 0.2 is conventionally considered to represent a small but meaningful improvement.
- **Two one-sided tests (TOST) for equivalence:** Used to test if the indices were statistically equivalent within a predefined margin (JM distance ±0.2).

Results

Symbolic regression produced two compact indices. Their structure and seasonal behaviour at the training site are analysed, and separability is quantified across eleven out-of-sample regions. Distributional analyses explain the observed gains and provide guidance for operational thresholding. No information from the validation regions was used during discovery, selection, or tuning.

Discovered index forms and structural properties

The vegetation index (SRVI) expression tree is shown in Fig. 2, with algebraic forms in Eqs. 11–12. The structure emphasises near-infrared relative to red while softly normalising by red, green, and SWIR1, a configuration that depresses values over bright soils and impervious surfaces and produces strongly negative responses over open water. The water index (SRWI) expression tree (Fig. 3) simplifies to a difference-over-sum contrast between (green+blue) and (NIR+SWIR1) (Eqs. 13–14), which remains positive for open water because of persistent NIR and SWIR1 absorption and negative for non-water classes, including urban and dry substrates. Both expressions are depth- and node-constrained, use protected operators, and admit only conservative unary functions, preserving interpretability and numerical robustness.

$$\text{SRVI} = \frac{(2.0 \times \text{NIR} - 3.0 \times \text{Red})}{(1.0 \times \text{NIR} + 1.0 \times \text{Red} + 0.5 \times \text{Green} + 0.5 \times \text{SWIR1})}, \tag{11}$$

$$\boxed{\text{SRVI} = \frac{2\,\text{NIR} - 3\,\text{Red}}{\text{NIR} + \text{Red} + 0.5\,(\text{Green} + \text{SWIR1})}}. \tag{12}$$

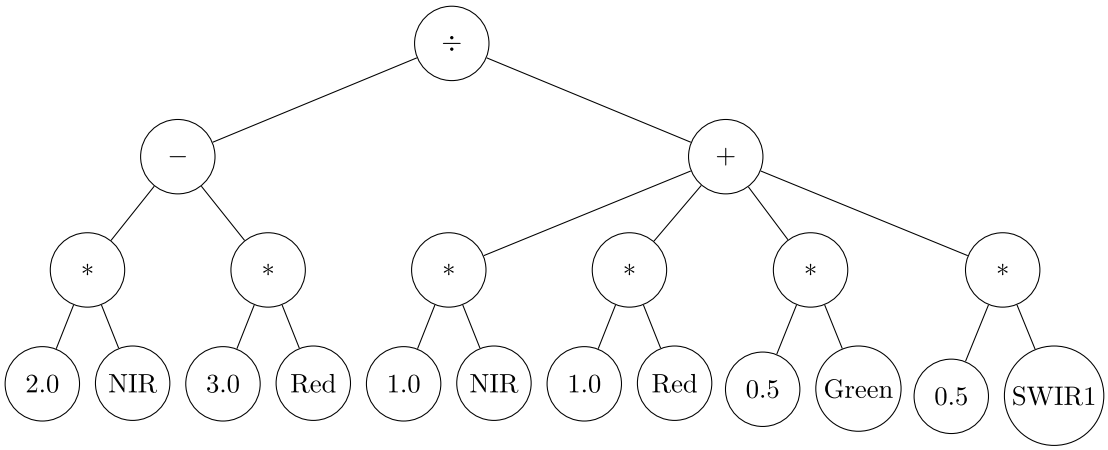


Fig. 2. Expression tree for the vegetation index (SRVI).

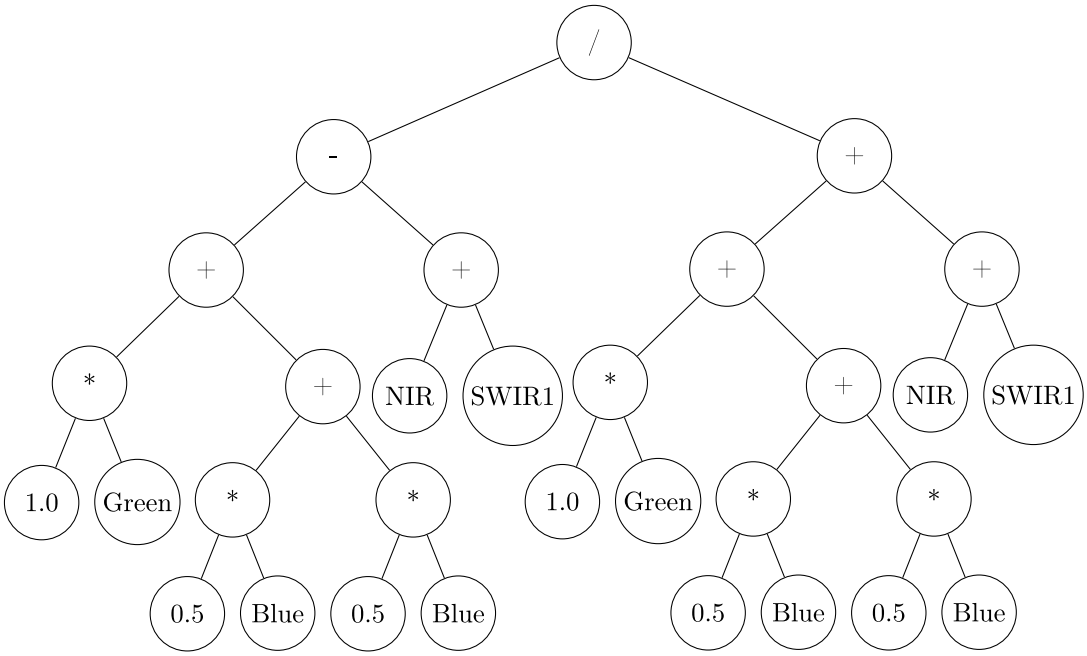


Fig. 3. Initial expression tree for the water index (SRWI).

$$\mathrm{SRWI} = \frac{(1.0 \cdot \mathrm{Green} + (0.5 \cdot \mathrm{Blue} + 0.5 \cdot \mathrm{Blue})) - (\mathrm{NIR} + \mathrm{SWIR1})}{(1.0 \cdot \mathrm{Green} + (0.5 \cdot \mathrm{Blue} + 0.5 \cdot \mathrm{Blue})) + (\mathrm{NIR} + \mathrm{SWIR1})}, \tag{13}$$

$$\boxed{\mathrm{SRWI} = \frac{(\mathrm{Green} + \mathrm{Blue}) - (\mathrm{NIR} + \mathrm{SWIR1})}{(\mathrm{Green} + \mathrm{Blue}) + (\mathrm{NIR} + \mathrm{SWIR1})}}. \tag{14}$$

Seasonal behaviour at the training site

Monthly behaviour over Limassol (Figs. 4 and 5) shows that the vegetation index captures realistic phenology with enhanced seasonal contrast relative to standard indices, including a pronounced late-winter to spring peak over croplands characteristic of Mediterranean rain-fed cycles. The water index maintains a stable, distinct positive signature for open water throughout the year and negative values for non-water classes. These properties, imposed only through weak phenological guidance in the objective, indicate that the learned ratios respect underlying spectral physics rather than overfitting idiosyncratic monthly patterns.

Statistical separability across global validation sites

Separability on the eleven out-of-sample regions, averaged across months to account for phenology, was quantified with the Jeffries–Matusita distance (Table 2). For vegetation versus non-vegetation, the vegetation

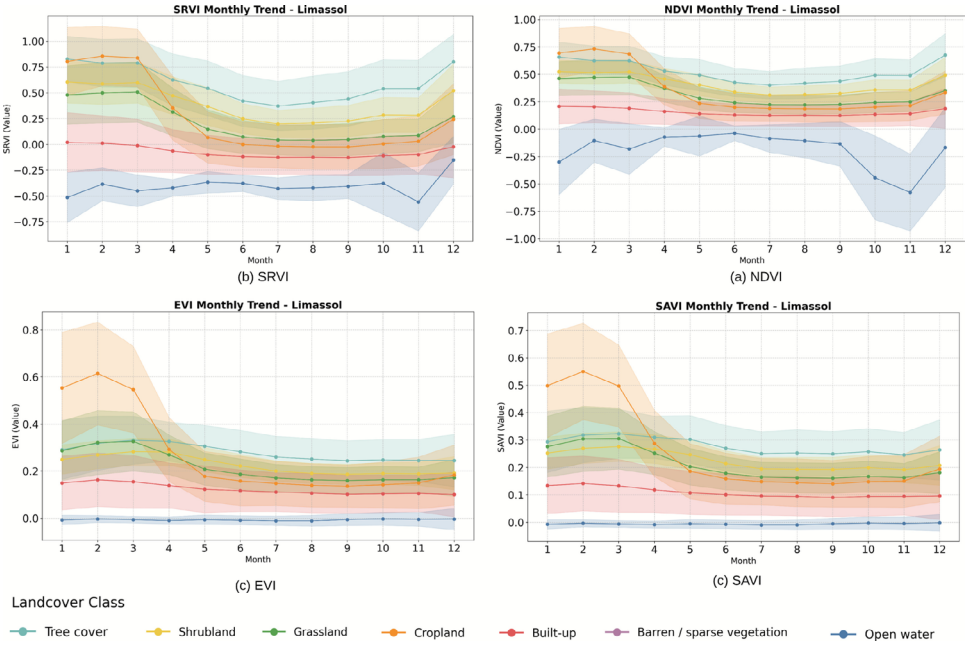


Fig. 4. Monthly phenological trends for vegetation indices at the Limassol training site.

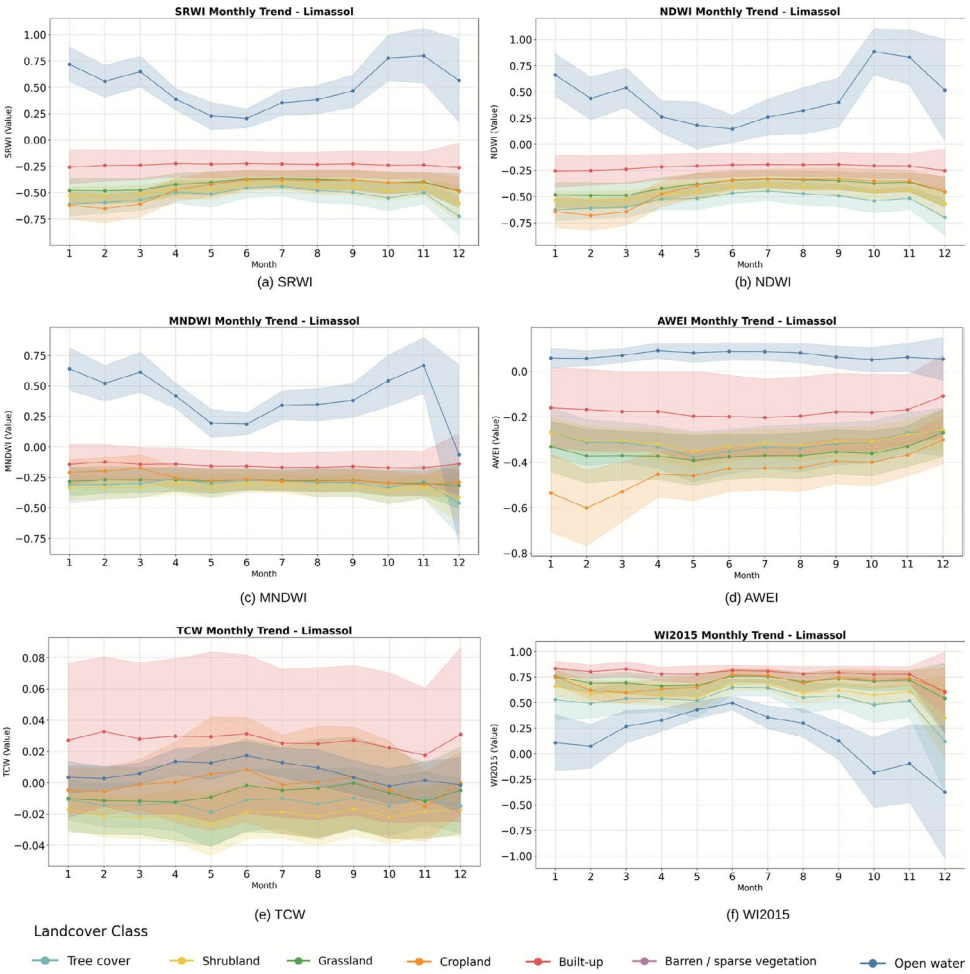


Fig. 5. Monthly phenological trends for water indices at the Limassol training site.

Compared index	JM Distance		*p*-value			Cohen's d
	Mean	Std	Wilcox	*t*-test	TOST	
Vegetation versus non-vegetation (SRVI: Mean JM = 1.15, SD = 0.51)						
NDVI	1.13	0.50	< 0.001	< 0.001	< 0.001	0.37
EVI	1.01	0.49	< 0.001	< 0.001	0.002	0.62
SAVI	1.10	0.51	0.001	0.002	< 0.001	0.27
NDRE	0.98	0.47	< 0.001	< 0.001	0.057	1.03
MSAVI2	1.06	0.48	< 0.001	< 0.001	< 0.001	0.42
Vegetation type separation (SRVI: Mean JM = 0.42, SD = 0.34)						
NDVI	0.41	0.32	0.299	0.158	< 0.001	0.18
EVI	0.37	0.33	0.090	0.059	< 0.001	0.24
SAVI	0.37	0.33	0.038	0.030	< 0.001	0.27
NDRE	0.37	0.26	0.038	0.007	< 0.001	0.34
MSAVI2	0.37	0.32	0.080	0.044	< 0.001	0.25
Water detection (SRWI: Mean JM = 1.53, SD = 0.33)						
NDWI	1.46	0.36	< 0.001	< 0.001	< 0.001	0.60
MNDWI	1.47	0.32	0.002	0.002	< 0.001	0.39
AWEI	1.48	0.42	0.654	0.143	< 0.001	0.18
WI2015	0.78	0.48	< 0.001	< 0.001	1.000	1.36
TCW	0.81	0.34	< 0.001	< 0.001	1.000	1.94

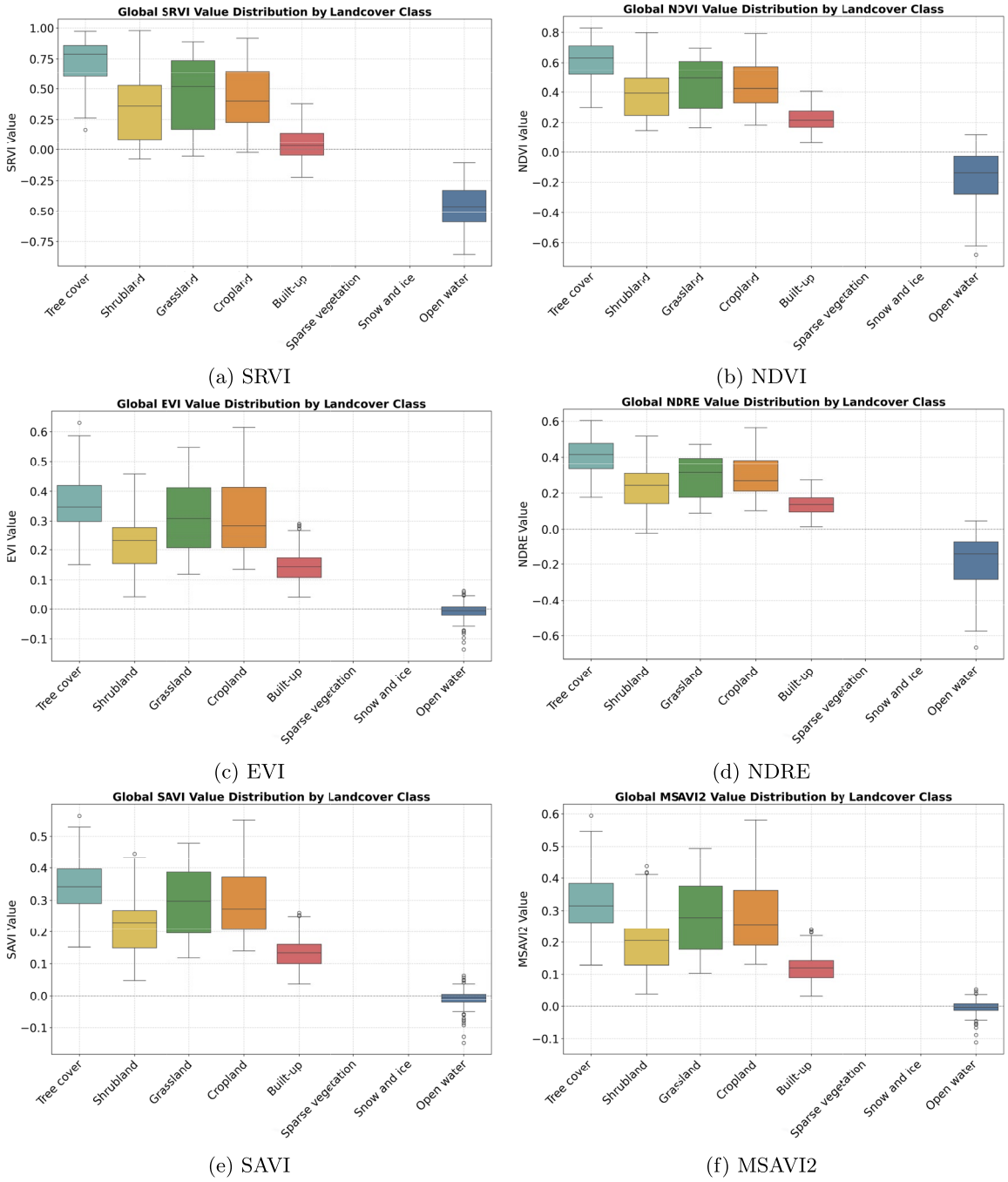
Table 2. Statistical comparison of index performance across eleven out-of-sample regions. Jeffries–Matusita (JM) ranges from 0 (indistinguishable) to 2 (fully separable).

index exceeds all benchmark indices with statistically significant paired improvements (Wilcoxon and paired t-tests, $p < 0.01$ in all cases) and practically meaningful effect sizes (e.g., Cohen's $d = 0.62$ versus EVI and $d = 0.42$ versus MSAVI2). For within-vegetation discrimination, consistent gains are observed relative to SAVI, NDRE, and MSAVI2 with small-to-moderate effect sizes and $p < 0.05$. For water detection, the water index outperforms NDWI and MNDWI with significant improvements and moderate effect sizes (e.g., $d = 0.60$ versus NDWI), and it vastly exceeds WI2015 and TCW. Performance is comparable to AWEI on average; however, the sign convention (water > 0) and compact four-band ratio simplify thresholding and transfer.

Global value distributions and integrated interpretation

Distributional analyses across land-cover classes (Figs. 6 and 7) clarify why the separability gains arise and how they should be operationalised. For vegetation mapping, the vegetation index exhibits a wider dynamic range than several benchmarks and maintains separation at upper quantiles while pushing non-vegetation (especially open water, built-up, and barren/sparse vegetation) further negative. This increases the separation margin in the binary task and reduces threshold sensitivity. Crucially, as observed in the global distribution plots, NDVI values for dense vegetation classes (Tree cover, Cropland) are highly compressed near the upper limit, confirming known saturation issues. In contrast, SRVI exhibits a significantly wider dynamic range for these same classes, providing empirical evidence that it retains sensitivity in high-biomass conditions. Second, while this study validated performance across eleven sites distributed globally, it does not constitute a continuous global thematic map. The results validate the potential for global application, but operational deployment would require further testing on cloud-computing platforms to assess consistency across continental scales. Within vegetation, medians and interquartile ranges for Tree cover, Grassland, Cropland, and Shrubland are more clearly separated than with several comparators, consistent with the small-to-moderate gains in Table 2. Operationally, a low preliminary threshold excludes most non-vegetation using negative and near-zero responses, after which an application-specific threshold can be selected via a small labelled subset or an automatic criterion such as Otsu's method. Edge cases include bright saline flats, highly reflective rooftops, and sub-pixel vegetation mixed with reflective substrates, where ancillary masking or sensor fusion is beneficial.

For water detection, the water index produces a tight, positive cluster for open water and consistently negative values for other classes, enabling a simple near-zero global threshold with minimal tuning. Compared with NDWI and MNDWI, the water cluster is narrower (lower variance), and built-up or shadowed pixels are pushed further negative, reducing false positives in urban and mountainous scenes. Shallow or highly turbid water remains positive because NIR and SWIR1 absorption persists; extremely bright turbidity or dense floating vegetation can move values toward zero, where a slightly positive threshold or a two-stage rule (water index plus an SWIR reflectance check) restores precision. Relative to AWEI, aggregate performance is similar, but the sign convention and compact four-band form simplify transfer. Post-processing that removes small components and shoreline speckle, and optionally re-admits coherent hydrologic components, further stabilises outputs across diverse environments.



(a) SRVI (b) NDVI

(c) EVI (d) NDRE

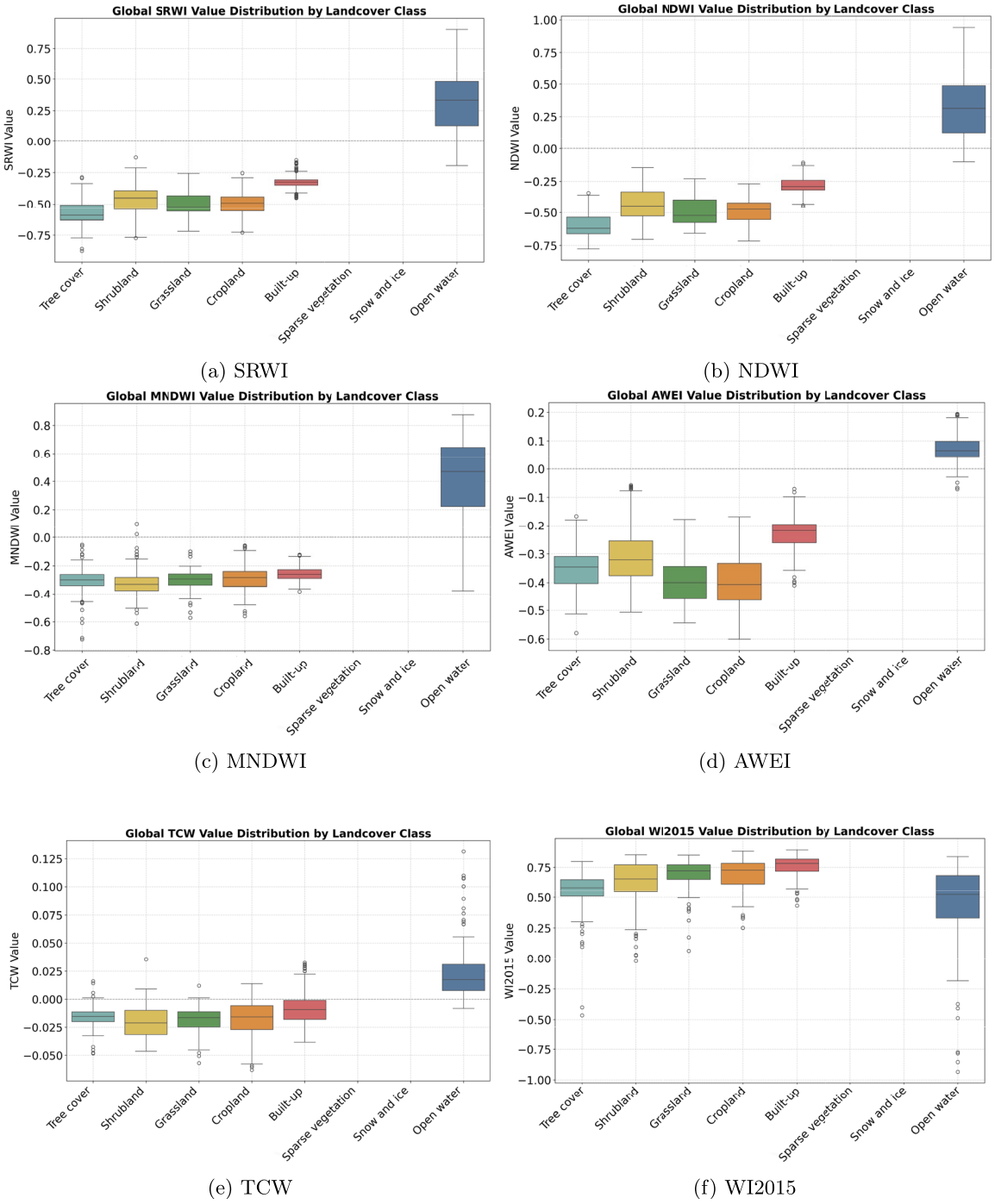
(e) SAVI (f) MSAVI2

Fig. 6. Global value distributions for the vegetation index and standard vegetation indices across land-cover classes. Box plots show median, interquartile range, and overall range.

Discussion

The findings demonstrate that symbolic regression can recover compact, interpretable formulas that capture spectral contrasts not well represented by conventional normalised-difference or linear-combination indices. The vegetation index emphasises near-infrared against red while softly normalising by red, green, and SWIR1, which systematically depresses values over bright soils, impervious surfaces, and water. The water index expresses a difference-over-sum between visible (green+blue) and shortwave components (NIR+SWIR1), preserving a robust positive response for open water and negative responses elsewhere. These designs are consistent with known absorption and reflectance behaviours and with the weak phenological guidance used during search, suggesting that the discovered forms align with underlying physics rather than exploiting spurious correlations.

Across eleven out-of-sample regions, separability improvements are statistically significant and practically meaningful. For vegetation versus non-vegetation, gains relative to established indices persist across biomes and seasons, and within-vegetation separability improves despite the challenge of overlapping canopy structures



(a) SRWI (b) NDWI

(c) MNDWI (d) AWEI

(e) TCW (f) WI2015

Fig. 7. Global value distributions for the water index and standard water indices across land-cover classes. Box plots show median, interquartile range, and overall range.

and mixed phenologies. For water detection, performance exceeds the common baselines, with narrower water distributions and stronger suppression of urban/shadow confounders; parity with AWEI on aggregate is accompanied by a simpler, sign-consistent formulation that eases threshold transfer. Together with the seasonal behaviour at the training site, these results indicate that expressions learned on a spectrally complex Mediterranean scene can generalise when evaluated out of sample across diverse landscapes.

Several aspects qualify these conclusions. First, optimisation targeted agreement with ESA WorldCover labels. While WorldCover is high quality, label noise, class boundary uncertainty, and local misclassifications propagate to the objective and can bias the learned trade-offs between parsimony and separability. Second, the analysis operated on Sentinel-2 Level-2A reflectance resampled to 10 m, with cloud/shadow filtering and robust per-band scaling. Residual atmospheric effects, adjacency, BRDF and view/illumination geometry differences, and resampling artefacts may still influence index values, particularly along sharp boundaries (shorelines, field

edges) and in high-relief terrain. Third, index thresholds were discussed operationally but not globally optimised; simple near-zero rules work well in aggregate, yet local adaptation (e.g., small positive offsets for highly turbid waters, or ancillary masks for saline flats and reflective rooftops) further improves precision. Finally, the search space excluded logarithmic, exponential, and trigonometric operators to favour interpretability and numerical stability; while this choice aided transfer, it may also preclude alternative compact expressions with competitive performance.

These caveats motivate several extensions. Ground-reference campaigns should relate the vegetation index to biophysical variables (e.g., LAI, FAPAR, chlorophyll proxies) and the water index to turbidity, depth, and colour, quantifying sensitivity and saturation regimes. Cross-sensor transfer should be evaluated explicitly–harmonised Landsat-8/9, Sentinel-3, and forthcoming missions–considering spectral response differences and band availability; simple coefficient re-fitting or constrained re-search may suffice. Scene-level corrections (topographic normalisation, BRDF adjustment, shadow handling) and spatial regularisation can be assessed to stabilise boundaries without eroding small features. Finally, training on alternative single-site "microcosms" and on multi-site mixtures would probe solution stability and identify families of near-optimal formulas with similar semantics but different robustness profiles.

Conclusion

This work introduces two compact spectral indices discovered via symbolic regression and validates their performance across eleven out-of-sample regions. The vegetation index increases separability between vegetation and non-vegetation and improves discrimination among vegetation types relative to widely used baselines. The water index provides consistently positive responses for open water and stronger suppression of urban/shadow confounders than common alternatives, while retaining a simple ratio structure amenable to portable thresholds.

Beyond the specific formulas, the study shows that data-driven expression discovery can yield interpretable indices that respect spectral physics, exhibit realistic seasonal behaviour, and generalise geographically when derived from a carefully chosen, spectrally heterogeneous site. These properties make the indices practical for global mapping pipelines and for applications such as land-use change detection, crop condition screening, flood mapping, and surface-water inventories.

Data availability

Copernicus Sentinel-2 Level-2A and ESA WorldCover 2021 data are freely available from their respective official portals. Code and derived data products are available from the corresponding author upon reasonable request.

Received: 6 October 2025; Accepted: 30 December 2025
Published online: 13 January 2026

References

1. IPBES, B. E. et al. Global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services. *IPBES Secretariat***1148** (2019).
2. Forster, P. M. et al. Indicators of Global Climate Change 2023: annual update of key indicators of the state of the climate system and human influence. *Earth System Science Data***16**, 2625–58 (2024).
3. Pettorelli, N., Bühne, H., Schulte to, Tulloch, A., et al. Satellite remote sensing of ecosystem functions: opportunities, challenges and way forward. *Remote Sensing in Ecology and Conservation***4**,71–93 (2018).
4. Xue, J. & Su, B. Significant remote sensing vegetation indices: A review of developments and applications. *Journal of sensors***2017**, 1353691 (2017).
5. Rouse, J. W., Haas, R. H., Schell, J. A. & Deering, D. W. Monitoring vegetation systems in the Great Plains with ERTS. *NASA Special Publication***351**, 309 (1974).
6. Huete, A. R. A soil-adjusted vegetation index (SAVI). *Remote Sensing of Environment***25**, 295–309 (1988).
7. Huete, A. et al. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. *Remote Sensing of Environment***83**, 195–213 (2002).
8. 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**, 1425–32 (1996).
9. Xu, H. Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery. *International Journal of Remote Sensing***27**, 3025–33 (2006).
10. Feyisa, G. L., Meilby, H., Fensholt, R. & Proud, S. R. Automated Water Extraction Index: A new technique for surface water mapping using Landsat imagery. *Remote Sensing of Environment***140**, 23–35 (2014).
11. Augusto, D. A., Barbosa, H. J. Symbolic regression via genetic programming. In: Proceedings. Vol. 1. Sixth Brazilian symposium on neural networks. IEEE. 173–8 (2000).
12. Zhong, J., Feng, L., Cai, W. & Ong, Y. S. Multifactorial genetic programming for symbolic regression problems. *IEEE transactions on systems, man, and cybernetics: systems***50**, 4492–505 (2018).
13. Cranmer, M. PySR: High-Performance Symbolic Regression in Python and Julia. *Astrophysics Source Code Library* 2ascl–2409 (024).
14. Schmidt, M. & Lipson, H. Distilling free-form natural laws from experimental data. *Science***324**, 81–5 (2009).
15. Wang, Y., Wagner, N. & Rondinelli, J. M. Symbolic regression in materials science. *MRS communications***9**, 793–805 (2019).
16. Makke, N. & Chawla, S. Interpretable scientific discovery with symbolic regression: a review. *Artificial Intelligence Review***57**, 2 (2024).
17. European Space Agency. Copernicus Sentinel-2. Accessed: 01/06/2025. (2024). https://sentinels.copernicus.eu/web/sentinel/missions/sentinel-2.
18. Zanaga, D., Van De Kerchove, R., Daems, D., et al. ESA WorldCover 10 m 2021 v200. (2022).
19. Gorelick, N. et al. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote sensing of Environment***202**, 18–27 (2017).
20. Qi, J., Chehbouni, A., Huete, A. R., Kerr, Y. H. & Sorooshian, S. A modified soil adjusted vegetation index. *Remote Sensing of Environment***48**, 119–26 (1994).
21. Gitelson, A. A. & Merzlyak, M. N. Quantitative estimation of chlorophyll a using reflectance spectra: Experiments with autumn chestnut and maple leaves. *Journal of Photochemistry and Photobiology B: Biology***22**, 247–52 (1994).

22. Fisher, A., Flood, N., Danaher, T. The water index (WI2015) for Sentinel-2 and Landsat-8. In: Proceedings of the 18th Australasian Remote Sensing and Photogrammetry Conference. Melbourne, Australia, 23–8 (2016).
23. Nedkov, R. Tasseled Cap Transformation for Sentinel-2 Images. *Journal of Geoinformatics and Spatial Analysis***1**, 87–95 (2017).

Acknowledgements
The authors would like to acknowledge the use of the Sentinel-2 data provided by the European Space Agency.

Author contributions
Conceptualization: C.C; Data analysis: C.C and S.P.N.; Data collection: C.C. Writing–original draft preparation: C.C., S.P.N. and M.M; Review and editing: C.C, D.G.H., and M.M.

Funding
This work was supported by the European Union's HORIZON Research and Innovation Programme by the 'EXCELSIOR': ERATOSTHENES: Excellence Research Centre for Earth Surveillance and Space-Based Monitoring of the Environment H2020 Widespread Teaming project (www.excelsior2020.eu). The 'EXCELSIOR' project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 857510, from the Government of the Republic of Cyprus through the Directorate General for the European Programmes, Coordination and Development and the Cyprus University of Technology.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to C.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.