remote sensing

MDPI

Article

Robust Soil Salinity Retrieval Under Small-Sample and High-Dimensional Hyperspectral Conditions via Physically Constrained Generative Augmentation

Shan Yu [1], Lide Su [2,*], Wala Du [3,4], Deji Wuyun [2], Han Gao [5], Liangliang Yu [6], Yuxin Zhao [6], A Ruhan [2] and Rong Li [2]

[1] College of Artificial Intelligence, Inner Mongolia Normal University, Hohhot 010022, China; yushan@imnu.edu.cn
[2] College of Geographic Science, Inner Mongolia Normal University, Hohhot 010022, China; wuyundeji@caas.cn (D.W.); 20172104155@mails.imnu.edu.cn (A.R.); 20224016019@mails.imnu.edu.cn (R.L.)
[3] Institute of Grassland Research, Chinese Academy of Agricultural Sciences, Hohhot 010010, China; duwala@caas.cn
[4] Arxan Forest and Grassland Disaster Prevention and Mitigation Field Scientific Observation and Research Station of Inner Mongolia Autonomous Region, Arxan 137400, China
[5] College of Ocean and Space Information, China University of Petroleum (East China), Qingdao 266580, China; gaohangeo@upc.edu.cn
[6] Bayannur Meteorological Bureau, Bayannur 015000, China; yullxy@126.com (L.Y.); a2291516@163.com (Y.Z.)
* Correspondence: 20224016062@mails.imnu.edu.cn

Highlights

What are the main findings?

- OBCA identifies a compact, cross-site robust VNIR feature subset (three-band synergies) that strengthens salinity sensitivity under heterogeneous field conditions.
- The physically constrained S-WGAN-GP with Teacher reliability screening produces high-confidence synthetic spectra that improve LOOCV performance and stabilize prediction.

What are the implications of the main findings?

- Reliable GAN-augmented training can mitigate small-sample bottlenecks while preserving spectral plausibility, enabling more transferable soil salinity inversion workflows.
- The resulting maps differentiate patchy, extensive, and edge-concentrated salinization patterns, supporting targeted management and site-specific remediation planning.

Abstract

Soil salinity mapping in heterogeneous irrigation districts faces a dual challenge: the high dimensionality of hyperspectral data leads to redundancy, while the scarcity of ground-truth samples restricts the generalization of data-driven models. Traditional regression methods often struggle to capture non-linear spectral responses under such "small-sample" conditions. To address these limitations, this study proposes a semi-supervised retrieval framework coupling Optimal Band Combination Analysis (OBCA) with a Spectral Wasserstein GAN with Gradient Penalty (S-WGAN-GP). We constructed a robust feature set via cross-scenario evaluation and developed a rigorous "Uncertainty-Aware Filtering" protocol to screen synthetic samples generated by a teacher mechanism. The OBCA screening revealed that salinity-sensitive features are robustly clustered in the Green (550–570 nm) and Near-Infrared (NIR, 880–950 nm) regions, with NIR bands demonstrating superior stability across different sites. The proposed S-WGAN-GP successfully densified the feature

Check for updates

Academic Editor: Nicolas Baghdadi

Received: 13 January 2026
Revised: 13 February 2026
Accepted: 25 February 2026
Published: 2 March 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

manifold by generating 1186 high-fidelity synthetic samples. By incorporating these augmented data, the inversion accuracy was substantially improved: the R^2 of the optimal SVR model increased from 0.36 (baseline) to 0.60 (+66.7%), and the RMSE decreased from 7.06 to 5.57 dSm^{-1}. This study confirms that physically constrained generative augmentation, when combined with rigorous quality control, effectively bridges the distribution gap in limited datasets. The proposed framework offers a transferable and accurate solution for fine-scale soil salinity monitoring in data-scarce arid regions.

Keywords: soil salinity inversion; UAV hyperspectral remote sensing; generative adversarial networks; semi-supervised learning; optimal band combination analysis

1. Introduction

Soil salinization is a critical eco-environmental constraint on sustainable agriculture and food security in arid and semi-arid regions, where coupled climate–hydrology–management interactions accelerate salt accumulation and amplify land degradation risks [1–3]. Recent syntheses further emphasize that salinity expansion and secondary salinization are intensifying in drylands and irrigated agroecosystems, making monitoring and management a priority for achieving food-security goals [4,5]. Soil salinity is commonly characterized by the electrical conductivity of the saturated paste extract (ECe), which is widely used as an operational indicator of crop salt-stress risk and management zoning in remote-sensing–driven salinity studies [6,7]. Therefore, fine-scale characterization of ECe spatial heterogeneity is essential for salinization assessment, zonal governance, and precision irrigation management, particularly in intensively irrigated districts with pronounced scale effects between point samples and pixel observations [8,9].

In recent years, UAV-based hyperspectral remote sensing—benefiting from centimeter-level spatial detail and continuous narrow-band spectral information—has demonstrated strong capability for capturing subtle spectral differences associated with soil salinity and secondary salinization in irrigated farmlands [10–15]. However, hyperspectral inputs remain high-dimensional and strongly collinear, and their performance is sensitive to noise and acquisition conditions, which can lead to band redundancy and unstable feature representations if not properly constrained [16]. Meanwhile, conventional empirical indices and fixed-band salinity indicators may suffer from comparability issues, naming/definition ambiguity, and limited validity domains across sensors, soil backgrounds, and moisture regimes—ultimately reducing transferability across regions and scenarios [6,17,18]. Consequently, identifying physically meaningful and ECe-sensitive band combinations from the full spectrum has become a key step to improve robustness and interpretability [19–23]. To enhance weak salinity-related spectral signatures and mitigate baseline/background effects, fractional-order differentiation (FOD) has been increasingly adopted as a spectral transformation technique in soil VIS–NIR/hyperspectral analysis. Compared with integer-order derivatives, low-order FOD provides a flexible way to emphasize subtle absorption/shape information while controlling noise amplification, thereby improving correlation with soil properties [24]. In this context, optimal band combination analysis and 2D/3D salinity-index construction—often coupled with fractional-order differentiation—have shown promise for mining synergistic band responses and improving salinity retrieval performance with clearer spectral interpretation [19]. Notably, recent UAV hyperspectral studies have demonstrated that OBCA can strengthen salinity–spectral correlations and improve mapping stability; yet, under small and spatially clustered field samples, regression generalization remains limited and sensitive to site-specific background factors [25].

For retrieval modeling, machine learning (ML) integrated with multi-source remote sensing has become a dominant paradigm for soil salinity mapping due to its capacity to capture non-linear spectral–salinity relationships and enable uncertainty-aware, explainable interpretation [26,27]. Nevertheless, model effectiveness is still constrained by the quantity, representativeness, and spatial balance of ground-truth: data-scarce environments and spatially clustered sampling often induce overfitting and poor cross-location generalization [28–30]. This limitation is particularly severe in irrigated districts (e.g., Hetao and similar systems), where salinity exhibits strong spatial heterogeneity and multi-scale coupling with soil texture, groundwater, and irrigation practices, and where pixel–point scale mismatch can further degrade inversion fidelity [6]. Recent efforts such as ground–UAV–satellite synergistic modeling and irrigation-district-scale inversion combined with process-based salt-migration simulation highlight the importance—but also the data demand—of scalable and stable salinity inversion [8,16,31]. Therefore, to mitigate sample scarcity and enable robust cross-regional learning, it is necessary to develop advanced sample generation and data augmentation strategies that can expand the effective feature space beyond limited field observations [29].

To address ground-truth scarcity, generative data augmentation has been increasingly explored in hyperspectral/remote-sensing learning [32,33]. Although GAN-based augmentation can improve data diversity and downstream accuracy, many studies focus primarily on classification gains, while the physical fidelity and trustworthiness of generated spectra (e.g., smooth spectral continuity, realistic absorption-feature morphology, and distributional alignment to site conditions) are less explicitly controlled [34,35]. Moreover, without rigorous quality-control protocols, synthetic samples may introduce domain shift or biased supervision signals that undermine model stability and transferability, especially for regression tasks [36,37]. This creates a key methodological gap for salinity inversion: under high-dimensional hyperspectral conditions, augmentation must be both (i) physically grounded in reflectance-domain behavior and (ii) reliability-controlled to prevent pseudo-label noise accumulation. Recent remote-sensing GAN studies increasingly emphasize physics-informed constraints (e.g., spectral-angle-based losses and domain/physical regularization) as essential to preserve intrinsic material signatures under varying acquisition conditions [38]. Therefore, a principled framework is needed that jointly links (a) salinity-sensitive band synergy discovery, (b) physically constrained spectral generation, and (c) uncertainty/consistency-aware pseudo-label filtering—so that augmented samples remain realistic, label-consistent, and transferable across heterogeneous irrigation scenes.

In view of these challenges, this study proposes a physically constrained generative framework tailored for robust soil salinity mapping in heterogeneous irrigation districts. As summarized in Figure 1, we develop a Spectral Wasserstein GAN with Gradient Penalty (S-WGAN-GP) coupled with a teacher-guided reliability filtering (TRF) module to generate physically plausible synthetic spectra and retain only trustworthy pseudo-samples for downstream inversion. In contrast to conventional augmentation pipelines that primarily pursue sample diversity, our framework explicitly embeds spectral physical constraints during generation and subsequently screens synthetic spectra using credibility- and distribution-consistency criteria with respect to real field observations before model training. The innovation of this work lies in transforming "more data" augmentation into "trustworthy, physically consistent" augmentation for hyperspectral salinity regression under small-sample settings: OBCA provides interpretable band synergies; S-WGAN-GP preserves spectral-shape integrity under reflectance-domain constraints; and TRF prevents confirmation bias by discarding low-credibility pseudo-labels. Specifically, this study aims to: (1) construct a generative model that preserves the physical integrity of hyperspectral salinity signatures; (2) design a teacher-based filtering protocol that removes low-credibility

pseudo-samples and mitigates error propagation; and (3) systematically evaluate the proposed strategy under strict cross-location validation to quantify its improvements in transferability and prediction accuracy across three sites in the Hetao Irrigation District, Inner Mongolia.

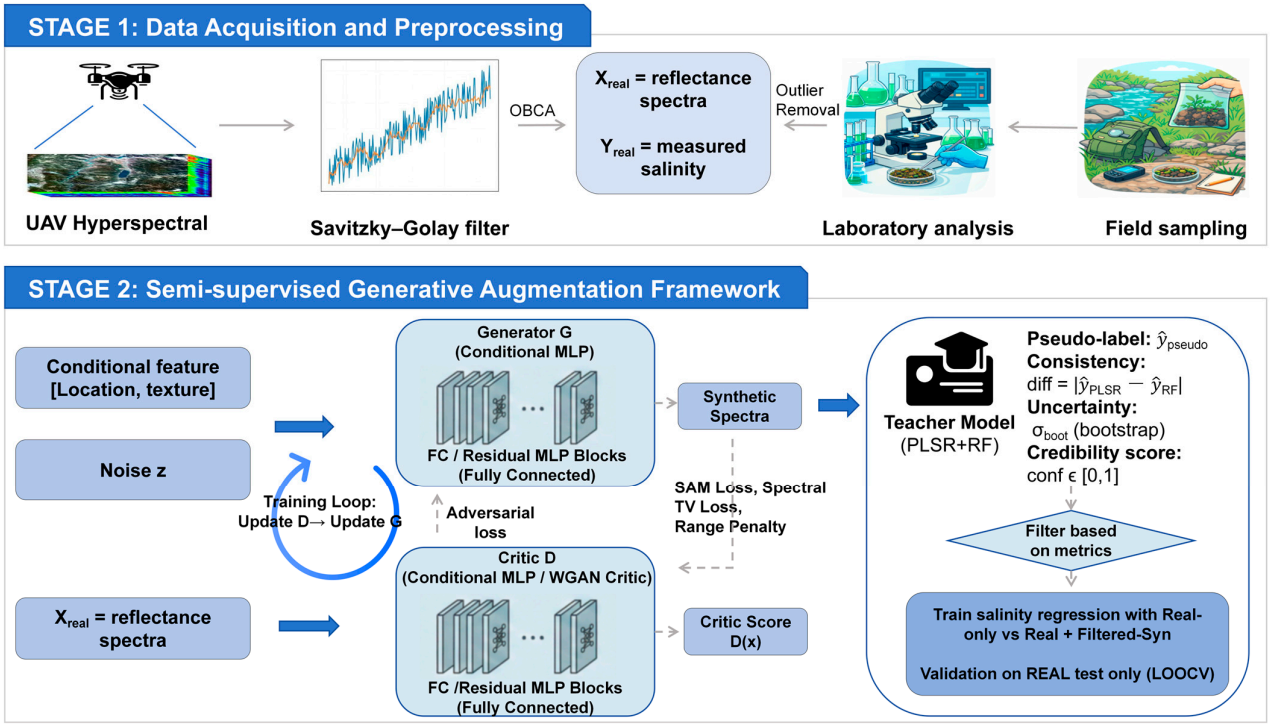


Figure 1. Overview of the proposed workflow. Stage 1 summarizes UAV hyperspectral acquisition, preprocessing, and the construction of the real labeled set (X_{real},Y_{real}) with outlier removal. Stage 2 illustrates the conditional S-WGAN-GP with reflectance-domain constraints (SAM loss, spectral TV loss, range penalty), followed by Teacher-based pseudo-labeling (PLSR + RF) and metric-based filtering. The final regression is trained using Real-only versus Real + Filtered-Syn, and all validations are performed on real samples only (LOOCV).

2. Materials and Methods

2.1. Study Area

The Hetao Irrigation District (HID) in Bayannur, western Inner Mongolia, China, was selected as the core experimental region (Figure 2). The HID (approximately 40°19′–41°18′N, 106°20′–109°19′E) lies along the northern side of the Yellow River's "Great Bend" (Ordos Loop) and is recognized as one of the largest arid irrigation districts in Northwest China [39]. Agricultural production in the HID is sustained by Yellow River diversion irrigation delivered through an extensive canal-based conveyance and field irrigation system, where non-growing-season (e.g., autumn/winter) flood irrigation is also widely practiced and has been mapped at district scale using remote sensing [40].The regional climate is arid to semi-arid, with low precipitation and strong atmospheric evaporative demand; recent district-scale studies report precipitation on the order of ~160 mm while potential evaporation can exceed ~2000 mm, indicating an evaporation–precipitation imbalance close to an order of magnitude [41]. This persistent hydroclimatic deficit, together with irrigation-induced shallow groundwater and strong capillary rise, favors near-surface salt accumulation and secondary salinization in cultivated lands, motivating both monitoring and management-oriented research in the HID [42].

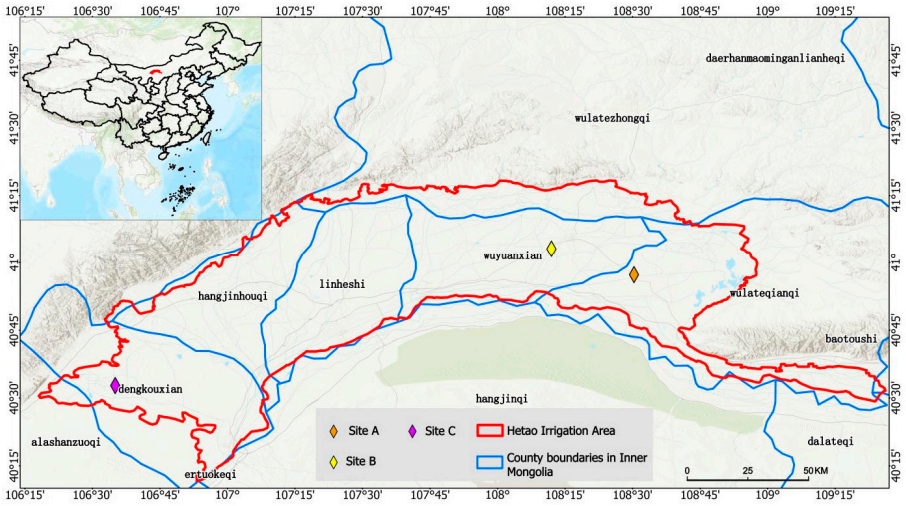


Figure 2. Overview map of the study area.

To comprehensively evaluate model generalization across varying geographic settings and salinization levels, we adopted the principle of spatial representativeness and typical salinity gradients, selecting three representative sub-areas distributed from west to east across the district (Figure 3):

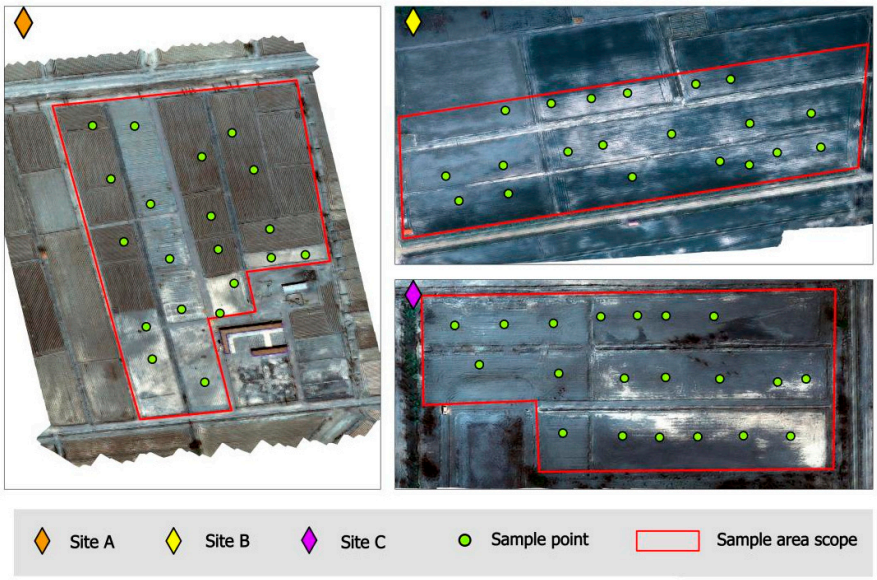


Figure 3. UAV images and sample point distribution.

Site A (downstream): Characterized by shallow groundwater and mixed land covers (cropland and shelterbelt/intercropped fields) with varying vegetation fractions, providing a challenging setting for robust retrieval under different surface conditions.

Site B (midstream core): A low-lying area with poor drainage where severe salinization is concentrated; salt crusts and salt patches are frequently observed on the soil surface, representing high-salinity extremes and strong spatial heterogeneity.

Site C (upstream): Dominated by sandy loam to loam textures, mainly covering alkaline wasteland and reclaimed cropland, representing relatively mild-to-moderate salinity conditions.

These three sites span distinct hydrogeological settings and soil textures within the HID and exhibit pronounced spatial heterogeneity in salinity, thus constituting an ideal

testbed for assessing the cross-location robustness and transferability of hyperspectral retrieval models.

2.2. Ground Sample Collection and Laboratory Analysis

Field sampling was conducted from 11 to 13 April 2023. All three study sites were bare-soil croplands in spring (unploughed and free of vegetation cover). During the sampling period, no precipitation events occurred, wind speeds were low, and near-surface illumination conditions remained comparatively stable. To ensure consistency between airborne and ground observations, a spatiotemporally synchronized sampling strategy was adopted: all ground sampling was completed within a ± 2 h window of the corresponding UAV overflight, thereby minimizing the influence of changes in solar elevation and short-term irradiance fluctuations on surface reflectance and improving the comparability between remotely sensed features and in situ salinity measurements.

Sampling followed a grid-based, spatially balanced design. Sampling plots were pre-defined in each site, and soil samples were collected at fixed intervals to form a spatially representative sampling network. In total, 60 surface soil samples were collected across the three sites (with approximately balanced numbers per site and/or covering the salinity gradient). All sampling locations were georeferenced using a Trimble R10 RTK GNSS (horizontal accuracy ± 1 cm, vertical accuracy ± 2 cm). The sampling depth was controlled at 5 ± 0.5 cm to characterize salinity conditions in the bare-soil surface layer. Immediately after collection, samples were sealed, protected from light, and transported to the laboratory under 4 °C refrigerated conditions to maintain physicochemical stability (Figure 4).



Figure 4. Salt sample organization.

Laboratory salinity was quantified using the electrical conductivity (EC) method. After air-drying, soils were sieved through a 2 mm mesh. A 1:2 (w/w) soil-to-distilled-water suspension was prepared, magnetically stirred for 30 min at 25 ± 1 °C, and allowed to settle for 1 h. The electrical conductivity of the supernatant was measured using a calibrated HI98331 digital conductivity meter and reported as EC_{1:2} (dSm^{-1}).

To align with the widely used agronomic indicator, EC_{1:2} was converted to the electrical conductivity of the saturated paste extract (ECe) using a published inter-method calibration relationship. Hogg and Henry [43] reported a near-linear relationship between ECe and EC_{1:2} across a wide range of soils, with slopes around 2.27–2.35 (e.g., ECe = 2.27EC1:2 − 0.08 for all soils and ECe = 2.35EC1:2 − 0.36 for medium-textured soils). Therefore, we adopted the practical approximation:

$$ECe = 2.3 \times EC_{1:2} \quad (1)$$

where 2.3 is a rounded coefficient within the reported calibration range.

Based on the salinity classification scheme of Richards (1954) [44], soil salinization was categorized as non-saline (ECe < 4 dSm^{-1}), slightly saline (4–8 dSm^{-1}), moderately saline (8–16 dSm^{-1}), strongly saline (16–30 dSm^{-1}), and saline soil (>30 dSm^{-1}).

2.3. Collection and Preprocessing of Unmanned Aerial Vehicle Hyperspectral Data

A DJI Matrice 300 RTK (Figure 5a) quadrotor UAV was employed as the airborne platform, providing high hovering stability and sufficient payload capacity, with a nominal hovering precision of approximately ±0.1 m. The UAV was equipped with an S185 hyperspectral camera (specifications in Table 1), covering the 450–950 nm spectral range with 125 contiguous bands (spectral resolution 8 nm, sampling interval 4 nm). The sensor adopts a snapshot imaging architecture that enables simultaneous exposure across all wavelengths, thereby effectively reducing motion-induced artifacts and minimizing inter-band misregistration (Figure 5b). This configuration is well suited for rapid acquisition and fine-scale monitoring at the farmland scale.



(**a**) (**b**)

Figure 5. (**a**) DJI Matrice 300 RTK and (**b**) S185 sensors.

Table 1. UAV sensor parameters.

Parameters	Specifications
Spectral range	450–950 nm
Spectral resolution	8 nm @ 532 nm
Sampling interval	4 nm
Number of channels	125
Measurement time	0.1–1000 ms
Digital resolution	12 bit
High-resolution imaging speed	5 Cubes/s
Weight	490 g

UAV flights were conducted during 11–13 April 2023 under stable illumination conditions (wind speed < 3 $m \cdot s^{-1}$, cloud cover < 10%). The flight altitude was 100 m, yielding a ground sampling distance of approximately 0.05 m. Data acquisition was restricted to 11:25–14:30, when solar elevation is relatively stable, and each sortie was limited to ≤20 min to mitigate the influence of short-term irradiance variability.

Radiometric reference measurements were acquired before and after each flight. A Labsphere Spectralon calibrated diffuse reflectance panel (nominal reflectance ≈ 99%) was used for empirical line calibration, converting image digital numbers (DN) to surface reflectance and thereby reducing uncertainties associated with sensor response differences and changing observation conditions.

Preprocessing comprised geometric correction, radiometric calibration, and mosaicking, resulting in hyperspectral reflectance products at 0.05 m spatial resolution. To suppress

random noise while preserving diagnostically important absorption features, the reflectance spectra were smoothed using a Savitzky–Golay filter (window length 9 bands, third-order polynomial). The data were retained at the original spectral sampling for subsequent feature construction and inversion modeling.

2.4. Sensitive Feature Extraction via the Optimal Band Combination Algorithm (OBCA)

Hyperspectral reflectance data are characterized by strong inter-band correlations, which often lead to multicollinearity and substantial spectral redundancy. To identify spectral features that are most responsive to soil salinity while avoiding reliance on pre-selected fixed bands, we employed the Optimal Band Combination Algorithm (OBCA) [45–47]. OBCA systematically explores the spectral feature space to identify sensitive band combinations and capture potential band-synergy effects in a data-driven manner.

Let $R(\lambda)$ denote surface reflectance at wavelength λ. For any ordered three-band combination (λ_i, λ_j, λ_k) satisfying $\lambda_i < \lambda_j < \lambda_k$, we define $R_i \equiv R(\lambda_i)$, $R_j \equiv R(\lambda_j)$ and $R_k \equiv R(\lambda_k)$. A family of three-band indices (TBI) was constructed to account for non-linear spectral responses associated with soil surface conditions (e.g., roughness and moisture background). The six TBI forms were selected to span commonly used spectral-index families and their three-band extensions. In remote sensing, ratio-type and normalized-difference-type indices are widely adopted because they enhance spectral contrast related to the target property while partially suppressing multiplicative brightness/illumination effects [48]. Three-band formulations further provide additional degrees of freedom to capture cross-band interactions (band synergy) and improve sensitivity under high-dimensional hyperspectral conditions [49]. Therefore, we adopted six representative forms (modified ratio/normalized-difference and interaction/combination variants) to define a compact yet comprehensive search space for salinity-sensitive triplets. Specifically, six index forms were considered:

$$TBI1 = (R_i - R_j)/R_k \quad (2)$$

$$TBI2 = R_i/(R_j + R_k) \quad (3)$$

$$TBI3 = R_i - 2R_j + R_k \quad (4)$$

$$TBI4 = (R_i - R_j)/(R_j - R_k) \quad (5)$$

$$TBI5 = (R_i - R_j)/(R_i + R_k) \quad (6)$$

$$TBI6 = (R_i - R_j)/(R_i + 2R_j + R_k) \quad (7)$$

For each candidate triplet and index form, the resulting index value x was evaluated against laboratory-measured soil salinity (ECe, denoted as y) using both Pearson's correlation coefficient r and Spearman's rank correlation coefficient ρ. Pearson's r quantifies linear association and is defined as:

$$r = \frac{\sum_{n=1}^{N}(x_n - \bar{x})(y_n - \bar{y})}{\sqrt{\sum_{n=1}^{N}(x_n - \bar{x})^2}\sqrt{\sum_{n=1}^{N}(y_n - \bar{y})^2}} \quad (8)$$

where N is the number of paired samples and $\bar{x}$ and $\bar{y}$ denote sample means. Spearman's ρ assesses monotonic association by applying Pearson correlation to ranked variables and is given by:

$$\rho = 1 - \frac{6\sum_{n=1}^{N} d_n^2}{N(N^2-1)}, \quad (9)$$

$$\rho = r(\mathrm{rank}(x), \mathrm{rank}(y)) \quad (10)$$

where d_n represents the rank difference between x_n and y_n for the n sample. For each index family, candidate triplets were ranked according to the absolute values of r and ρ, and the most informative combinations were retained as sensitive candidates.

To visualize the distribution of correlation strength across the spectral domain, correlation matrices for each index family were summarized and presented as heatmaps. Rather than selecting a "best" triplet from an individual scene—which can be sensitive to scene-specific noise and the extreme-value effect inherent in large combinational searches—we further enhanced robustness through a cross-scene fusion strategy, as described in Section 2.5.

2.5. Cross-Scene Fusion of OBCA Outputs and Construction of a Robust Band Set

To enhance generalization across heterogeneous site conditions and mitigate scene-specific extreme-value effects, OBCA outputs from the global dataset and three individual sites (Site A–C) were jointly fused to construct a robust spectral band set. For each scene s and each three-band index (TBI) family f, candidate triplets were first ranked according to the absolute values of Pearson's $|r|$ and Spearman's $|\rho|$. The top N triplets from each ranking (with $N = 3$ in this study) were retained, and all bands appearing in these triplets were pooled as initial candidates.

To evaluate band-level robustness, each candidate band b was assigned a composite robustness score defined as:

$$score(b) = w_p \cdot freq_P(b) + w_S \cdot freq_S(b) + \alpha \cdot site_support(b) + \beta \cdot feature_support(b) + \gamma \cdot entropy(b) \quad (11)$$

where $freq_P(b)$ and $freq_S(b)$ denote the occurrence counts of band b among the Pearson- and Spearman-based top-N triplets, respectively. The term $site_support(b)$ represents the number of scenes in which band b appears, while $feature_support(b)$ indicates the number of TBI families supporting band b. The entropy term entropy(b) $\in$ [0,1] is a normalized Shannon entropy computed from the distribution of band occurrences across scenes, which favors bands that are consistently selected rather than dominated by a single scene.

To further ensure stability, candidate bands were required to satisfy the constraints:

$$site_support(b) \geq 2,\ feature_support(b) \geq 2 \quad (12)$$

thereby retaining only bands supported by multiple scenes and multiple index families. Because neighboring hyperspectral bands are often highly correlated, redundancy was further reduced using a greedy selection strategy with a minimum band-spacing constraint:

$$\left| b - b' \right| \geq (\Delta + 1), \forall b' \in B \quad (13)$$

where B denotes the final selected band set and $\Delta = 1$, corresponding to a minimum separation of approximately 8 nm given the 4 nm spectral sampling interval of the UAV hyperspectral sensor used in this study. The resulting robust band set was subsequently employed for feature construction and as the spectral input to the proposed S-WGAN-GP-based augmentation and inversion framework.

2.6. Physically Constrained S-WGAN-GP and Teacher Pseudo-Labeling

To alleviate the scarcity of labeled hyperspectral salinity samples while preserving physical plausibility, we propose a two-stage framework consisting of (i) a physically constrained conditional Spectral WGAN with Gradient Penalty (S-WGAN-GP) for synthetic spectrum generation and (ii) a teacher—based pseudo-labeling module with uncertainty-aware screening. The overall objective is to construct a reliable augmented training set that

improves generalization without accumulating errors from low-quality synthetic spectra. Throughout this framework, all evaluations are conducted strictly on real observations, and the conditional variables do not contain any salinity-related information, thereby avoiding label leakage.

2.6.1. Conditional S-WGAN-GP with Spectral Physical Constraints

Let $x \in \mathbb{R}^B$ denote a hyperspectral reflectance spectrum with B spectral bands, and let c denote an auxiliary condition vector encoding contextual information (e.g., spatial coordinates and texture-related features). A conditional S-WGAN-GP is trained to model the conditional spectral distribution p $(x|c)$.

The generator G maps Gaussian noise $z \sim$ N(0, I) and condition c to a synthetic spectrum:

$$\widetilde{x} = G(z, c), \tag{14}$$

The critic D scores paired inputs (x, c) to estimate spectral realism. Following WGAN-GP, the critic is optimized by

$$\mathcal{L}_D = \mathbb{E}_{x\sim P_r}[D(x,c)] - \mathbb{E}_{\widetilde{x}\sim P_g}\left[D\left(\widetilde{x},c\right)\right] + \lambda_{gp}\mathbb{E}_{\widetilde{x}\sim P_z}\left(\|\nabla_{\widetilde{x}} D\left(\widetilde{x},c\right)\|_2 - 1\right)^2 \tag{15}$$

where P_r and P_g denote the real and generated spectral distributions, respectively. The interpolated sample is defined as $\hat{x} = \epsilon x + (1-\epsilon)\widetilde{x}$ with $\epsilon \sim$ U (0,1), and the gradient penalty is computed with respect to $\hat{x}$.

To ensure physical fidelity of the generated spectra, the generator is trained with additional spectral constraints:

$$L_G = -\mathbb{E}\left[D\left(\widetilde{x},c\right)\right] + \lambda_{sam}L_{SAM}\left(\widetilde{x},x_{ref}\right) + \lambda_{tv}L_{TV}\left(\widetilde{x}\right) + \lambda_{rng}L_{range}\left(\widetilde{x}\right) \tag{16}$$

Spectral-angle constraint (degree). Let $X_{ref} = \left\{r_j\right\}_{j=1}^{N_{ref}}$ be a real reference spectral bank (constructed from real labeled spectra, optionally per-site). The spectral angle mapper in degrees is defined as:

$$SAM_{deg}(a,b) = \frac{180}{\pi}\arccos\left(\frac{a^T b}{\|a\|_2\|b\|_2 + \epsilon}\right) \tag{17}$$

where ϵ is a small constant for numerical stability. For each generated spectrum $\widetilde{x}$, we measure its closest physical match in the bank:

$$SAM_{\mathrm{deg}}(\bar{x}; X_{\mathrm{ref}}) = \min_{r\in X_{ref}} SAM_{\mathrm{deg}}(\bar{x}, r) \tag{18}$$

and define

$$\mathcal{L}_{SAM}\left(\widetilde{x}; X_{ref}\right) = SAM_{deg}\left(\widetilde{x}; X_{ref}\right) \tag{19}$$

Equivalently, a robust aggregation such as the mean of the top-k nearest references may be used; the above "min" definition is fully reproducible and consistent with degree-based thresholds.

Smoothness constraint. Spectral smoothness is enforced along the wavelength dimension via first-order total variation:

$$\mathcal{L}_{TV}(\bar{x}) = \frac{1}{B-1}\sum_{b=1}^{B-1}|\bar{x}_{b+1} - \bar{x}_b| \tag{20}$$

Range constraint. Reflectance values are constrained within [0,1] using a soft penalty:

$$\mathcal{L}_{range}(\bar{x}) = \frac{1}{B}\sum_{b=1}^{B}(\max(0, -\bar{x}_b) + \max(0, \bar{x}_b - 1)) \quad (21)$$

It is emphasized that the conditional variables c provide contextual guidance only and do not include the target salinity, ensuring no label leakage. The framework operates purely in the spectral domain and is consistent with the downstream inversion setting.

2.6.2. GAN-Based Guardrails for Synthetic Candidate Selection

After convergence, a large synthetic pool is generated by sampling z and c and computing $\widetilde{x}$ = G (z, c). To remove unrealistic or physically implausible samples prior to pseudo-labeling, deterministic guardrails are applied by jointly enforcing critic realism and spectral fidelity:

$$K\left(\widetilde{x}\right) = \mathbb{I}\left(D\left(\widetilde{x}, c\right) \geq \tau_D\right) \cdot \mathbb{I}\left(SAM\left(\widetilde{x}, x_{ref}\right) \leq \tau_{SAM}\right), \quad (22)$$

The candidate set is defined as

$$S_{cand} = \left\{\widetilde{x} \,\middle|\, K\left(\widetilde{x}\right) = 1\right\} \quad (23)$$

In practice, τ_D and τ_{SAM} are determined using quantile-based criteria on the generated pool (optionally per site), which is robust to scale differences and avoids manual tuning. This step yields a high-confidence synthetic candidate set for subsequent pseudo-labeling.

2.6.3. Teacher Pseudo-Labeling with Uncertainty-Aware Screening

Pseudo-labels are assigned exclusively by a teacher model trained only on real labeled samples. The Teacher integrates a partial least squares regression (PLSR) model and a compact random forest regressor (RF_{small}) to capture both linear spectral–salinity relationships and non-linear interactions. For each synthetic spectrum $\widetilde{x} \in S_{cand}$, the Teacher outputs the pseudo-label:

$$\hat{y}_T = f_T\left(\widetilde{x}\right) \quad (24)$$

Bootstrap uncertainty. To quantify reliability, we perform bootstrap refitting on real labeled data. Specifically, we retrain the Teacher M times using bootstrap resamples, yielding predictions $\left\{\hat{g}^{(m)}(\bar{x})\right\}_{m=1}^{M}$. The predictive uncertainty is defined as:

$$\sigma_T\left(\widetilde{x}\right) = Std\left(\left\{f_T^m\left(\widetilde{x}\right)\right\}_{m=1}^{M}\right) \quad (25)$$

Synthetic samples with excessive uncertainty are removed:

$$K_u\left(\widetilde{x}\right) = \mathbb{I}\left(\sigma_T\left(\widetilde{x}\right) \leq \tau_\sigma\right),\ S_{syn} = \left\{\widetilde{x}, \hat{y}_T \,\middle|\, K_u\left(\widetilde{x}\right) = 1\right\} \quad (26)$$

Consistency and credibility score. In addition to σ_T, we compute a consistency indicator between PLSR and RF_{smal}:

$$\text{diff}\left(\widetilde{x}\right) = \left|\hat{y}_{PLS}\left(\widetilde{x}\right) - \hat{y}_{RF}\left(\widetilde{x}\right)\right| \quad (27)$$

We further derive a credibility score conf $(\widetilde{x}) \in [0,1]$ that favors samples with small diff and low uncertainty. Let Q_p denote the p-th quantile computed on the candidate pool (optionally per site). We define normalized diagnostics:

$$\widetilde{d}(\bar{x}) = \min\left(\frac{diff(\widetilde{x})}{Q_p(diff)}, 1\right) \tag{28}$$

$$\widetilde{s}(\bar{x}) = \min\left(\frac{\sigma_T(\widetilde{x})}{Q_p(\sigma_T)}, 1\right) \tag{29}$$

and the credibility score:

$$conf(\widetilde{x}) = 1 - (w_{diff}\widetilde{d}(\widetilde{x}) + w_{std}\widetilde{s}(\widetilde{x})) \tag{30}$$

$$\text{with } w_{diff} + w_{std} = 1 \tag{31}$$

This definition is scale-free, bounded in [0,1], and directly matches the intended behavior (high confidence when both disagreement and uncertainty are low).

Site-adaptive acceptance envelope. Finally, we apply a per-site acceptance envelope using quantile-based thresholds on $\{\mathrm{SAM}_{\mathrm{deg}}, D\left(\widetilde{x}, c\right), conf\left(\widetilde{x}\right), \sigma_{\mathrm{T}}\}$ computed over the candidate pool. For a given site s, the final keep rule is:

$$\bar{x}kept \Longleftrightarrow D(\widetilde{x}, c) \geq \tau_D^{(s)} \wedge SAM_{deg}(\widetilde{x}; X_{ref}^{(s)}) \leq \tau_{SAM}^{(s)} \wedge conf(\widetilde{x}) \geq \tau_{conf}^{(s)} \Lambda \sigma_T(\widetilde{x}) \leq \tau_\sigma^s \tag{32}$$

The resulting filtered synthetic set is merged with real samples for downstream salinity inversion modeling, while all reported evaluations are performed strictly on real observations.

2.7. Salinity Inversion Modeling

Soil salinity inversion is formulated as a supervised regression task that learns a mapping from hyperspectral spectral features to field-measured salinity:

$$\hat{y} = f(x),\ x \in \mathbb{R}^d \tag{33}$$

where x denotes the input feature vector and y is the predicted soil salinity (ECe). In this study, x is composed of spectral features only, i.e., OBCA-selected reflectance bands (or band-derived spectral features). This design preserves physical interpretability and reduces the risk of shortcut learning caused by site-specific spatial correlations or auxiliary cues. Notably, while auxiliary variables (e.g., coordinates or texture descriptors) may be used as conditioning guidance in the generative stage, all inversion models in Section 2.7 are trained and evaluated strictly on spectral inputs to ensure a fair and physically consistent comparison.

To ensure robustness and comparability, we implement multiple representative regression models under a unified preprocessing and validation protocol.

(1) Partial Least Squares Regression (PLSR)

PLSR projects x into a low-dimensional latent subspace that maximizes the covariance with y, and then fits a linear model in the latent space:

$$t = Xw, \hat{y} = tb \tag{34}$$

where t is the latent score vector, w is the projection direction, and b is the regression coefficient.

(2) Support Vector Regression with RBF kernel (SVR-RBF)

SVR estimates a maximum-margin regression function in a kernel-induced feature space. Using the radial basis function (RBF) kernel,

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2\right) \tag{35}$$

SVR optimizes an ϵ-insensitive loss with regularization to balance fitting accuracy and generalization.

(3) Random Forest (RF)

RF is an ensemble of decision trees trained on bootstrap resamples with random feature subspace selection. The final prediction is obtained by averaging predictions across trees:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} f_t(x), \tag{36}$$

where T is the number of trees and f_t denotes the t-th tree regressor.

(4) Extreme Gradient Boosting (XGBoost, XGB)

XGBoost is a gradient-boosted decision tree method that builds an additive model by sequentially fitting regression trees to the residuals of prior trees. The prediction is expressed as

$$\hat{y} = \sum_{k=1}^{K} f_k(x), f_k \in \mathcal{F}, \tag{37}$$

where K is the number of trees and $\mathcal{F}$ denotes the space of regression trees. Training minimizes a regularized objective:

$$\mathcal{L} = \sum_i \updownarrow(y_i, \hat{y}_i) + \sum_k \Omega(f_k), \tag{38}$$

in which $\updownarrow$ is a differentiable loss (e.g., squared error) and Ω penalizes model complexity to improve generalization.

All models are fitted on a consistent feature space. When required by the learning algorithm (e.g., SVR), features are standardized using statistics computed from the training data within each validation split. Tree-based models (RF and XGB) are trained on the same spectral inputs and do not require strict feature scaling, but they are kept within the same preprocessing pipeline for consistency.

(5) Cross-validation protocol

Model performance is evaluated using leave-one-out cross-validation (LOOCV) on the real labeled samples. Given N real samples, LOOCV conducts N runs: in the i-th run, one sample is held out for testing and the remaining $N-1$ samples are used for training. Predictions are aggregated over all runs to compute overall metrics:

$$\hat{y}_i = f^{(-i)}(x_i),\ i = 1, \ldots, N, \tag{39}$$

where $f^{(-i)}$ denotes the model trained with the i-th sample excluded. This protocol maximizes training data usage per fold and provides an unbiased estimate of generalization performance under limited ground-truth availability.

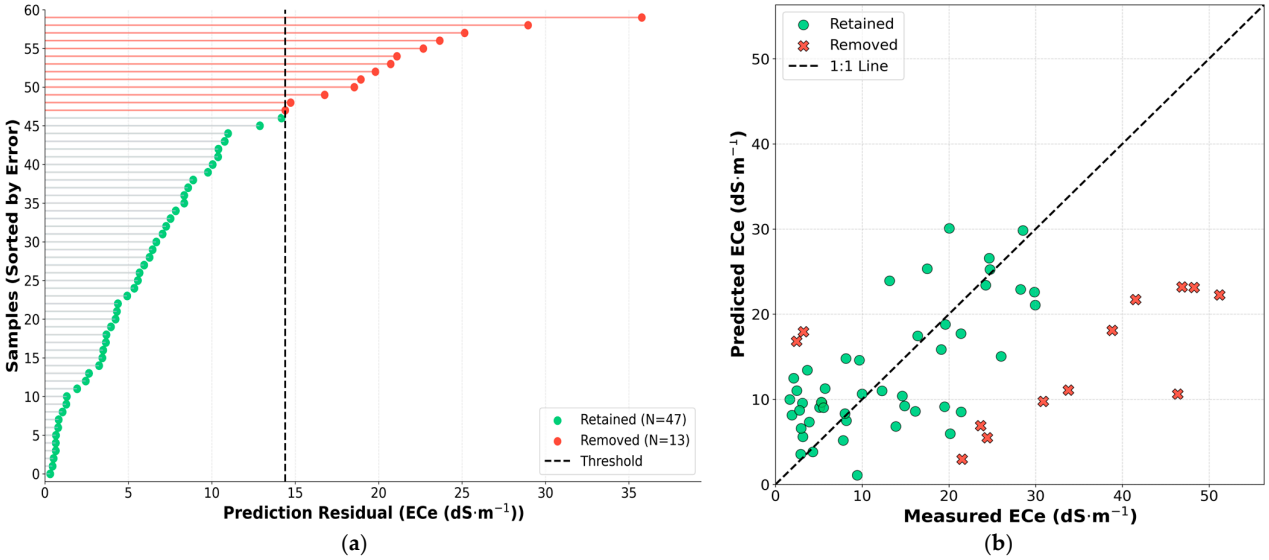
3. Results

3.1. SVR-Residual Screening Yields a High-Consistency Real Reference Set

3.1.1. Data Preprocessing and Outlier Removal Framework

Given the inherent complexity of field environments and the high sensitivity of hyperspectral sensors, the raw dataset inevitably contains a small proportion of outliers that do not conform to the general spectral response patterns. These outliers may arise from random environmental variability rather than systematic errors. To mitigate their potential influence on model training, this study employed an SVR-based Residual Analysis Framework to identify and filter out these inconsistent samples, ensuring that the model focuses on learning the dominant spectral features of soil salinity.

As illustrated in Figure 6a, a Support Vector Regression (SVR) model with 5-fold cross-validation was employed to predict the electrical conductivity (ECe) of all 60 original samples. The absolute difference between the measured and predicted ECe was defined as the prediction residual. Samples with excessive residuals indicate a significant "spectral-chemical inconsistency", meaning their spectral signatures do not align with the general physical laws governing the dataset.



(**a**) (**b**)

Figure 6. (**a**) Residual ranking plot from the 5-fold cross-validated SVR model. Samples are sorted by absolute prediction residual $|ECe_{pred} - ECe_{obs}|$; the vertical dashed line indicates the residual threshold (≈14.5 dSm^{-1}) used to flag outliers. Green points denote retained samples (N = 47), and red points denote removed samples (N = 13). (**b**) Measured–predicted consistency plot for the same SVR-CV results. The dashed line represents the 1:1 reference; retained samples cluster closer to the 1:1 line, whereas removed samples show larger deviations, indicating spectral–chemical inconsistency and motivating their exclusion.

To avoid subjective threshold selection, the residuals were ranked in ascending order and a data-driven knee-point detection approach (maximum curvature criterion, Kneedle-type principle) was applied to identify the turning point of the residual distribution. The algorithmically determined knee occurs at approximately 14.5 dSm^{-1}, beyond which residuals increase sharply and increasingly reflect atypical observations.

Consequently, the 13 samples (21.7%) exceeding this threshold were identified as outliers and removed. Figure 6b further visualizes the consistency check. The removed samples (red markers) are predominantly located far from the 1:1 reference line. Notably, several outliers appear in the high-salinity range (>30 dSm^{-1}), suggesting that extremely saline soils may exhibit spectral saturation or non-linear response patterns that differ from

the general population. The retained 47 samples (green markers) demonstrate strong alignment with along the 1:1 line, confirming the validity of the cleaning process.

3.1.2. Statistical Characteristics of the High-Quality Dataset

After applying the cleaning framework, the statistical distribution of the retained dataset (N = 47) was analyzed to ensure its representativeness. Figure 7 presents the "Raincloud Plot" of soil salinity across the three sampling sites (A, B, and C), combining probability density (violin), statistical intervals (box), and raw data distribution (scatter).

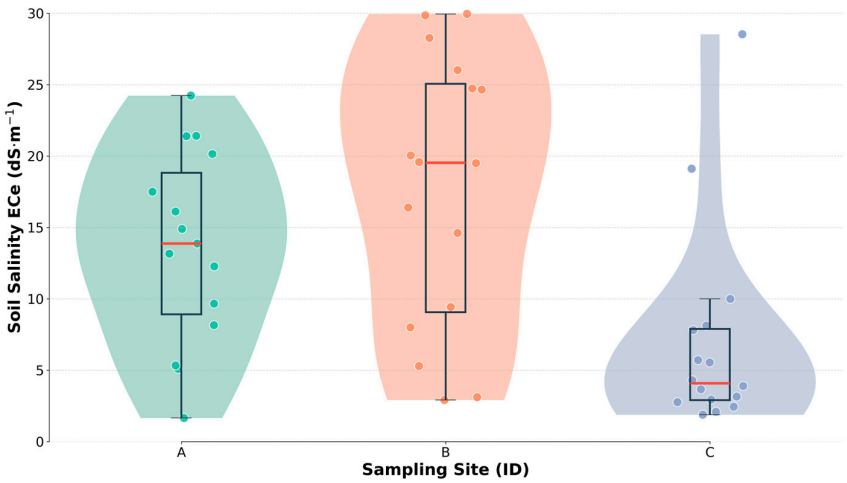


Figure 7. Distribution of cleaned soil salinity (ECe) samples across the three sampling sites. The colored shaded areas represent the probability density distribution of the data (violin plots), and the correspondingly colored circles represent individual sample measurements (green for Site A, orange for Site B, and blue for Site C). The internal box plots indicate the median (red horizontal line), interquartile range (box), and overall data range (whiskers) for each site.

Site A exhibits the widest range of salinity with a broad distribution shape, indicating high spatial variability within the site.

Site B shows a relatively uniform distribution with a higher median value (approx. 20 dSm^{-1}), representing the highly saline zone.

Site C is concentrated in the lower salinity range with a compact distribution density, representing the low-salinity zone.

Table 2 provides a quantitative comparison before and after data cleaning. As previously analyzed, the original dataset (N = 60) contained high variability. However, the removal of outliers effectively optimized the data structure: the standard deviation was reduced from 13.35 to 8.97, and the coefficient of variation (CV) decreased from 79.12% to 72.43%. This reduction in dispersion signifies an enhanced signal-to-noise ratio while preserving a sufficient gradient (Min: 1.65, Max: 29.95 dSm^{-1}) for subsequent modeling.

Table 2. Descriptive statistics of soil salinity (ECe) before and after outlier removal.

Dataset	N	Min	Max	Mean	Std	CV(%)
Original	60	1.65	51.22	16.87	13.35	79.12
Cleaned	47	1.65	29.95	12.39	8.97	72.43

3.2. OBCA Reveals Salinity-Sensitive Band Synergies and a Cross-Site Robust NIR-Dominant Feature Set

3.2.1. Global Spectral Response Analysis Based on OBCA

To transcend the information limitations inherent in single-band analysis, six distinct forms of three-band salinity indices (TBI1-TBI6) were constructed. By traversing the

entire spectral band space, both Pearson (r) and Spearman (ρ) correlation coefficients were calculated to capture linear and non-linear relationships simultaneously.

Tables 3 and 4 summarize the optimal combinations and their corresponding correlation coefficients derived from the global dataset. The results demonstrate that the OBCA-optimized three-band combinations significantly enhanced the spectral signal of soil salinity. Notably, TBI1, formulated as $(Ri - Rj)/R_k$, exhibited superior performance, achieving the highest Spearman correlation of 0.753 and a Pearson correlation of 0.717. It is worth noting that for most indices (TBI1, TBI2, TBI5, TBI6), the Spearman coefficient is consistently higher than the Pearson coefficient (e.g., TBI1: 0.753 > 0.717), indicating that the relationship between spectral features and soil salinity is predominately non-linear.

Table 3. Optimal band combinations and correlation coefficients ($|r|$) for the six constructed spectral indices (TBI1–TBI6) based on the global dataset.

Feature	Combination	PearsonAbs
TBI1	R_{30}, R_{38}, R_{124}	0.717387
TBI2	R_{29}, R_{31}, R_{38}	0.668571
TBI3	R_4, R_{26}, R_{38}	0.558812
TBI4	R_6, R_{24}, R_{74}	0.613876
TBI5	R_{29}, R_{38}, R_{125}	0.705115
TBI6	R_{29}, R_{38}, R_{125}	0.688971

Table 4. Optimal band combinations and correlation coefficients ($|\rho|$) for the six constructed spectral indices (TBI1–TBI6) based on the global dataset.

Feature	Combination	SpearmanAbs
TBI1	R_{30}, R_{38}, R_{124}	0.75266
TBI2	R_{30}, R_{31}, R_{38}	0.702012
TBI3	R_5, R_{25}, R_{37}	0.569843
TBI4	R_6, R_{23}, R_{76}	0.600729
TBI5	R_{29}, R_{38}, R_{125}	0.739015
TBI6	R_{29}, R_{38}, R_{125}	0.707216

In terms of spectral mechanisms, the optimal combinations were not randomly distributed but showed clear patterns. Taking the best-performing TBI1 as an example, the selected bands were R_{30} (567 nm), R_{38} (599 nm), and R_{124} (946 nm). This suggests that the interaction between the visible green-yellow region (sensitive to soil color and vegetation/algae stress) and the Near-Infrared (NIR) plateau (sensitive to soil moisture and salt crust structure) effectively suppresses background noise while amplifying salinity-related spectral features.

To visualize the distribution of sensitive bands within the spectral space, Figure 8 depicts the 3D Correlation Cubes for the TBI features.

Orthogonal slice visualization reveals a distinct "Spectral Clustering Phenomenon". As shown in the cubes for TBI1, TBI5, and TBI6, high-correlation regions (indicated by deep red cores) are highly concentrated. Specifically, the axes of these hotspots align consistently around 560–600 nm and 940–950 nm. This clustering indicates that, despite mathematical variations among different TBI forms, they consistently capture intrinsic spectral response mechanisms. The high similarity between the optimal bands of TBI1 (R_{30}, R_{38}, R_{124}) and TBI5 (R_{29}, R_{38}, R_{125}) further confirms the robustness of these specific spectral regions in characterizing soil salinity.

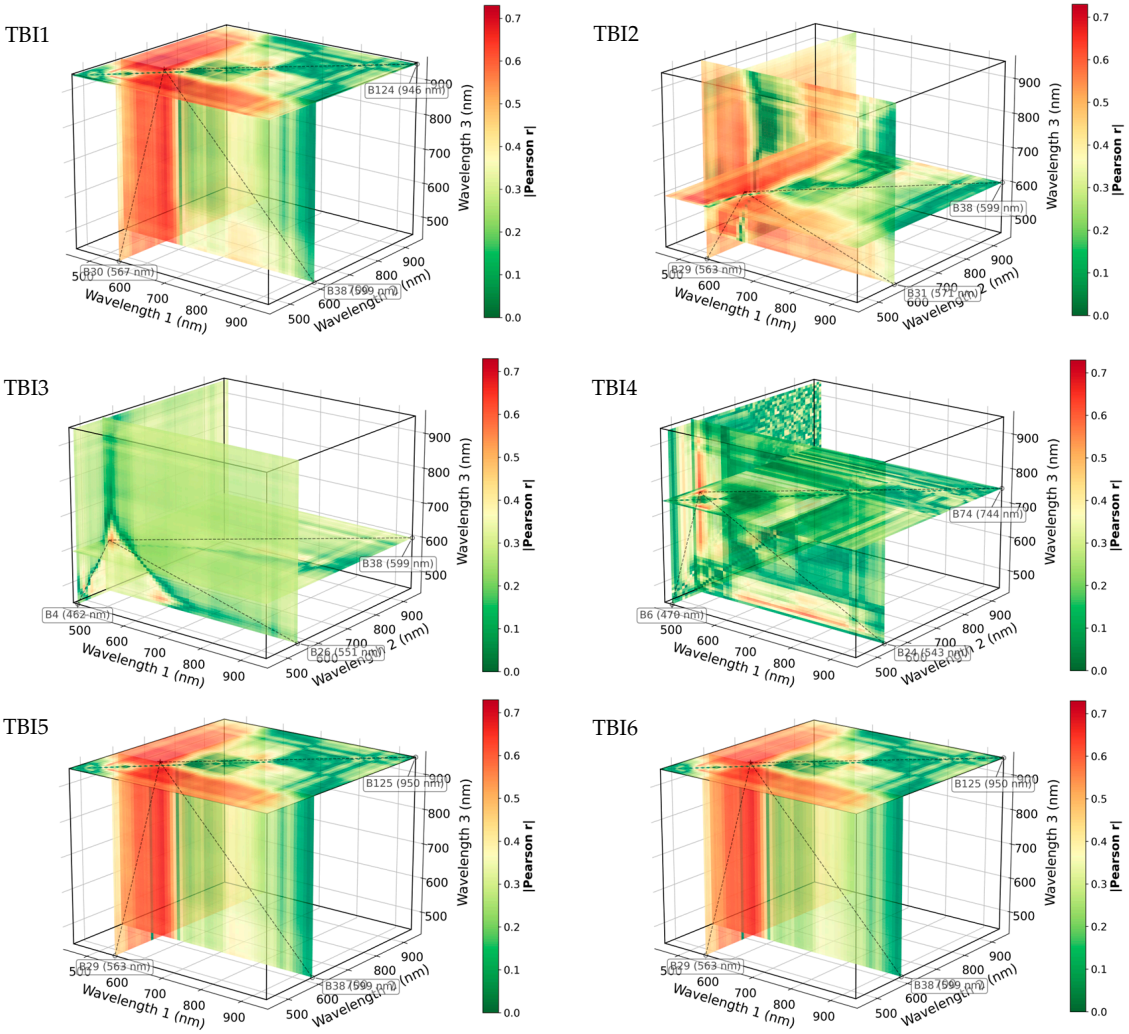


Figure 8. Slice display of the correlation between soil spectral characteristics and soil salinity based on the Optimal Band Combination Analysis (OBCA) algorithm.

3.2.2. Cross-Scenario Robustness Evaluation and Final Feature Subset Determination

Relying exclusively on global correlations may predispose models to overfitting specific site conditions. To isolate features that maintain stability across spatially heterogeneous environments, a Cross-Scenario Robustness Evaluation Framework was implemented across three independent sampling sites (Site A, B, and C).

Figure 9 presents the Cross-Site Consistency Heatmap, visualizing the selection frequency of candidate bands across different environments. The analysis identified two distinct spectral response patterns:

Global Consistency: Specific bands in the NIR region, such as R_{125} (950 nm), demonstrated consistent selection across all three sites (Site A: 1, Site B: 3, Site C: 4). This indicates that the far-NIR region captures fundamental spectral absorption features of soil salinity—likely related to lattice water in salt crystals—that remain invariant to local environmental changes.

Local Dominance: Conversely, other bands showed dominant support within specific contexts. Most notably, R_{109} (885 nm) exhibited the highest singular frequency, being selected 5 times in Site A (high variability zone). Similarly, R_{102} (857 nm) and R_{25} (547 nm) proved critical for distinguishing salinity gradients in Site B and Site C, respectively.

To quantify the comprehensive quality of each band, a composite Robustness Score was calculated, integrating correlation magnitude, cross-site recurrence, and feature diversity. Figure 10 displays the ranking of the final selected bands based on this score.

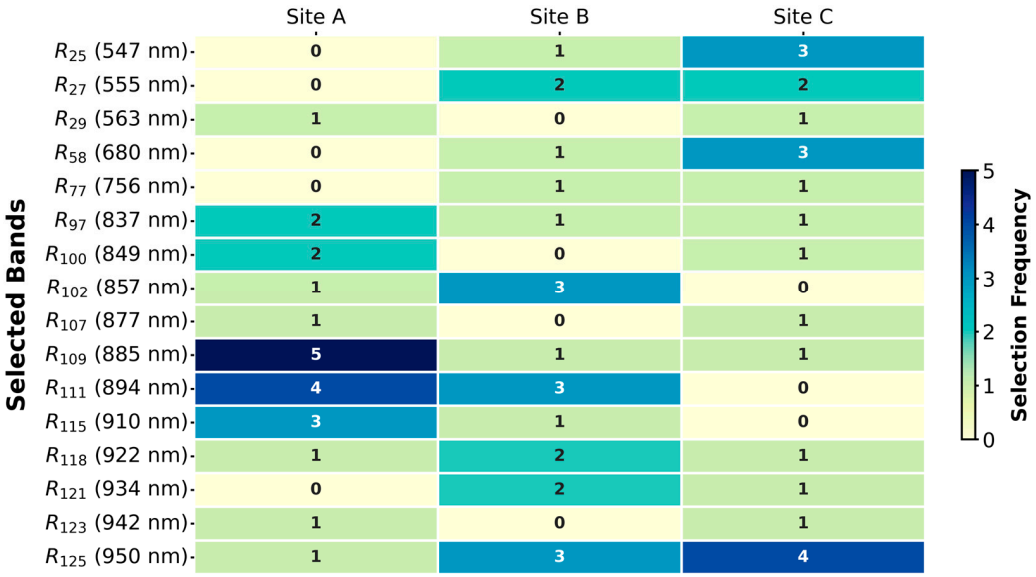


Figure 9. Site-wise selection frequency of the robust OBCA band set.

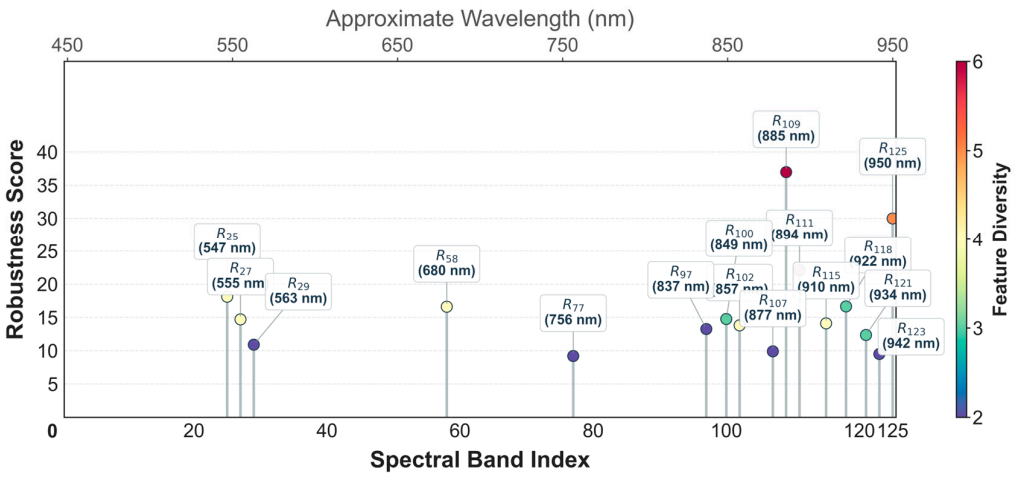


Figure 10. Robustness Score Profile of Selected Bands.

The results indicate that the optimal features are clustered in the Near-Infrared (NIR) region. As shown in the lollipop chart, R_{109} (885 nm) achieved the highest robustness score (represented by the tallest marker), followed closely by R_{125} (950 nm). This confirms that NIR bands possess superior penetration and stability compared to visible bands when facing complex surface heterogeneity. Based on this robustness ranking and applying a minimum spectral interval constraint (to mitigate collinearity redundancy), a final subset of salinity-sensitive bands was determined. These bands represent the "Spectral Consensus" of the study area and serve as the input variables for subsequent machine learning inversion models (Table 5).

Table 5. The final subset of 16 robust salinity-sensitive bands selected based on cross-scenario evaluation.

Band ID	Center Wavelength (nm)	Spectral Region	Robustness Score	Primary Feature Support
R_{25}	547	Visible (Green)	18	TBI2, TBI3
R_{27}	555	Visible (Green)	15	TBI1, TBI5
R_{29}	563	Visible (Green)	11	TBI2, TBI5, TBI6
R_{58}	680	Red Edge	16.5	TBI4
R_{77}	756	NIR	9	TBI2
R_{97}	837	NIR	13	TBI3, TBI4
R_{100}	849	NIR	14	TBI2
R_{102}	857	NIR	14.8	TBI1, TBI4
R_{107}	877	NIR	10	TBI3
R_{109}	885	NIR	37.5	All (High Diversity)
R_{111}	894	NIR	25.5	TBI1, TBI3, TBI5
R_{115}	910	NIR	22	TBI1, TBI6
R_{118}	922	NIR	23	TBI2, TBI5
R_{121}	934	NIR	12	TBI4
R_{123}	942	NIR	9.5	TBI2, TBI6
R_{125}	950	NIR	30	All (High Diversity)

3.3. S-WGAN-GP Produces Spectrally Faithful, Manifold-Aligned, High-Confidence Pseudo-Labeled Samples

3.3.1. Training Stability and Spectral Fidelity Analysis

The training stability of a Wasserstein GAN is a key indicator of whether the model has learned the data distribution without mode collapse. Figure 11a reports the adversarial training dynamics of the Critic (D) and Generator (G), while Figure 11b tracks the Spectral Angle Mapper (SAM) between generated spectra and the reference bank during training. Figure 12 further compares the mean reflectance curves of real and generated spectra to assess spectral fidelity at the waveform level.

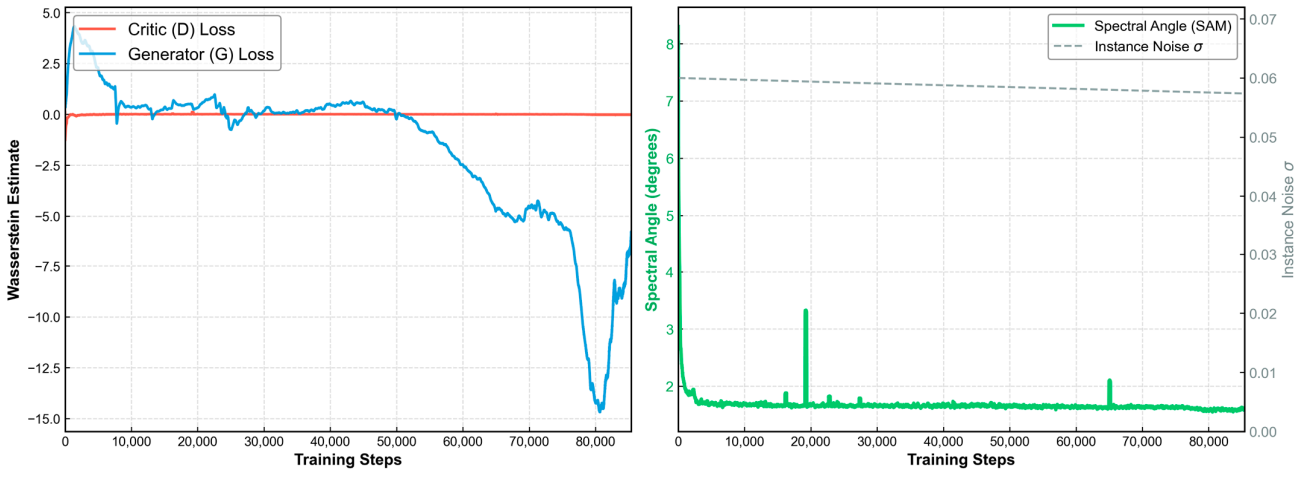


Figure 11. Training dynamics and spectral fidelity evaluation of the S-WGAN-GP model. (**a**) Evolution of the Generator (G) and Critic (D) losses and the Spectral Angle Mapper (SAM) during training steps; (**b**) Comparison of the mean spectral reflectance between real observed samples (solid blue line) and GAN-generated synthetic samples (dashed red line).

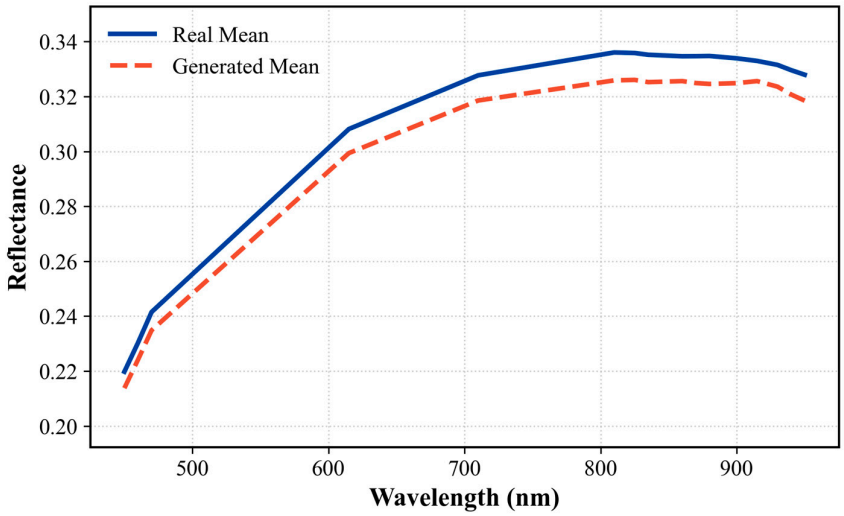


Figure 12. Comparison of the mean spectral reflectance between real observed samples (solid blue line) and GAN-generated synthetic samples (dashed red line).

As shown in Figure 11a, the Critic loss rapidly stabilizes and remains close to zero over most of the training steps, suggesting that the gradient penalty effectively enforces the Lipschitz constraint and prevents critic divergence. In contrast, the Generator loss exhibits the typical non-monotonic behavior of WGAN-GP optimization: after an initial adjustment stage, it oscillates within a relatively stable range, and a pronounced re-balancing event is observed in the late phase (around 8×10^4 steps), where the generator loss becomes substantially more negative and then rebounds. This late-stage transition is consistent with an adversarial refinement process, where the generator adapts to a sharper decision boundary imposed by the critic and further improves the match to the target spectral manifold, rather than collapsing to trivial outputs.

From a physical consistency perspective, Figure 11b shows that SAM drops sharply at the beginning of training and then maintains a low baseline (approximately 1.6–1.8°) for the majority of the optimization trajectory. Only a few short-lived spikes occur, followed by immediate recovery to the low-level regime, indicating that occasional aggressive updates do not cause persistent degradation in spectral shape. Notably, SAM is an angle-based similarity measure and, when used on calibrated reflectance spectra, is relatively insensitive to illumination and albedo (brightness) effects; thus it primarily evaluates spectral shape consistency rather than absolute magnitude matching. Meanwhile, the instance noise parameter (σ) decays gradually over training, which is known to smooth the critic's decision surface and improve adversarial stability; the coexistence of a stable low SAM regime supports the robustness of the learned spectral generator.

The mean-spectrum comparison in Figure 12 further confirms spectral fidelity. The generated mean reflectance curve closely follows the real mean across 450–950 nm, preserving the overall trend from the visible to near-infrared range and maintaining consistent slope transitions. A mild systematic amplitude bias is observed, with the generated mean being slightly lower than the real mean (i.e., a small underestimation of albedo). This mild magnitude offset is plausible for UAV hyperspectral reflectance over bare soil because reflectance magnitude is sensitive to residual radiometric/illumination variability (e.g., temporal irradiance changes during flight) and to reflectance anisotropy (BRDF effects) driven by sun–sensor geometry and micro-scale roughness/shadowing, whereas spectral shape can remain comparatively stable. Importantly, this discrepancy mainly manifests as an amplitude shift rather than a distortion of spectral morphology, implying that inter-band relationships and spectral-shape characteristics are well preserved. Such morphology-level

agreement is critical for hyperspectral salinity inversion because downstream features (e.g., OBCA-selected band vectors and shape-sensitive indices) depend strongly on relative spectral structure. Therefore, the results in Figures 11 and 12 collectively suggest that the proposed S-WGAN-GP produces physically plausible spectra with stable training behavior and high spectral-shape fidelity, providing a reliable basis for subsequent pseudo-labeling and regression modeling.

3.3.2. Data Manifold Alignment and Feature Space Expansion

Beyond waveform-level spectral fidelity, a generative model is practically valuable in semi-supervised learning only if it can improve coverage of the data manifold implied by abundant unlabeled observations. Figure 13 visualizes the distributions of real labeled samples, unlabeled samples, and GAN-generated synthetic samples in a two-dimensional embedding space obtained with t-distributed stochastic neighbor embedding (t-SNE), which preserves local neighborhood relationships and is commonly used for qualitative inspection of high-dimensional data geometry.

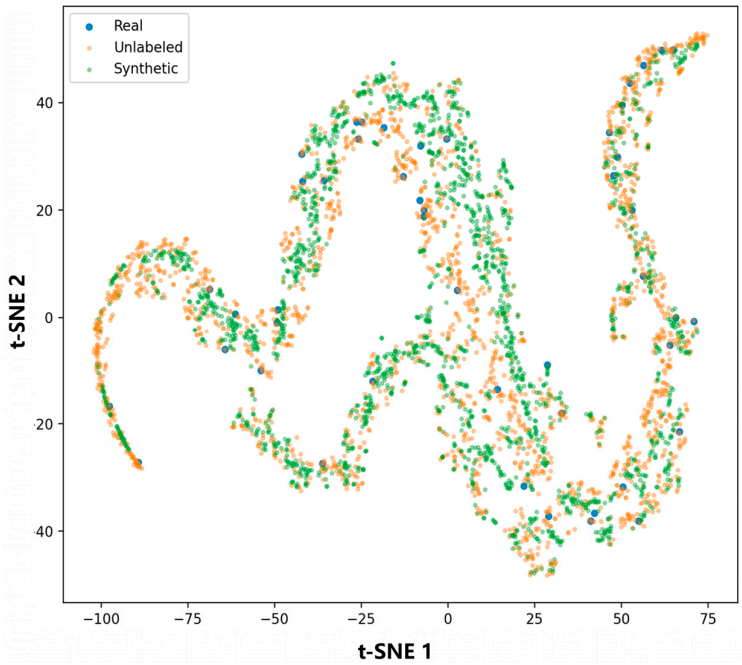


Figure 13. Visualization of the data manifold alignment in the 2D feature space using t-Distributed Stochastic Neighbor Embedding (t-SNE). The plot illustrates the distribution coverage of real labeled samples (blue), unlabeled samples (orange), and the generated synthetic samples (green), highlighting the effective bridging of the feature gap. Two-dimensional t-SNE embedding of real, unlabeled, and synthetic spectra. The x–y axes represent embedding coordinates (dimensionless); only relative distances/neighborhood structure are meaningful.

In Figure 13, the unlabeled set (orange) forms a highly non-linear, elongated manifold with multiple curved branches, indicating substantial variability in spectral patterns across the scene. The real labeled set (blue) is extremely sparse relative to this structure and occupies only a few localized regions, suggesting limited coverage of the spectral variability that the downstream regressor will encounter. By contrast, the synthetic set (green) is largely co-located with the unlabeled manifold and follows its curved topology rather than appearing as isolated clusters or uniformly scattered points. This spatial co-occurrence implies that the generator does not simply replicate a few labeled examples; instead, it produces samples that are consistent with the broader unlabeled distribution in the embedding space.

Importantly, the synthetic samples densify regions where labeled points are absent, thereby reducing the coverage gap between the sparse labeled set and the unlabeled manifold. Such coverage expansion is beneficial for regression in low-label regimes because it provides additional training support around the unlabeled structures, which can help the model learn smoother decision functions and reduce extrapolation risk when predicting on previously unseen spectral patterns. We emphasize that t-SNE provides a qualitative diagnostic rather than a quantitative divergence metric; therefore, the key takeaway from Figure 13 is the distributional overlap and manifold-following behavior of synthetic samples relative to the unlabeled data, supporting their potential utility for subsequent pseudo-labeling and downstream inversion.

3.3.3. Reliability of Pseudo-Labels and Uncertainty Quantification

Because GAN-generated spectra are unlabeled, their practical value in the proposed framework depends on whether the Teacher can assign reliable pseudo-labels while preventing noise accumulation. We therefore adopted a staged screening protocol that couples GAN-based physical guardrails with Teacher credibility and uncertainty diagnostics.

We first generated a synthetic pool of N = 13,999 spectra. A global realism gate based on the critic score was applied to remove low-realism samples, retaining 13,719 spectra with Dscore $\geq$ 4.0719. Here, Dscore denotes the WGAN-GP critic output D(x), which is an unbounded real-valued score (not a probability); higher scores indicate that a sample is more consistent with the real-data distribution under the learned 1-Lipschitz critic. To determine the cutoff in a reproducible, data-driven manner, we computed critic scores on the real labeled spectra after GAN training and set the acceptance threshold as a conservative lower-tail quantile of the real-score distribution (i.e., requiring synthetic samples to achieve critic scores comparable to typical real samples). This procedure yields Dscore $\geq$ 4.0719 for the present dataset. Next, to enforce spectral fidelity, we performed a site-wise selection by keeping the top K = 3000 candidates per site with the smallest spectral angle to the site-specific reference bank (SAM_deg), resulting in 9000 candidates in total. The Teacher then assigned pseudo-labels $\hat{S}$ and simultaneously produced credibility indicators, including an uncertainty estimate S_std (bootstrap-based predictive standard deviation) and a consistency term diff = |S_{PLS} − S_{RF}|. Finally, synthetic samples were retained only if they simultaneously satisfied a site-adaptive acceptance envelope: (i) sufficiently low spectral mismatch (SAM_deg below an upper bound), (ii) sufficiently high realism (Dscore above a lower bound), (iii) sufficient prediction credibility (conf above a lower bound), and (iv) bounded uncertainty (S_std below an upper bound). This multi-constraint screening produced 1635 high-confidence pseudo-labeled samples before thinning. A quota-controlled thinning step was applied to reduce redundancy while preserving a balanced site-wise representation, resulting in N = 1186 final synthetic samples (SiteA/B/C: $N_{syn,final}$ = 472/432/282 corresponding $N_{real,original}$ = 15/16/16). The resulting site-wise acceptance bounds are summarized in Table 6.

Table 6. Site-wise quality envelope of the final retained synthetic samples after stage filtering.

Site	N_Real_Original	N_Syn_Final	SAM_Deg (Max)	Dscore (Min)	Conf (Min)	S_Std (Max)
A	15	472	$\leq$0.556	$\geq$4.226	$\geq$0.472	$\leq$10.479
B	16	432	$\leq$0.895	$\geq$4.181	$\geq$0.474	$\leq$10.331
C	16	282	$\leq$0.738	$\geq$4.381	$\geq$0.546	$\leq$7.35

The pseudo-label distribution (Figure 14a) shows a structured pattern with a global mean of 12.18 dSm^{-1}, mainly concentrated in the moderate range (~10–14 dSm^{-1}) while still covering the high-salinity tail, indicating that the retained synthetic set provides a

broad yet controlled supervisory signal rather than being dominated by extremes. Teacher diagnostics further support reliability: confidence scores are centered at 0.56 (Figure 14b), and the uncertainty distribution is largely unimodal with mean S_std = 6.67 (Figure 14c), suggesting stable pseudo-labeling behavior and effective suppression of noisy supervision. Spatial maps of pseudo-labeled salinity (Figure 15) offer an additional plausibility check, exhibiting coherent patterns consistent with site conditions (patchy heterogeneity in Site A, extensive high salinity in Site B, and a low-salinity baseline with localized accumulations in Site C). Table 6 summarizes the site-adaptive acceptance envelope used to retain reliable pseudo-labeled spectra, where SAM_deg(max) constrains spectral-shape deviation, Dscore(min) enforces adversarial realism, conf(min) requires minimum Teacher credibility, and S_std(max) bounds predictive uncertainty; together, these thresholds define the per-site quality region for physically plausible and label-consistent synthetic samples.

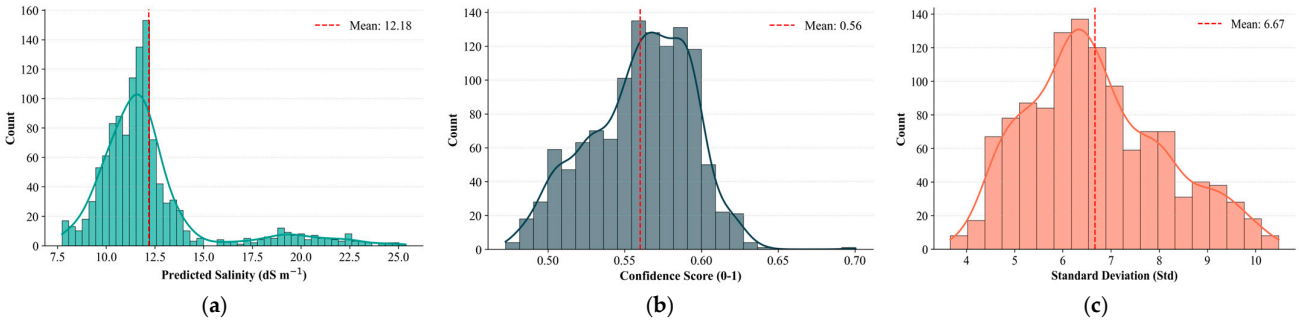


Figure 14. (**a**) Frequency distribution of the predicted soil salinity (ECe) for the generated samples; (**b**) Distribution of the prediction confidence scores from the Teacher framework; (**c**) Distribution of the prediction uncertainty (standard deviation), indicating the stability of the pseudo-labels.

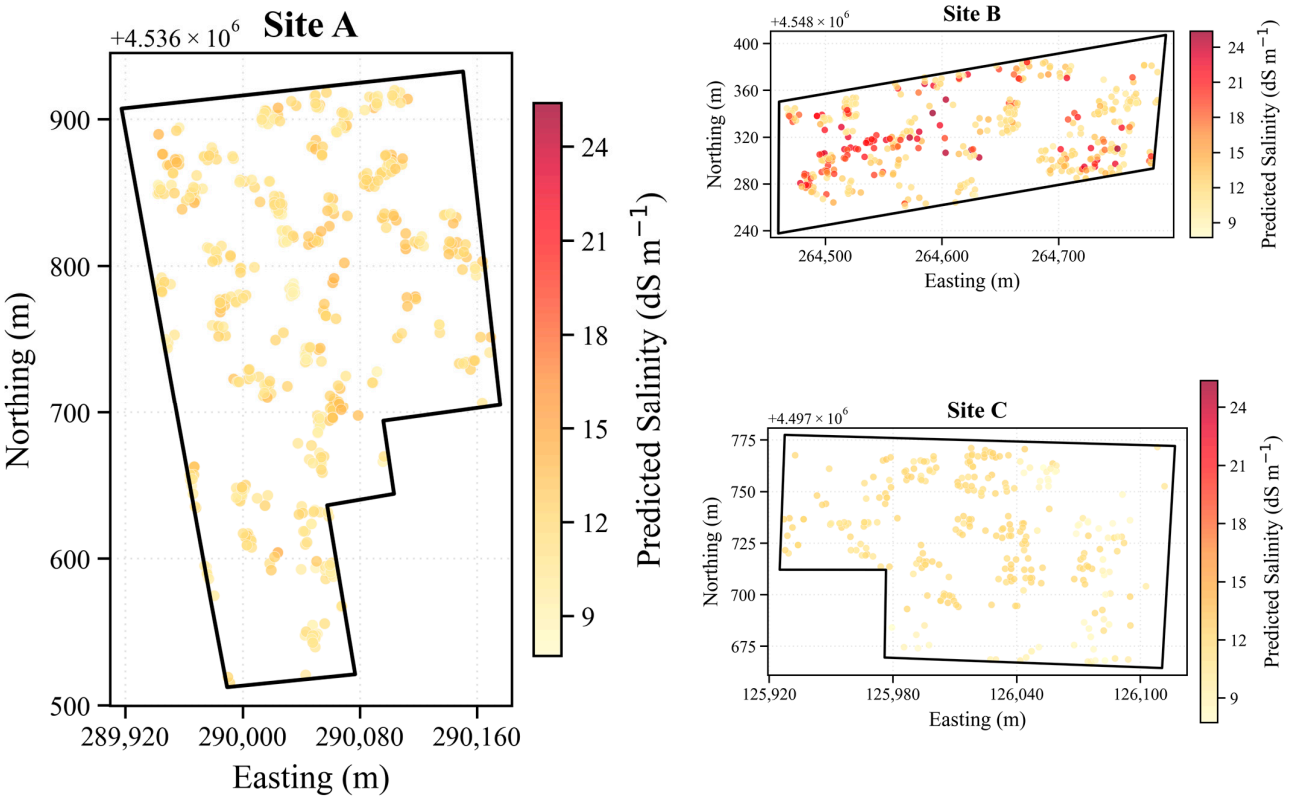


Figure 15. Spatial distribution maps of predicted soil salinity (ECe) across the three experimental areas (Site A, Site B, and Site C) derived from the final semi-supervised inversion model. The solid black lines delineate the boundaries of the respective sampling sites.

3.4. Mapping Augmented Training Improves LOOCV Inversion and Enables Spatially Coherent Salinity Mapping

3.4.1. Evaluation Strategies and Model Training

To rigorously quantify the contribution of GAN-generated data, all models were evaluated using leave-one-out cross-validation (LOOCV) on the real labeled set only, which prevents data leakage and is suitable for small-sample assessment. Two training settings were compared:

(i) Real-only (Baseline): models trained on the original labeled set (N = 47).

(ii) Real + Filtered Synthetic (Proposed): models trained on the same real set plus high-confidence synthetic samples generated by S-WGAN-GP and retained by the staged reliability screening (total N = 1233).

3.4.2. Performance Comparison and Accuracy Analysis

Table 7 and Figure 16 summarize LOOCV results for four regressors (SVR, RF, XGB, PLSR). Under the Real-only baseline, all models show limited generalization due to the sparse training set (all $R^2 < 0.40$, RPD = 1.15–1.28), indicating insufficient ability to capture the non-linear mapping from hyperspectral features to soil salinity.

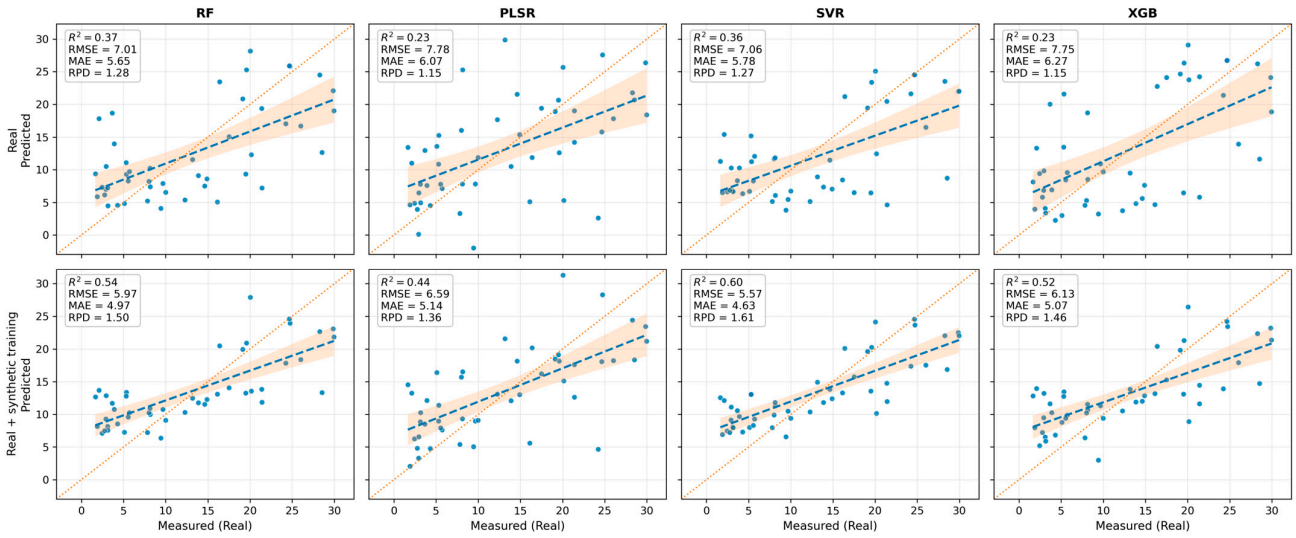


Figure 16. Measured versus predicted soil salinity (ECe) for four regression models (RF, PLSR, SVR, and XGB) evaluated on real observations using LOOCV. The top row shows the supervised baseline trained with real samples only, whereas the bottom row shows models trained with the combined real + synthetic dataset. Each blue dot represents a held-out real sample; the orange dotted line denotes the 1:1 reference line, the blue dashed line indicates the linear fitted trend, and the light orange shaded area represents the uncertainty band (e.g., 95% confidence interval) of the fit. Performance statistics (R^2, RMSE, MAE, and RPD) are reported in each panel.

In contrast, the Real + Filtered Synthetic setting yields consistent improvements across all models (Table 7; Figure 16). Augmented training densifies the feature space and reduces overfitting, leading to tighter clustering of predictions around the 1:1 line for held-out real samples (Figure 16).

Best-performing model (SVR). SVR benefits most from augmentation, with R^2 increasing from 0.36 to 0.60 and RMSE decreasing from 7.06 to 5.57 $dS \cdot m^{-1}$, while RPD improves to 1.61 (Table 7). This indicates that the proposed filtering-preserved synthetic set provides informative coverage of the spectral manifold and improves generalization to unseen real observations.

Table 7. Accuracy comparison of inversion models using LOOCV based on Baseline (Real only) and Augmented (Real + Synthetic) strategies.

Strategy	Training Data Size (N)	Model	R^2	RMSE ($dS \cdot m^{-1}$)	MAE	RPD	Improvement (R^2)
Real-only training (Baseline)	47 (Original Only)	SVR	0.36	7.06	5.78	1.27	-
		RF	0.37	7.01	5.65	1.28	-
		XGB	0.28	7.52	6.08	1.19	-
		PLSR	0.23	7.78	6.07	1.15	-
Real + synthetic training (Proposed)	1233 (Original + Generated)	SVR	0.6	5.57	4.63	1.61	+66.7%
		RF	0.54	5.97	4.97	1.5	+45.9%
		XGB	0.52	6.13	5.07	1.46	+85.7%
		PLSR	0.44	6.59	5.14	1.36	+91.3%

3.4.3. Spatial Mapping of Soil Salinity

Using the best-performing SVR trained under the Real + Filtered Synthetic setting (R^2 = 0.60, RPD = 1.61), we generated pixel-wise salinity (ECe) maps for all three sites (Figure 17). The resulting maps exhibit spatially coherent patterns rather than salt–noise speckle, suggesting that the augmented framework improves not only point-scale accuracy but also the geographic plausibility of the predicted salinity surfaces.

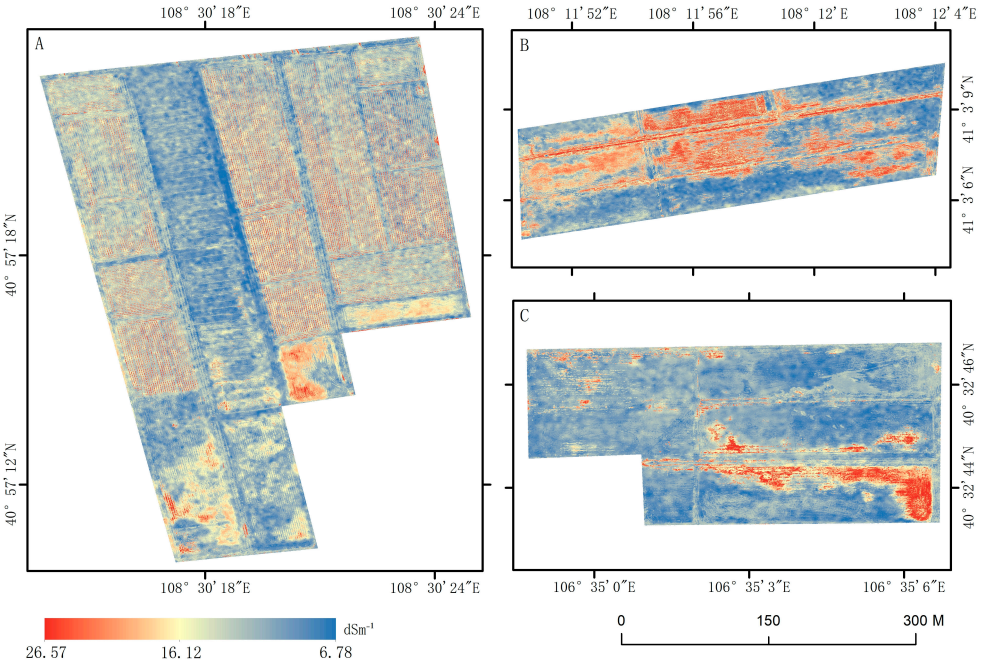


Figure 17. High-resolution spatial distribution maps of predicted soil salinity (ECe) across the three study sites.

Specifically, Site A shows strong heterogeneity with alternating low-to-moderate patches and localized high-salinity clusters, consistent with a mosaic-type salinization landscape. Site B is dominated by persistently high ECe over large contiguous areas, matching its severe, extensive salinization background. Site C shows a low-salinity baseline with narrow, structured bands of elevated ECe along edges/linear features, consistent with localized accumulation controlled by boundary effects and micro-relief.

4. Discussion

4.1. Physicochemical Interpretation of Spectral Sensitivity and Robust Feature Selection

Accurate salinity inversion fundamentally depends on identifying spectrally sensitive features with clear physicochemical meaning. Based on the OBCA screening and cross-site robustness evaluation, we found that the optimal bands are not randomly distributed; instead, they concentrate in the visible green region (≈550–570 nm) and the near-infrared (NIR) region (≈880–950 nm). This "spectral clustering" pattern aligns with the optical behavior of salt-affected soils, where salinization modifies surface brightness, particle-scale scattering, and water–salt coupling processes. Laboratory spectroscopy further indicates that saline soils with salt efflorescence/crusts often exhibit elevated reflectance in the VIS–NIR domain due to smoother crusted surfaces and altered scattering, and salt composition can reshape diagnostic spectral behavior.

The sensitivity in the green bands (e.g., around 560 nm) can be interpreted as a proxy for salinity-driven changes in surface color and micro-roughness. In irrigated drylands, salt accumulation often increases surface reflectance through salt crystallization and crusting, and may co-vary with residue/microstructure conditions that modulate visible reflectance. This VIS response is also strongly modulated by moisture state; moisture–reflectance interactions can change the apparent brightness and the visibility of salinity-related features, which partially explains why fixed-band indices may show limited transferability across scenes with different wetness/backgrounds. More importantly, the consistent selection of NIR bands approaching the water-absorption neighborhood (e.g., near 900–950 nm) across sites suggests that robust salinity signals are tightly coupled to soil moisture dynamics and salt hygroscopicity: salinity changes soil water potential and water binding status, which indirectly reshapes the NIR spectral response and stabilizes the salinity–reflectance relationship under variable field conditions [50]. In particular, prior spectral analyses of saline soils frequently report diagnostic behavior around the ~980 nm absorption neighborhood and NIR peaks/valleys that are sensitive to moisture-related absorption and saline-mineral effects, supporting the physical plausibility of our NIR-centered robust band synergies.

By coupling OBCA with robustness scoring, we aimed to suppress "pseudosensitive" bands that arise from site-specific noise, illumination variability, or transient surface states, thereby retaining only those features that are reproducible and physically interpretable across scenarios. Compared with prior work that selects sensitive bands within a single scene, our cross-site robustness scoring explicitly penalizes extreme-value effects and scene-specific artifacts, improving the likelihood that retained bands capture transferable salinity mechanisms rather than incidental acquisition/background noise. To further strengthen the physical interpretability in future work, semi-analytical radiative transfer modeling provides a complementary pathway to link observed spectral features to soil absorption/scattering processes in a more explainable way. (We add this as future work rather than claiming full mechanistic attribution in the current dataset.)

4.2. Manifold Densification as the Mechanism of Generative Data Augmentation

A core difficulty in salinity mapping over heterogeneous irrigation districts is the small-sample dilemma: sparse field observations cannot adequately cover the high-dimensional, non-linear spectral manifold [51]. The observed performance gain of the semi-supervised augmented scheme (e.g., R2 improving from 0.36 to 0.60) can be understood as manifold densification: generative models increase the sampling density in underrepresented regions of the feature space, which regularizes downstream regression and improves generalization.

In our framework, the WGAN-GP-based generator was used to produce 1186 synthetic spectra, which, when combined with 47 real samples, yielded 1233 training samples in total. Conceptually, these generated samples act as "connective tissue" between sparse

real clusters, reducing blind spots and enabling the regressor to learn a smoother mapping from reflectance to salinity. This interpretation is consistent with the WGAN-GP principle that gradient penalty stabilizes training by enforcing a 1-Lipschitz critic, which empirically reduces training pathologies and supports learning smoother distribution geometry.

However, "more synthetic data" is not automatically beneficial in hyperspectral regression: recent GAN literature in remote sensing emphasizes that physically informed constraints (e.g., spectral-angle distance/shape constraints and reflectance-domain regularization) are often necessary to keep generated spectra faithful to intrinsic material signatures under varying acquisition conditions. Therefore, our gain is not attributed to sample inflation alone, but to "physically constrained manifold densification": the generator is guided to populate plausible spectral neighborhoods while later stages prevent low-quality generations from entering training.

4.3. Uncertainty-Aware Stage Filtering to Prevent Noise Propagation in Semi-Supervised Learning

While generative augmentation expands feature coverage, the decisive factor is the reliability of pseudo-labels assigned to synthetic samples [52]. Naïvely accepting all pseudo-labeled samples can trigger confirmation bias: early teacher errors become reinforced as additional "training truth", degrading performance. This vulnerability of self-training to incorrect pseudo-label noise and error accumulation has been widely discussed in semi-supervised learning literature. To address this, we employed a stage-wise filtering strategy that uses uncertainty/consistency signals as gates, retaining only synthetic samples that lie in the teacher's reliable regions and rejecting those likely to carry noisy supervision.

Specifically, our protocol operationalizes "reliability" through a multi-criterion filter (e.g., spectral similarity/quality constraints and model-based uncertainty/consistency screening), which aligns with recent uncertainty-aware semi-supervised regression developments that explicitly weight or filter pseudo-labels based on predictive uncertainty to mitigate unreliable pseudo-supervision.

Overall, this stage filtering mechanism enables a practical balance: the GAN explores and populates new manifold regions, while uncertainty/consistency screening ensures that only trustworthy pseudo-supervision is injected into the student model. Importantly, this design connects the "physical plausibility" gates (spectral-shape fidelity) with "label credibility" gates (teacher uncertainty/consistency), which is essential for regression tasks where small label errors can directly bias continuous outputs.

4.4. Advantage and Limitations

Despite the small-sample setting, the proposed framework offers three practical advantages. First, OBCA provides a compact and more transferable spectral representation by explicitly searching salinity-sensitive three-band synergies, helping alleviate redundancy and collinearity that commonly destabilize hyperspectral regression. Second, the S-WGAN-GP backbone improves training stability via gradient penalty, while the added spectral constraints and pre-Teacher guardrails explicitly prioritize spectral fidelity over naive sample inflation—important because GAN-based remote-sensing augmentation can otherwise introduce physically implausible patterns and domain-shifted pseudo-supervision. Third, the Teacher module screens synthetic candidates using uncertainty/consistency diagnostics, reducing error accumulation when pseudo-labels are imperfect and improving robustness for cross-location learning, a key challenge highlighted in recent soil salinity remote-sensing/ML syntheses.

Notwithstanding these strengths, several limitations merit attention. First, to enhance spectral–chemical consistency, we removed outliers after screening; however, this may reduce sensitivity to extreme salinity conditions (e.g., very high ECe), where spec-

tral detectability can be complicated by surface state changes (crust thickness, moisture regime) and non-linear saturation-like effects. Second, although the synthetic sample size increased, the diversity of the seed real samples (N = 47) remains the fundamental bottleneck: GANs cannot reliably extrapolate beyond the support of the observed distribution. Third, some filtering thresholds (e.g., critic-score cutoffs and site-adaptive envelopes) are empirically calibrated and may require re-tuning under different sensors, seasons, or radiometric/BRDF conditions, which can change reflectance magnitude and local neighborhood structure. Finally, while our constraints emphasize spectral-shape plausibility, residual radiometric/illumination variability may still produce mild amplitude biases; incorporating stronger radiometric normalization and physics-based constraints is a clear direction for future work. In addition, our validation is limited to three UAV sites (A–C) with sparse field samples; thus, the reported accuracy should be interpreted as site-validated performance and cannot be directly generalized to large areas that are completely unsampled. More rigorous assessment of out-of-area generalization requires spatially explicit evaluation (e.g., region-held-out or spatially blocked cross-validation), because random or point-wise CV can underestimate prediction error in spatially structured data.

5. Conclusions

In this study, we tackled the small-sample bottleneck in UAV hyperspectral soil salinity mapping by integrating OBCA-based feature discovery with a physically constrained semi-supervised augmentation framework (S-WGAN-GP with teacher-guided reliability filtering), and validated the approach across three heterogeneous sites in the Hetao Irrigation District. The main conclusions are:

(1) Salinity-sensitive and transferable features. OBCA reduced spectral redundancy and identified salinity-sensitive bands primarily in the green and near-infrared regions; within the studied sites, NIR features around ~950 nm exhibited the most consistent cross-site behavior.

(2) A trustworthy augmentation mechanism. S-WGAN-GP generated shape-consistent spectra under reflectance-domain constraints, while teacher-guided filtering retained only high-confidence pseudo-labeled samples. Scientifically, the observed gains support a "physically constrained manifold densification + reliability-controlled pseudo-label selection" mechanism, rather than simple inflation of synthetic sample size.

(3) Improved generalization under the tested conditions. Relative to real-only training (N = 47), incorporating filtered synthetic samples improved LOOCV performance on real observations; the best SVR model increased R^2 from 0.36 to 0.60 (RPD = 1.61) and produced spatially coherent salinity maps across the three sites.

Limitations and outlook. Some screening thresholds may require re-tuning under different sensors/seasons or radiometric/BRDF conditions, and strict screening may reduce sensitivity to extreme salinity tails. Future work will expand multi-season, multi-region sampling and adopt region-held-out/spatially blocked evaluation to quantify large-area generalization under spatial autocorrelation, which is recommended for spatial prediction tasks. We will also further integrate physics-based spectral modeling (e.g., RTM-informed constraints) to better characterize extreme regimes.

Author Contributions: Conceptualization, S.Y. and L.S.; methodology, S.Y., L.S. and W.D.; software, S.Y.; validation, S.Y., W.D. and D.W.; formal analysis, S.Y. and L.S.; investigation, S.Y., W.D., D.W. and H.G.; resources, L.S., L.Y. and Y.Z.; data curation, S.Y., D.W. and H.G.; writing—original draft preparation, S.Y.; writing—review and editing, L.S., A.R., R.L., H.G. and L.Y.; visualization, S.Y. and D.W.; supervision, L.S.; project administration, S.Y.; funding acquisition, S.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Independent Research Program of Inner Mongolia Normal University (grant number 2025JYJFZX001); the Science and Technology Breakthrough Project of Inner Mongolia Autonomous Region (grant number 2025KJTW0026); the Science and Technology Program of the Department of Science and Technology of Inner Mongolia Autonomous Region (grant numbers 2024KJHZ0002 and 2024KJHZ0007); the Open Fund of the Inner Mongolia Academy of Forestry Sciences (grant number KF2025ZD08); the Key Special Project of the "Science and Technology to Revitalize Inner Mongolia" Action (grant number 2020ZD0028); and the project "Integrated Ecological Rehabilitation Technologies and Multi-Scenario Demonstration for Saline-Alkali Land in the Hetao Plain" (grant number 2022YFHH0144).

Data Availability Statement: The data presented in this study are available on reasonable request from the corresponding author. The data are not publicly available due to privacy restrictions related to sensitive georeferenced location information.

Acknowledgments: The authors gratefully acknowledge the support of the Inner Mongolia Autonomous Region Key Laboratory of Ecological Environment Collaborative Intelligence and the Arxan Forest and Grassland Disaster Prevention and Mitigation Field Scientific Observation and Research Station of Inner Mongolia Autonomous Region for providing the research platform support.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Wang, J.; Ding, J.; Wang, Y.; Ge, X.; Lizaga, I.; Chen, X. Soil salinization in drylands: Measure, monitor, and manage. *Ecol. Indic.* **2025**, *175*, 113608. [CrossRef]
2. Sahbeni, G.; Ngabire, M.; Musyimi, P.K.; Székely, B. Challenges and Opportunities in Remote Sensing for Soil Salinization Mapping and Monitoring: A Review. *Remote Sens.* **2023**, *15*, 2540. [CrossRef]
3. Shokri, N.; Hassani, A.; Sahimi, M. Multi-Scale Soil Salinization Dynamics From Global to Pore Scale: A Review. *Rev. Geophys.* **2024**, *62*, e2023RG000804. [CrossRef]
4. Jiang, Z.; Ding, J.; Li, Z.; Liu, J. Present knowledge and future challenges in remote sensing for soil salinization monitoring: A review of bibliometric analysis. *Int. J. Remote Sens.* **2024**, *46*, 247–272. [CrossRef]
5. Wang, F.; Han, L.; Liu, L.; Bai, C.; Ao, J.; Hu, H.; Li, R.; Li, X.; Guo, X.; Wei, Y. Advancements and Perspective in the Quantitative Assessment of Soil Salinity Utilizing Remote Sensing and Machine Learning Algorithms: A Review. *Remote Sens.* **2024**, *16*, 4812. [CrossRef]
6. Chen, H.; Wu, J.; Xu, C. Monitoring Soil Salinity Classes through Remote Sensing-Based Ensemble Learning Concept: Considering Scale Effects. *Remote Sens.* **2024**, *16*, 642. [CrossRef]
7. Chindong, J.M.; Ouzemou, J.-E.; Laamrani, A.; El Battay, A.; Hajaj, S.; Rhinane, H.; Chehbouni, A. A Multi-Sensor Machine Learning Framework for Field-Scale Soil Salinity Mapping Under Data-Scarce Conditions. *Remote Sens.* **2025**, *17*, 3778. [CrossRef]
8. Jia, J.; Chen, C.; Liu, Q.; Ding, B.; Ren, Z.; Jia, Y.; Bai, X.; Du, R.; Chen, Q.; Wang, S.; et al. Soil salinity monitoring model based on the synergistic construction of ground-UAV-satellite data. *Soil Use Manag.* **2023**, *40*, e12980. [CrossRef]
9. Zhao, Y.; Miao, Q.; Shi, H.; Li, X.; Yan, J.; Yang, S.; Hou, C.; Yu, C.; Feng, W.; Hao, J. Inversion of soil salinization at the branch canal scale in the Hetao Irrigation District based on improved spectral indices. *Agric. Water Manag.* **2025**, *316*, 109608. [CrossRef]
10. Das, A.; Bhattacharya, B.K.; Setia, R.; Jayasree, G.; Sankar Das, B. A novel method for detecting soil salinity using AVIRIS-NG imaging spectroscopy and ensemble machine learning. *ISPRS J. Photogramm. Remote Sens.* **2023**, *200*, 191–212. [CrossRef]
11. Li, J.; Ding, J.; Wang, J.; Tan, J.; Ge, X.; Cui, K.; Liu, Y. Daily-scale dynamic monitoring of soil salinization in arid cotton fields based on multi-source remote sensing spatiotemporal fusion. *Int. J. Appl. Earth Obs. Geoinf.* **2025**, *144*, 104940. [CrossRef]
12. Lazzeri, G.; Milewski, R.; Foerster, S.; Moretti, S.; Chabrillat, S. Early Detection of Soil Salinization by Means of Spaceborne Hyperspectral Imagery. *Remote Sens.* **2025**, *17*, 2486. [CrossRef]
13. Song, Q.; Gao, X.; Song, Y.; Li, Q.; Chen, Z.; Li, R.; Zhang, H.; Cai, S. Estimation and mapping of soil texture content based on unmanned aerial vehicle hyperspectral imaging. *Sci. Rep.* **2023**, *13*, 14097. [CrossRef] [PubMed]
14. Tan, J.; Ding, J.; Wang, Z.; Han, L.; Wang, X.; Li, Y.; Zhang, Z.; Meng, S.; Cai, W.; Hong, Y. Estimating soil salinity in mulched cotton fields using UAV-based hyperspectral remote sensing and a Seagull Optimization Algorithm-Enhanced Random Forest Model. *Comput. Electron. Agric.* **2024**, *221*, 109017. [CrossRef]
15. Zhang, J.; Ding, J.; Tan, J.; Wang, J.; Zhang, Z.; Wang, Z.; Ge, X. Monitoring soil salinization in Arid cotton fields using Unmanned Aerial Vehicle hyperspectral imagery. *Int. J. Appl. Earth Obs. Geoinf.* **2025**, *140*, 104584. [CrossRef]
16. Chen, H.; Wu, J.; Xu, C. A deep learning-based method for deep soil salinity prediction: Considering the driving mechanisms of salinity profiles. *Geoderma* **2025**, *464*, 117615. [CrossRef]

17. Bandak, S.; Movahedi-Naeini, S.A.; Mehri, S.; Lotfata, A. A longitudinal analysis of soil salinity changes using remotely sensed imageries. *Sci. Rep.* **2024**, *14*, 10383. [CrossRef]
18. Tan, J.; Ding, J.; Han, L.; Ge, X.; Wang, X.; Wang, J.; Wang, R.; Qin, S.; Zhang, Z.; Li, Y. Exploring PlanetScope Satellite Capabilities for Soil Salinity Estimation and Mapping in Arid Regions Oases. *Remote Sens.* **2023**, *15*, 1066. [CrossRef]
19. Yang, J.; Guo, B.; Zhang, R. The Optimal Estimation Model for Soil Salinization Based on the FOD-CNN Spectral Index. *Remote Sens.* **2025**, *17*, 2357. [CrossRef]
20. Ma, M.; Hao, Y.; Huang, Q.; Liu, Y.; Xiu, L.; Gao, Q. Soil Salinity Estimation by 3D Spectral Space Optimization and Deep Soil Investigation in the Songnen Plain, Northeast China. *Sustainability* **2024**, *16*, 2069. [CrossRef]
21. Li, J.; Zhang, T.; Shao, Y.; Ju, Z. Comparing Machine Learning Algorithms for Soil Salinity Mapping Using Topographic Factors and Sentinel-1/2 Data: A Case Study in the Yellow River Delta of China. *Remote Sens.* **2023**, *15*, 2332. [CrossRef]
22. Zhai, J.; Wang, N.; Hu, B.; Han, J.; Feng, C.; Peng, J.; Luo, D.; Shi, Z. Estimation of Soil Salinity by Combining Spectral and Texture Information from UAV Multispectral Images in the Tarim River Basin, China. *Remote Sens.* **2024**, *16*, 3671. [CrossRef]
23. Shi, S.; Wang, N.; Chen, S.; Hu, B.; Peng, J.; Shi, Z. Digital mapping of soil salinity with time-windows features optimization and ensemble learning model. *Ecol. Inform.* **2025**, *85*, 102982. [CrossRef]
24. Hong, Y.; Chen, Y.; Yu, L.; Liu, Y.; Liu, Y.; Zhang, Y.; Liu, Y.; Cheng, H. Combining Fractional Order Derivative and Spectral Variable Selection for Organic Matter Estimation of Homogeneous Soil Samples by VIS–NIR Spectroscopy. *Remote Sens.* **2018**, *10*, 479. [CrossRef]
25. Zhu, C.; Ding, J.; Zhang, Z.; Wang, Z. Exploring the potential of UAV hyperspectral image for estimating soil salinity: Effects of optimal band combination algorithm and random forest. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2022**, *279*, 121416. [CrossRef] [PubMed]
26. Aksoy, S.; Sertel, E.; Roscher, R.; Tanik, A.; Hamzehpour, N. Assessment of soil salinity using explainable machine learning methods and Landsat 8 images. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *130*, 103879. [CrossRef]
27. Ge, H.; Han, Y.; Xu, Y.; Zhuang, L.; Wang, F.; Gu, Q.; Li, X. Estimating soil salinity using multiple spectral indexes and machine learning algorithm in Songnen Plain, China. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 7041–7050. [CrossRef]
28. Sirpa-Poma, J.W.; Satge, F.; Resongles, E.; Pillco-Zola, R.; Molina-Carpio, J.; Flores Colque, M.G.; Ormachea, M.; Pacheco Mollinedo, P.; Bonnet, M.P. Towards the Improvement of Soil Salinity Mapping in a Data-Scarce Context Using Sentinel-2 Images in Machine-Learning Models. *Sensors* **2023**, *23*, 9328. [CrossRef]
29. Sarkar, S.K.; Rudra, R.R.; Sohan, A.R.; Das, P.C.; Ekram, K.M.M.; Talukdar, S.; Rahman, A.; Alam, E.; Islam, M.K.; Islam, A. Coupling of machine learning and remote sensing for soil salinity mapping in coastal area of Bangladesh. *Sci. Rep.* **2023**, *13*, 17056. [CrossRef]
30. Kaplan, G.; Gašparović, M.; Alqasemi, A.S.; Aldhaheri, A.; Abuelgasim, A.; Ibrahim, M. Soil salinity prediction using Machine Learning and Sentinel—2 Remote Sensing Data in Hyper—Arid areas. *Phys. Chem. Earth Parts A/B/C* **2023**, *130*, 103400. [CrossRef]
31. Li, J.; Ding, J.; Tan, J.; Wang, J.; Han, L.; Cui, K. Estimating soil salinity in cotton fields using UAV multispectral remote sensing and SSA–SVM optimised machine learning model. *Int. J. Digit. Earth* **2025**, *18*, 2548378. [CrossRef]
32. Feng, J.; Gao, Z.; Shang, R.; Zhang, X.; Jiao, L. Multi-Complementary Generative Adversarial Networks with Contrastive Learning for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–18. [CrossRef]
33. Zhang, M.; Wang, Z.; Wang, X.; Gong, M.; Wu, Y.; Li, H. Features kept generative adversarial network data augmentation strategy for hyperspectral image classification. *Pattern Recognit.* **2023**, *142*, 109701. [CrossRef]
34. Hao, S.; Xia, Y.; Ye, Y. Generative Adversarial Network with Transformer for Hyperspectral Image Classification. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 1–5. [CrossRef]
35. Yu, Z.; Cui, W. Robust hyperspectral image classification using generative adversarial networks. *Inf. Sci.* **2024**, *666*, 120452. [CrossRef]
36. Zhu, Y.; Su, H.; Xu, P.; Xu, Y.; Wang, Y.; Dong, C.H.; Lu, J.; Le, Z.; Yang, X.; Xuan, Q.; et al. Data augmentation using continuous conditional generative adversarial networks for regression and its application to improved spectral sensing. *Opt. Express* **2023**, *31*, 37722–37739. [CrossRef]
37. Zhou, M.; Li, Y. Digital Mapping and Scenario Prediction of Soil Salinity in Coastal Lands Based on Multi-Source Data Combined with Machine Learning Algorithms. *Remote Sens.* **2024**, *16*, 2681. [CrossRef]
38. Silva-Silverio, A.; Gómez-Gil, P.; Sánchez-Argüelles, D.O. Conditional GAN Approaches on Regression Labels: A State-of-the-Art Review. *WIREs Data Min. Knowl. Discov.* **2025**, *15*, e70050. [CrossRef]
39. Wen, Y.; Wan, H.; Shang, S. A monthly distributed water and salt balance model for agro- and non-agro areas with shallow groundwater table in the Hetao irrigation district. *J. Hydrol.* **2023**, *616*, 128811. [CrossRef]
40. Qian, X.; Qi, H.; Shang, S.; Wan, H.; Rahman, K.U.; Wang, R. Deep Learning-based Near-real-time Monitoring of Autumn Irrigation Extent at Sub-pixel Scale in a Large Irrigation District. *Agric. Water Manag.* **2023**, *284*, 108335. [CrossRef]

41. Liu, Y.; Li, X.; Liu, X.; Han, F.; Yang, B.; Zheng, Y.; Chen, Z.; Du, H.; Liu, Y.; Hu, L.; et al. Optimizing irrigation and planting strategies to prevent non-point source pollution in the Hetao Irrigation District of Inner Mongolia, China. *Sci. Total Environ.* **2024**, *957*, 177757. [CrossRef] [PubMed]
42. Fu, C.; Xue, J.; Chen, J.; Cui, L.; Wang, H. Evaluating spatial and temporal variations of soil water, heat, and salt under autumn irrigation in the Hetao Irrigation District based on distributed SHAW model. *Agric. Water Manag.* **2024**, *293*, 108707. [CrossRef]
43. Hogg, T.J.; Henry, J.L. Comparison of 1:1 and 1:2 Suspensions and Extracts with the Saturation Extract in Estimating Salinity in Saskatchewan Soils. *Can. J. Soil Sci.* **1984**, *64*, 699–704. [CrossRef]
44. U.S. Salinity Laboratory. *Diagnosis and Improvement of Saline and Alkali Soils*; U.S. Department of Agriculture: Washington, DC, USA, 1954.
45. Hong, Y.; Munnaf, M.A.; Guerrero, A.; Chen, S.; Liu, Y.; Shi, Z.; Mouazen, A.M. Fusion of visible-to-near-infrared and mid-infrared spectroscopy to estimate soil organic carbon. *Soil Tillage Res.* **2022**, *217*, 105284. [CrossRef]
46. Tian, Y.C.; Yao, X.; Yang, J.; Cao, W.X.; Hannaway, D.B.; Zhu, Y. Assessing newly developed and published vegetation indices for estimating rice leaf nitrogen concentration with ground- and space-based hyperspectral reflectance. *Field Crops Res.* **2011**, *120*, 299–310. [CrossRef]
47. Li, F.; Mistele, B.; Hu, Y.; Chen, X.; Schmidhalter, U. Optimising three-band spectral indices to assess aerial N concentration, N uptake and aboveground biomass of winter wheat remotely in China and Germany. *ISPRS J. Photogramm. Remote Sens.* **2014**, *92*, 112–123. [CrossRef]
48. NASA's Applied Remote Sensing Training Program. Spectral Indices for Land and Aquatic Applications. Available online: https://www.earthdata.nasa.gov/learn/trainings/spectral-indices-land-aquatic-applications (accessed on 9 February 2026).
49. Yang, H.; Li, F.; Hu, Y.; Yu, K. Hyperspectral indices optimization algorithms for estimating canopy nitrogen concentration in potato (*Solanum tuberosum* L.). *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *102*, 102416. [CrossRef]
50. Lin, N.; Ma, X.; Sui, Y.; Zhu, R.; Liu, H.; Wu, M.; Jiang, R. Soil salinity estimation based on satellite hyperspectral and synthetic aperture radar remote sensing image fusion. *Agric. Water Manag.* **2026**, *323*, 110076. [CrossRef]
51. Lyu, H.; Grafton, M.; Ramilan, T.; Irwin, M.; Sandoval, E. Synthetic hyperspectral reflectance data augmentation by generative adversarial network to enhance grape maturity determination. *Comput. Electron. Agric.* **2025**, *235*, 110341. [CrossRef]
52. Jo, Y.; Kahng, H.; Kim, S.B. Deep semi-supervised regression via pseudo-label filtering and calibration. *Appl. Soft Comput.* **2024**, *161*, 111670. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.