

Comparative Evaluation of Six Spectral Water Indices and Machine-Learning Cross-Sensor Harmonization of Sentinel-2 and Landsat 8/9 Water-Area Retrievals for the Chah-Nimeh 3 Reservoir, Sistan, Iran

Hadi Siasar^{1*}✉

¹ Assistant Professor, Department of Agricultural Engineering Payame Noor University, Tehran, Iran

*Corresponding Authors: Email: (hadisiasar@pnu.ac.ir)

Received: 04 January 2026

Accepted: 06 February 2026

Available online: 05 March 2026

Abstract: The Chah-Nimeh reservoirs are the only reliable surface-water supply of the Sistan region in south-eastern Iran, and their monitoring depends almost entirely on satellite imagery. This study compares six spectral water indices (NDWI, MNDWI, WRI, TCW, MBWI and NDMI) computed from monthly Sentinel-2 (2018–2025, 85 months) and Landsat 8/9 (2017–2025, 103 months) composites of the Chah-Nimeh 3 reservoir, and tests whether machine learning can harmonize the two sensors. Within each sensor, NDWI, MNDWI and WRI were practically interchangeable ($r > 0.99$), whereas NDMI departed most ($r = 0.79$ for Sentinel-2). Between sensors, agreement was limited: for MNDWI the Landsat areas were on average 10.0 km² smaller than the Sentinel-2 areas, with an RMSE of 14.5 km², a correlation of 0.61 and 95% limits of agreement of -30.6 to 10.6 km². A random forest classifier that uses only Landsat information detected the unreliable Landsat months ($|\Delta A| > 5$ km²) with an area under the ROC curve of 0.79. Seven regression models (MLR, SVR, RF, XGBoost, SVR-GWO, XGB-PSO and LSTM) were then trained to convert Landsat indices into Sentinel-2-equivalent areas. Under a random split all models appeared skilful (RF: $R^2 = 0.42$; LSTM: $R^2 = 0.97$), but under a chronological split only the LSTM retained positive efficiency (NSE = 0.30) while all tabular models failed (NSE < 0). The Sentinel-2 record shows a significant decline of the water surface (Mann–Kendall $Z = -8.1$), from an annual mean of 41.8 km² in 2020 to 11.8 km² in 2023, followed by a partial recovery. The results warn against random splitting of autocorrelated satellite time series and show that Landsat and Sentinel-2 water areas of small, dust-affected reservoirs should not be merged without quality screening.

Keywords: Surface water mapping, Spectral water indices, Sentinel-2, Landsat, Sensor harmonization, Chah-Nimeh, Sistan

1. Introduction

Surface water extent is one of the most widely used indicators of water availability in arid regions. Since the launch of the free and open Landsat and Sentinel archives, water bodies have routinely been delineated with spectral water indices, and cloud platforms such as Google Earth Engine (Gorelick et al., 2017) have made long, consistent time series accessible to water managers. Global products such as the Global Surface Water dataset (Pekel et al., 2016) demonstrate the potential of these archives, but they are generated with a single algorithm and may not be optimal for small reservoirs in dusty, hyper-arid environments.

A large number of indices have been proposed. The Normalized Difference Water Index (NDWI; McFeeters, 1996) contrasts green and near-infrared reflectance. The Modified NDWI (MNDWI; Xu, 2006) replaces the near-infrared with the shortwave infrared band and suppresses built-up noise. The Water Ratio Index (WRI; Shen and Li, 2010), the Tasseled Cap Wetness component (TCW; Crist, 1985), the Multi-Band Water Index (MBWI; Wang et al., 2018) and the Normalized Difference Moisture Index (NDMI; Gao, 1996) exploit different band combinations. Comparative studies show that the ranking of indices depends on the

sensor, the water body and the atmospheric conditions, so a local evaluation is usually required.

Combining sensors is attractive because Sentinel-2 offers 10–20 m resolution and a 5-day revisit while Landsat provides a much longer record. Harmonized surface-reflectance products such as HLS (Claverie et al., 2018) aim at a seamless combination, but they do not remove the differences that arise from spatial resolution, band placement and the number of cloud-free observations available in a month. For a reservoir of about 10–50 km², a single contaminated Landsat scene can bias a monthly composite by tens of percent, especially in the Sistan basin where the summer 'wind of 120 days' produces frequent dust storms (Rashki et al., 2012).

The Chah-Nimeh reservoirs are four natural depressions near Zabol that are filled from the Hirmand (Helmand) River and supply drinking and irrigation water to more than 400,000 people in the Sistan plain. The desiccation of the Hamoun wetlands and the recurrent droughts of the basin have made their monitoring a priority (Sharifikia, 2013). The objectives of this study were (i) to compare six spectral water indices for Sentinel-2 and Landsat 8/9 over the Chah-Nimeh 3 reservoir, (ii) to quantify the agreement

between the two sensors and to detect unreliable Landsat months with a classifier, and (iii) to test whether machine learning and deep learning models can convert Landsat index values into Sentinel-2-equivalent water areas, both under a random split and under a chronological split.

2. Materials and Methods

2.1. Study Area

Chah-Nimeh 3 (approximately 30.83° N, 61.69° E, about 475–495 m a.s.l.) is one of four interconnected reservoirs located south-east of Zabol, in Sistan and Baluchestan province, Iran (Fig. 1). The reservoirs are filled by diversions from the Hirmand River, which originates in Afghanistan, and act as the strategic water store of the region. The climate is hyper-arid, with mean annual precipitation below 60 mm, potential evapotranspiration above 3,000 mm and strong summer winds that cause dust storms and high evaporation losses.

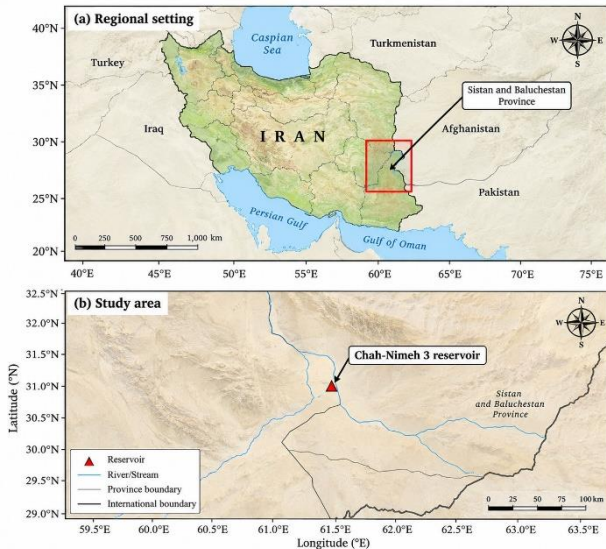


Figure 1. Location of the Chah-Nimeh 3 reservoir in the Sistan region, south-eastern Iran

2.2. Satellite Data and Water Indices

Monthly water-surface areas derived from six spectral indices were used. The Sentinel-2 record covers December 2018 – December 2025 (85 monthly values) and the Landsat 8/9 record January 2017 – December 2025 (103 monthly values); 82 months are common to both sensors. The series were produced in Google Earth Engine from atmospherically corrected surface-reflectance collections, using monthly cloud-masked composites and a fixed threshold for each index. The indices are defined as

$$NDWI = (\rho_{\text{green}} - \rho_{\text{NIR}}) / (\rho_{\text{green}} + \rho_{\text{NIR}}) \quad (1)$$

$$MNDWI = (\rho_{\text{green}} - \rho_{\text{SWIR1}}) / (\rho_{\text{green}} + \rho_{\text{SWIR1}}) \quad (2)$$

$$NDMI = (\rho_{\text{NIR}} - \rho_{\text{SWIR1}}) / (\rho_{\text{NIR}} + \rho_{\text{SWIR1}}) \quad (3)$$

together with WRI (the ratio of the sum of green and red to the sum of near- and shortwave infrared reflectance), the Tasselled Cap Wetness component and

the Multi-Band Water Index, which combines the green, red, near-infrared and both shortwave infrared bands. Because MNDWI is the most widely used index for open water in arid regions, the Sentinel-2 MNDWI area was taken as the reference variable in the modelling experiments.

2.3. Agreement Analysis

For each index, the agreement between sensors was quantified on the common months by the mean difference (bias), the RMSE, the Pearson correlation and the slope of the ordinary least-squares regression of the Sentinel-2 area on the Landsat area. Bland and Altman (1986) plots with 95% limits of agreement (mean difference ± 1.96 SD) were used to visualize the structure of the disagreement.

2.4. Detection of Unreliable Landsat Months

A month was labelled unreliable when the absolute difference between the Landsat and Sentinel-2 MNDWI areas exceeded 5 km², which occurred in 50 of the 82 common months. Three classifiers (logistic regression, random forest and XGBoost) were trained to identify these months from Landsat-only information: the six index areas, their range across indices (called the index spread below) and the sine and cosine of the month. Performance was assessed by stratified five-fold cross-validation with the area under the ROC curve (AUC) and the accuracy.

2.5. Cross-Sensor Regression Models

The harmonization models estimate the Sentinel-2 MNDWI area from ten Landsat predictors: the six index areas, the index spread, the median of the six indices and the sine and cosine of the month. Seven models were compared: multiple linear regression (MLR), support vector regression (SVR; Drucker et al., 1997), random forest (RF; Breiman, 2001), XGBoost (Chen and Guestrin, 2016), an SVR tuned by the grey wolf optimizer (SVR-GWO; Mirjalili et al., 2014), an XGBoost tuned by particle swarm optimization (XGB-PSO; Kennedy and Eberhart, 1995) and a long short-term memory network (LSTM; Hochreiter and Schmidhuber, 1997) that receives the previous six months of Landsat indices. The metaheuristic searches used 20 agents and 40 iterations with the validation RMSE as fitness (MEALPY; Van Thieu and Mirjalili, 2023), and five LSTM networks with different seeds were averaged.

Two experiments were carried out. In experiment A the months were assigned randomly to training (60%, $n = 49$), validation (20%, $n = 16$) and test (20%, $n = 17$) sets, which is the usual practice when a static transfer function is calibrated. In experiment B the split was chronological: training up to December 2022 ($n = 46$), validation in 2023 ($n = 12$) and test in 2024–2025 ($n = 24$). Comparing the two experiments shows how much of the apparent skill of a random split is produced by the temporal autocorrelation of the series. Performance was

measured with R^2 , RMSE, MAE, NSE (Nash and Sutcliffe, 1970) and Willmott's index (Willmott, 1981).

3. Results

3.1. Behaviour of the Indices

Fig. 2 shows the two records. The Sentinel-2 series is smooth and physically plausible: the water area increased to 49.2 km² in May 2020, fell to 6.2 km² in February 2024 and then partly recovered. The annual mean MNDWI area decreased from 41.8 km² (2020) to 11.8 km² (2023), a loss of 72%, and the Mann–Kendall test confirms a highly significant negative trend ($Z = -8.14$, $p < 0.001$, Sen's slope = -4.97 km²/year). The Landsat series follows the same general evolution but contains strong month-to-month fluctuations: in 2025, for example, the Landsat MNDWI area alternates between 1.5 and 20 km² while Sentinel-2 varies only between 17 and 21 km².

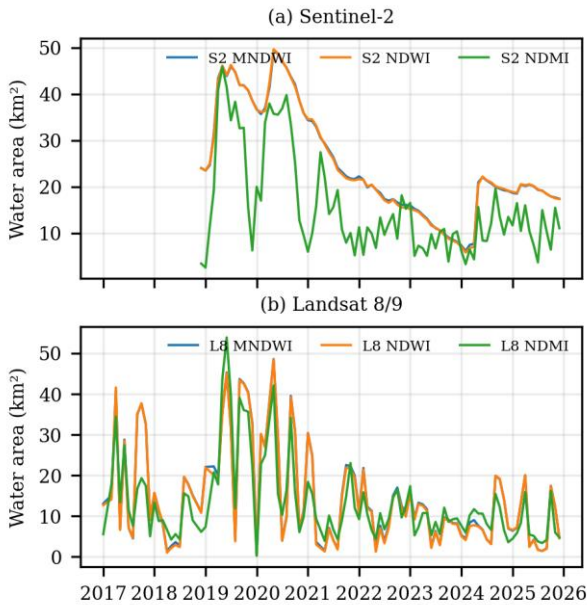


Figure 2. Monthly water-surface area of the Chah-Nimeh 3 reservoir from three indices of (a) Sentinel-2 and (b) Landsat 8/9

Within each sensor, NDWI, MNDWI and WRI produced almost identical areas ($r > 0.999$ for both sensors). TCW and MBWI deviated moderately ($r = 0.88$ and 0.98 for Sentinel-2), and NDMI was the most different index ($r = 0.79$ for Sentinel-2 and 0.88 for Landsat). NDMI is a vegetation moisture index, so it is sensitive to the moist margins of the reservoir rather than to open water. The correlation matrix (Fig. 5) also shows that the differences among the indices of one sensor are much smaller than the differences between the two sensors.

3.2. Agreement Between Sensors

Table 1. Agreement between Landsat 8/9 and Sentinel-2 water areas for the 82 common months (km²)

Index	S2 mean	L8 mean	Bias (L8-S2)	RMSE	r	Slope
NDWI	24.36	14.20	-10.17	14.56	0.611	0.61
MNDWI	24.48	14.48	-10.00	14.45	0.608	0.60
WRI	24.39	14.44	-9.95	14.41	0.611	0.61

TCW	19.38	13.97	-5.42	11.49	0.585	0.54
MBWI	21.48	11.36	-10.12	14.31	0.554	0.55
NDMI	15.14	13.27	-1.87	9.18	0.636	0.65

Table 1 summarizes the agreement. For all six indices, Landsat systematically reported smaller areas than Sentinel-2, with biases between -1.9 and -10.2 km². Correlations ranged from 0.55 (MBWI) to 0.64 (NDMI), and the regression slopes of about 0.54 – 0.65 show that the Landsat areas also underestimate the amplitude of the variations. For the reference index (MNDWI), the bias was -10.00 km² with an RMSE of 14.45 km². The Bland–Altman plot (Fig. 4) gives 95% limits of agreement of -30.6 to 10.6 km², which is wider than the mean water area of the reservoir in the dry years. The scatter plots (Fig. 3) show that the disagreement is not a simple scale factor: a large group of months lies far below the 1:1 line, where Landsat reports almost no water while Sentinel-2 reports 15 – 25 km².

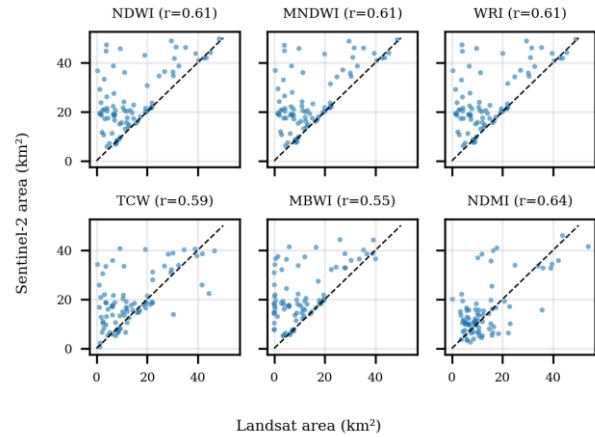


Figure 3. Sentinel-2 versus Landsat 8/9 water areas for the six indices (dashed: 1:1 line)

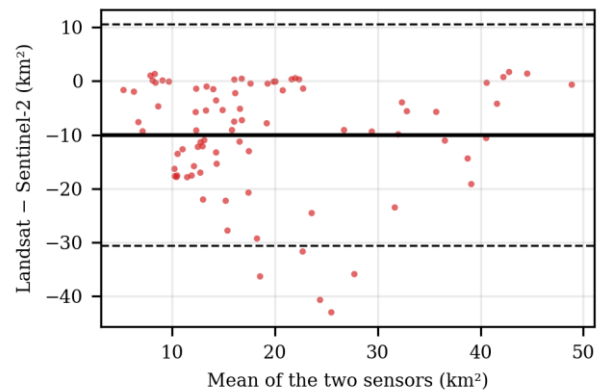


Figure 4. Bland–Altman plot for the MNDWI areas (solid: mean difference; dashed: 95% limits of agreement)

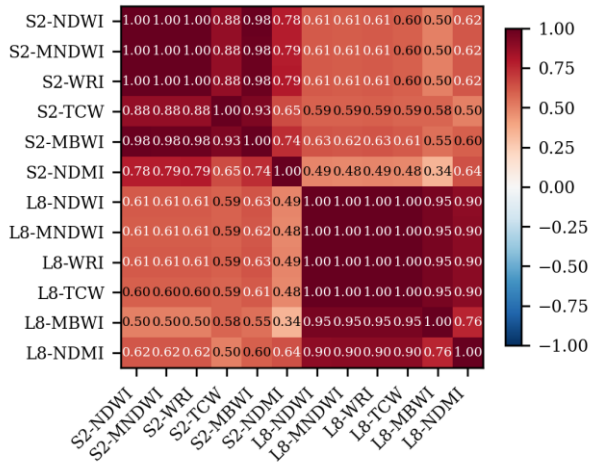


Figure 5. Correlation matrix of the twelve index-sensor combinations over the common months

3.3. Detection of Unreliable Landsat Months

Table 2. Cross-validated detection of unreliable Landsat months ($|\Delta A| > 5 \text{ km}^2$)

Classifier	AUC	Accuracy (%)	Hits	False alarms
Logistic	0.738	69.5	40	15
RF	0.791	74.4	40	11
XGB	0.768	73.2	40	12

50 of 82 months were labelled unreliable; five-fold stratified cross-validation.

Half of the common months (50 of 82) differed by more than 5 km^2 between the sensors. The random forest identified them from Landsat information alone with an AUC of 0.79 and an accuracy of 74%, followed by XGBoost (AUC = 0.77) and logistic regression (AUC = 0.74). The most informative variable was the spread between the six Landsat indices: when the indices disagree with each other, the composite is usually contaminated. This provides a simple internal quality flag that can be computed without any reference sensor (Fig. 6).

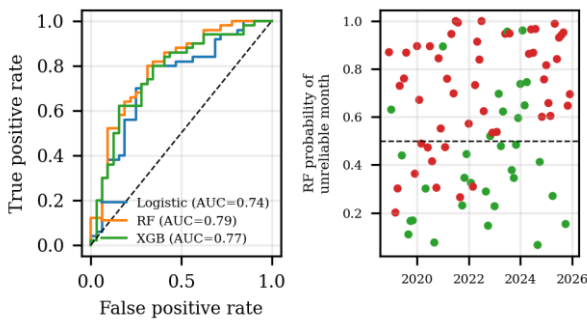


Figure 6. (a) ROC curves of the three classifiers and (b) random forest probability of an unreliable Landsat month (red: labelled unreliable)

3.4. Cross-Sensor Harmonization

Table 3. Test-period performance of the harmonization models under a random split (A) and a chronological split (B); RMSE in km^2

Model	R ² (A)	RMSE (A)	NSE (A)	R ² (B)	RMSE (B)	NSE (B)
MLR	0.552	8.66	0.496	0.390	16.25	-
SVR	0.350	10.53	0.254	0.023	10.75	-4.104
RF	0.420	9.67	0.372	0.016	9.79	-3.231

XGBoost	0.400	10.75	0.223	0.026	9.98	-3.399
SVR-GWO	0.412	9.76	0.360	0.030	7.07	-1.207
XGB-PSO	0.381	10.70	0.230	0.001	11.56	-4.900
LSTM	0.966	2.49	0.958	0.343	3.97	0.304

Under the random split (Table 3, columns A), all models produced positive efficiencies. The LSTM appeared extremely accurate ($R^2 = 0.966$, $\text{RMSE} = 2.49 \text{ km}^2$, $\text{NSE} = 0.958$), followed by MLR ($\text{NSE} = 0.50$), RF (0.37), SVR-GWO (0.36), SVR (0.25), XGB-PSO (0.23) and XGBoost (0.22). Under the chronological split (columns B), the picture changed completely: every tabular model produced a negative NSE, i.e. a worse forecast than the mean of the test period, and only the LSTM remained positive ($\text{NSE} = 0.30$, $\text{RMSE} = 3.97 \text{ km}^2$). The linear model degraded most ($\text{NSE} = -10.7$), because it extrapolated the strong bias of the training years into a period in which the reservoir was much smaller.

The explanation is that the Sentinel-2 series is strongly autocorrelated. With a random split, each test month lies between two training months, and a model can reach a good score simply by interpolating between neighbouring values, even without a real relationship to the Landsat inputs. This is confirmed by the very large difference between the two experiments and by the fact that the LSTM, which explicitly uses a six-month window of Landsat data, is the only model that keeps some skill when it must extrapolate in time. Studies that harmonize sensors or calibrate remote-sensing relationships with random cross-validation should therefore report a chronological evaluation as well.

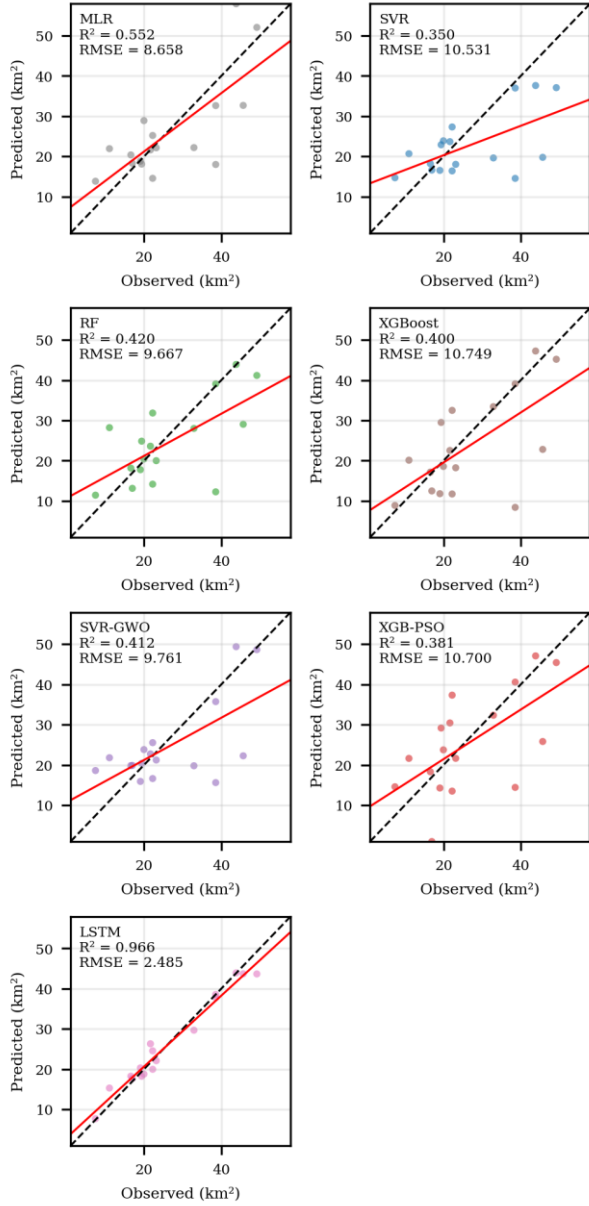


Figure 7. Observed versus predicted Sentinel-2 MNDWI area in the test period of experiment A (random split)

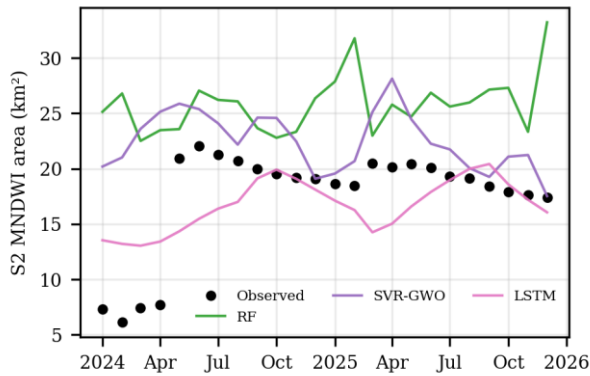


Figure 8. Observed and predicted Sentinel-2 MNDWI area in the test period of experiment B (2024–2025)

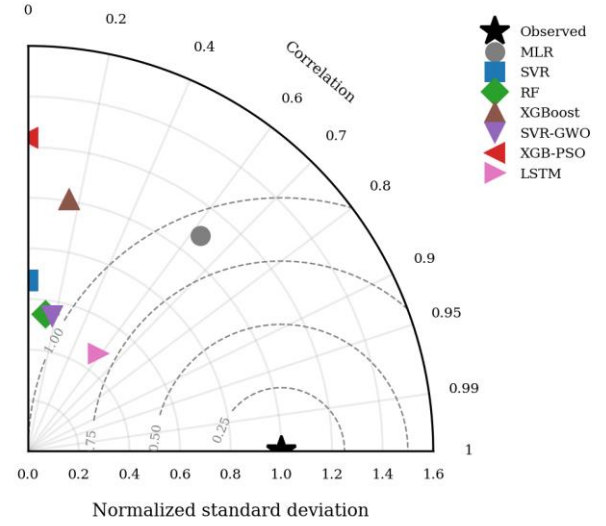


Figure 9. Taylor diagram of the chronological test period (experiment B)

The metaheuristic searches reduced the validation RMSE of the standardized target from 0.452 to 0.444 (SVR-GWO) and from 0.524 to 0.434 (XGB-PSO) in experiment A (Fig. 10). The tuned models were slightly better than their untuned counterparts in experiment A (SVR: 0.25 $\rightarrow$ 0.36), but tuning could not compensate for the loss of transferability in experiment B. The permutation importance (Fig. 11) confirms that the index spread and NDMI, rather than the individual water indices, carried most of the information: the models mainly learn when a Landsat composite can be trusted.

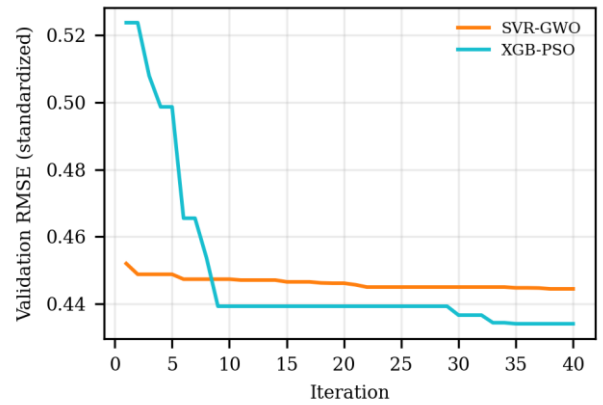


Figure 10. Convergence of the GWO and PSO searches in experiment A

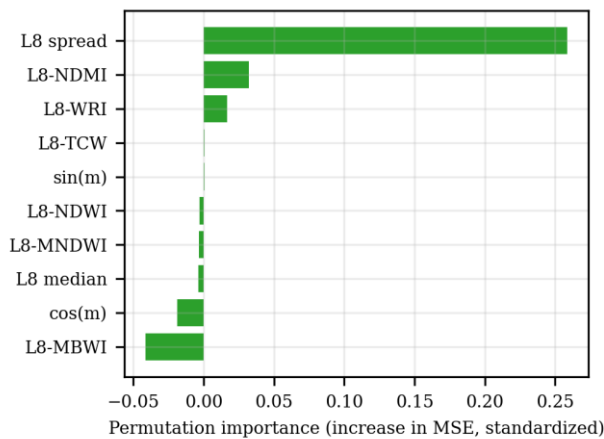


Figure 11. Permutation importance of the random forest inputs (experiment A)

4. Discussion

Three practical conclusions follow from the comparison. First, for this reservoir the choice among NDWI, MNDWI and WRI is irrelevant: the three indices delineate the same pixels. The choice of sensor and of the compositing strategy matters far more, which reverses the emphasis of many index-comparison studies. Second, Landsat monthly composites over a small reservoir in a dust-prone basin are frequently unreliable. The reason is the low number of cloud- and dust-free Landsat observations per month compared with Sentinel-2, so that a single contaminated scene dominates the monthly composite. The index spread provides an effective internal quality flag, and screening months with a classifier before merging the two sensors is a simple and transferable remedy.

Third, the two split experiments illustrate a general pitfall of machine learning with environmental time series. A random split produced apparently excellent models, including an LSTM with R^2 above 0.96, while the same models failed when they had to predict a later period. Because the majority of cross-sensor calibration studies use random cross-validation, part of the reported accuracies is likely to be optimistic. A chronological evaluation, as recommended for hydrological models, is the honest test of a transfer function that is intended to extend a record backwards or forwards in time.

The study has limitations. Only one reservoir was analysed, and the labels used for the classifier were defined with a fixed threshold of 5 km²; other thresholds would change the class balance. The number of valid scenes per month, cloud fraction and aerosol optical depth were not available and would very likely improve the classifier. Finally, the Sentinel-2 series was used as the reference, although it is not error-free; validation against in-situ water levels or high-resolution imagery would be desirable.

5. Conclusions

Six spectral water indices and two sensors were compared for the Chah-Nimeh 3 reservoir in Sistan. NDWI, MNDWI and WRI were interchangeable within each sensor, whereas NDMI and TCW behaved differently. Landsat 8/9 monthly areas were on average 10.0 km² smaller than Sentinel-2 areas, with limits of agreement of -31 to 11 km², so the two records cannot be merged directly. A random forest detected unreliable Landsat months from the spread between indices with an AUC of 0.79. Harmonization models performed well under a random split but, except for the LSTM (NSE = 0.30), failed under a chronological split, which shows that random cross-validation strongly overestimates the transferability of such models. The Sentinel-2 record documents a loss of about 72% of the water surface between 2020 and 2023 and a partial recovery in 2024–2025, and it should be the basis of operational monitoring of the reservoir.

6. Data Availability

The monthly index time series analysed in this study were extracted in Google Earth Engine from the Sentinel-2 and Landsat 8/9 surface-reflectance collections and are available from the authors on request, together with the analysis code.

7. References

- Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, 327(8476), 307–310.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794).
- Claverie, M., Ju, J., Masek, J. G., Dungan, J. L., Vermote, E. F., Roger, J.-C., Skakun, S. V., & Justice, C. (2018). The Harmonized Landsat and Sentinel-2 surface reflectance data set. *Remote Sensing of Environment*, 219, 145–161.
- Crist, E. P. (1985). A TM tasseled cap equivalent transformation for reflectance factor data. *Remote Sensing of Environment*, 17(3), 301–306.
- Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., & Vapnik, V. (1997). Support vector regression machines. *Advances in Neural Information Processing Systems*, 9, 155–161.
- Gao, B.-C. (1996). NDWI — A normalized difference water index for remote sensing of vegetation liquid water from space. *Remote Sensing of Environment*, 58(3), 257–266.
- Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 202, 18–27.
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780.

- Kennedy, J., & Eberhart, R. (1995). Particle swarm optimization. In *Proceedings of ICNN'95 – International Conference on Neural Networks* (Vol. 4, pp. 1942–1948). IEEE.
- McFeeters, S. K. (1996). The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features. *International Journal of Remote Sensing*, 17(7), 1425–1432.
- Mirjalili, S., Mirjalili, S. M., & Lewis, A. (2014). Grey wolf optimizer. *Advances in Engineering Software*, 69, 46–61.
- Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I — A discussion of principles. *Journal of Hydrology*, 10(3), 282–290.
- Pekel, J.-F., Cottam, A., Gorelick, N., & Belward, A. S. (2016). High-resolution mapping of global surface water and its long-term changes. *Nature*, 540, 418–422.
- Rashki, A., Kaskaoutis, D. G., Rautenbach, C. J. deW., Eriksson, P. G., Qiang, M., & Gupta, P. (2012). Dust storms and their horizontal dust loading in the Sistan region, Iran. *Aeolian Research*, 5, 51–62.
- Shen, L., & Li, C. (2010). Water body extraction from Landsat ETM+ imagery using adaboost algorithm. In *18th International Conference on Geoinformatics* (pp. 1–4). IEEE.
- Sharifikia, M. (2013). Environmental challenges and drought hazard assessment of Hamoun Desert Lake in Sistan region, Iran, based on the time series of satellite imagery. *Natural Hazards*, 65, 201–217.
- Van Thieu, N., & Mirjalili, S. (2023). MEALPY: An open-source library for latest meta-heuristic algorithms in Python. *Journal of Systems Architecture*, 139, 102871.
- Wang, X., Xie, S., Zhang, X., Chen, C., Guo, H., Du, J., & Duan, Z. (2018). A robust Multi-Band Water Index (MBWI) for automated extraction of surface water from Landsat 8 OLI imagery. *International Journal of Applied Earth Observation and Geoinformation*, 68, 73–91.
- Willmott, C. J. (1981). On the validation of models. *Physical Geography*, 2(2), 184–194.
- Xu, H. (2006). Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery. *International Journal of Remote Sensing*, 27(14), 3025–3033.