Supplementary Material

for

AquaClim-Atlas India (1990–2025): A Geospatial Database of Aquaculture Transitions and Climate Exposure Across Coastal India

List of Supplementary

S1: Database Generation Process Overview and Variable Description

Data Structure

AquaClim is a balanced panel dataset of village-year observations (1990–2025) constructed from harmonized remote sensing, census, and climate data. The dataset is structured at the village level with repeated annual observations.

- Unit of observation: Village × Year
- Spatial unit: Administrative village boundaries
- Temporal coverage: 1990–2025

Data Processing Steps

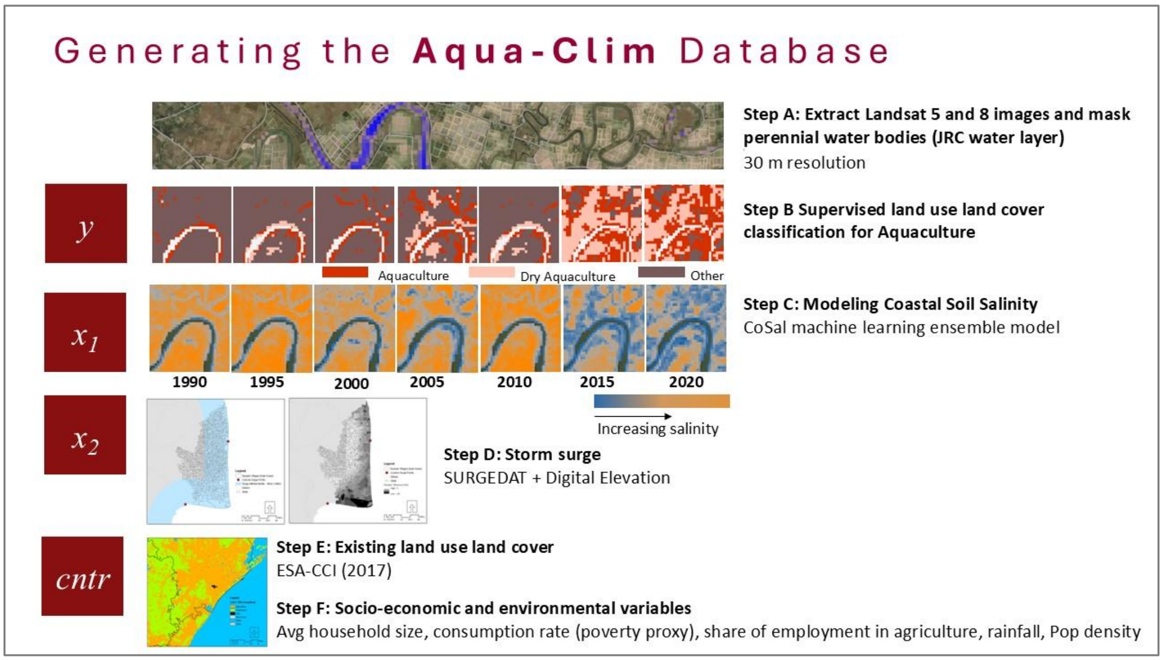


Figure S1: Steps taken to generate the database. We developed a longitudinal panel dataset at the village-level, covering all 31,088 coastal villages across all 27 coastal districts of the three states (6 in Odisha, 9 in Andhra Pradesh, and 12 in Tamil Nadu). This dataset integrates: (A) Remote-sensing-derived supervised land use classification of aquaculture using Landsat 5 (1990-2012) and Landsat 8 (2013-2025) satellite images; (B) Environmental stress modeling of salinity change, using the machine-learning and remote-sensing based CoSal-SA model (Jain et al. 2025); (C) Environmental shock impact of storm surges using SURGEDAT data base that documents all surge events from 1880-2015 (Keim 2015); (D) Geospatial attributes of villages, such as elevation and distance from the sea using the 2001 Census of India Boundaries (Census of India 2001) and Copernicus Digital Elevation Model GLO-30m (Copernicus 2013); (E) Rainfall using 100 year 0.25 x 0.25 Gridded data from Indian Meteorological Department (IMD, Climate Monitoring and Prediction Group 2024); (F) Population density using Global Human Settlement data (European Commission 2016); and (G) Land use/land cover of agricultural and urban lands using Copernicus Global Land Cover Layers: CGLS-LC100 Collection 3 for 2015 (Buchhorn et al. 2020), with the consideration that the former land use is more likely to be converted than the latter.

Aquaculture Variables
Step 1: Landsat Composites for all districts for all years (1990-2025) in Google Earth Engine
Step 2: Mask permanent water (Google earth engine)
Step 3: Conduct supervised land use land cover classification for two time periods (Landsat 5 period and Landsat 8 period) and export GeoTiffs from GEE.
Step 4: Overlay village boundaries and extract area under active and dry aquaculture and other land uses in ArcGIS Pro (ArcPy Code)

1. Aqua_ha (absolute land area classified as aquaculture in hectares)
2. DryAqua_ha (absolute land area classified as fallow aquaculture / abandoned or dry aquaculture ponds in hectares)
3. Aqua_perc (% of village area classified as aquaculture)
4. DryAqua_perc (% of village area classified as dry aquaculture pond)
5. pct_change_aqua (percentage change in aquaculture land area from last recorded year).
6. pct_change_dryaqua (percentage change in dry aquaculture land area from last recorded year).

Soil Salinity Variables
Step 1: Landsat composites for all years (1990-2025)
Step 2: Harmonize all Landsat 5 composite bands with 2025 Landsat 8 composite bands (Python code)
Step 3: Export multi-band GeoTiffs from GEE to R
Step 4: Mask Aquaculture Pixels, Run CoSal model on each GeoTIFF and exposrt Salinity GeoTiffs in R
Step 5: Overlay village boundaries on GeoTIFF and calculate % high salibity for each year in ArcGIS Pro (ArcPy codes)

1. Saline_ha (absolute land area in hectare above 1900 μs/cm electrical conductivity)
2. Saline_perc_norm (percentage village land area above 1900 μs/cm electrical conductivity – normalized for the two time periods following a discontinuity analysis).
3. avg_salinity_5yr (average of 5 years of percentage area above 1900 μs/cm electrical conductivity – measure of peristsent salinity)
4. pct_change_salinity (percentage change in land area above 1900 μs/cm electrical conductivity from last recorded year).

Storm exposure variables
Step 1: Extract all SURGEDAT point geolocations, dates (starting 1885), and surge height for coastal areas of interest
Step 2: Extract Digital Elevation Model from GEE to ArcPro
Step 3: Offset each surge geolocation based on the depth of surge and width of surge (coastal length) based on SURGEDAT or EMDAT or other secondary sources of information.
Step 4: Identify villages within the offset areas from the cyclone eye and any point in the village below the maximum surge height (in ArcGIS pro).

1. first_surge_year (first year when a village was affected by a storm surge; if never then NA)
2. postSurge (cumulative binary variable for storm surge: remains 0 until affected, turns 1 after being affected in a year by a storm surge).
3. Saline_Storm_Category (combined categorical variable with salinity levels threshold at median Salinity_perc_norm across all villages in that year):
    - high salinity storm affected
    - low salinity storm affected
    - high salinity no storm
    - low salinity no storm

Rainfall
Step 1: Total and average annual rainfall in a village is estimated using the 100 year 0.25 x 0.25 Gridded rainfall data from the Indian Meteorological Department in R.

- Rainfall_mm (average annual rainfall in a village)

Geophysical variables
Step 1: Estimate the distance of the centroid of the village from the closest point on the coastline, and categorise the villages based on this distance, as near or far from the coast.
Step 2: Digital elevations of all pixels falling within the village boundary are averaged and an average elevation from the sea is estimated and assigned.

- NEAR_DIST (Distance of the village centroid from the sea in meters)
- Sea_Dist (categorised by very near, near, medium, far, very far from the sea, based on quantiles)
- DEM_avg (average elevation from the sea calculated by averaging all digital elevation values of all pixels within a village)

Land-use contextual variables
Step 1: Extract land use land cover from ESA-CCI (2017) and export Geotiff from GEE
Step 2: Overlay with village boundaries and calculate hectares and % area under agriculture and urban (Note: These are, at present, time invariant variables, however, can be made time varying using other relevant available datasets. Other land uses such as forest etc. could also be included although at this stage those have not been included).

- agriculture_area_ha (area in hectares under agriculture in the base year)
- agriculture_percent (percentage of village area under agriculture in the base year)
- urban_area_ha (urbanised area in hectares in the base year)
- urban_percent (urbanised area percentage in the base year)

Population / socioeconomic variables
Step 1: Population estimates were obtained from the Global Human Settlement Layer (GHSL) population grids at 5-year intervals between 1990 and 2025.
Step 2: Population counts were extracted by overlaying village boundaries on gridded population rasters using area-integrated zonal aggregation, producing village-level population totals (Population) for each time step.
Step 3: Population values were interpolated between available time points using linear interpolation. This resulted in a complete annual village-level population time series for 1990–2025.

- Population
- No_HH (number of households)

Variable Classification
Time-varying variables
Includes annually varying environmental and socioeconomic indicators:

- Aquaculture extent (Aqua_ha, Aqua_perc)
- Salinity exposure (Saline_ha, Saline_perc_norm)
- Climate variables (Rainfall_mm)
- Lagged and smoothed indicators
- Population

Time-invariant variables
Structural geographic and administrative covariates:

- Elevation (DEM_avg)
- Distance to coast (NEAR_DIST)
- Total village area (total_area_ha)
- Administrative identifiers (State, District)

Variable Definitions (Codebook)

variable	description	unit	type	variability
Aqua_ha	Area under aquaculture ponds (hectares)	hectares	continuous	Time-varying
DryAqua_ha	Area under dry aquaculture ponds (hectares)	hectares	continuous	Time-varying
Aqua_perc	Aquaculture area as percentage of village area	percent	percentage	Time-varying
DryAqua_perc	Abandoned aquaculture area as percentage of village area	percent	percentage	Time-varying
pct_change_aqua	percentage change in aquaculture land area from last recorded year	percent	percentage	Time-varying
pct_change_dryaqua	percentage change in dry aquaculture land area from last recorded year	percent	percentage	Time-varying
Saline_ha	Area affected by salinity (>1900 µs/cm) (hectares)	hectares	continuous	Time-varying
Saline_perc	Percentage area affected by salinity (>1900 µs/cm)	percent	percentage	Time-varying
Saline_perc_norm	Normalized salinity index (standardized)	index	index	Time-varying
avg_salinity_5yr	5-year rolling mean salinity index (average of 5 years of percentage area above 1900 µs/cm electrical conductivity – measure of persistent salinity)	index	rolling index	Time-varying
pct_change_salinity	percentage change in land area above 1900 µs/cm electrical conductivity from last recorded year	index	rolling index	Time-varying
first_surge_year	first year when a village was affected by a storm surge; if never then NA	year	categorical	Static
postSurge	cumulative binary variable for storm surge: remains 0 until affected, turns 1 after being affected in a year by a storm surge		categorical	Time-varying
Saline_Storm_Category	combined categorical variable with salinity levels threshold at median Salinity_perc_norm across all villages in that year: ○ high salinity storm affected ○ low salinity storm affected ○ high salinity no storm ○ low salinity no storm	index	categorical	Time-varying
Rainfall_mm	Annual rainfall (mm)	mm	continuous	Time-varying
DEM_avg	Mean elevation of village (meters)	meters	continuous	Static
NEAR_DIST	Distance to nearest coastline or reference point (meters)	meters	continuous	Static
Sea_Dist	Distance from the sea categorised by very near, near, medium, far, very far from the sea	index	categorical	Static

agriculture_area_ha	area in hectares under agriculture in the base year	hectares	continuous	Static
agriculture_percent	percentage of village area under agriculture in the base year	percent	percentage	Static
urban_area_ha	urbanised area in hectares in the base year	hectares	continuous	Static
urban_percent	urbanised area percentage in the base year	percent	percentage	Static
Population	Total population of village	persons	count	Static

Time varying variables

variable	**mean**	**sd**	**min**	**max**	**miss**
Year	2007.5	10.3883	1990	2025	0
Aqua_ha	8.897876	68.43224	0	6154.65	20.50023
DryAqua_ha	24.36761	93.37626	0	10517.76	20.50023
Saline_ha	24.59798	109.3543	0	23747.29	21.44075
Aqua_perc	1.076198	4.639457	0	99.48478	20.82073
DryAqua_perc	5.674983	10.9938	0	99.98207	20.6171
Saline_perc_norm	1.10E-16	0.999999	-0.60356	12.20067	21.68995
avg_salinity_5yr	-0.00372	0.853894	-0.60356	9.703772	4.367173
Lag_Aqua	1.048464	4.540065	0	99.36171	23.59762
Lag_Saline	-0.00486	0.990984	-0.60356	12.20067	24.46719
pct_change_aqua	-6.28581	117.0966	-100	3880.952	89.99873
pct_change_dryaqua	40.97687	259.9173	-100	9000	67.86398
pct_change_salinity	41.84599	308.8378	-100	8265.82	71.80888
Rainfall_mm	1297.359	425.478	14.73088	3097.946	3.972594
Population	2336.493	13745.58	0	1102339	0

Time invariant variables

variable	mean	sd	min	max	miss
total_area_ha	455.498	719.9925	1.043787	35470.97	0
Shape_Area	4581463	7323124	10437.87	3.55E+08	0
DEM_avg	66.68106	247.2073	-9999	1292.092	0
NEAR_DIST	35788.09	25742.91	19.37926	144333.8	0
agriculture_percent	70.22813	30.06526	0	99.99968	7.50772
urban_percent	2.63338	6.837364	0	98.46622	0.003217

Data Construction Notes

- All raster-derived variables were aggregated to village polygons using zonal statistics.
- Percentage variables are normalized by total village area.
- Lagged variables are constructed using within-village temporal shifts.

Missing Data

Missingness is generally low and arises from:

- Remote sensing cloud contamination
- Boundary mismatches in early years
- Sparse census covariates in select districts

S2: Aquaculture land-use classification

Data Sources

The AquaClim-Atlas draws on the atmospherically corrected US Geological Survey Landsat 5 TM (1990-2012) and Landsat 8 OLI (Level 2 Collection 2 Tier 1) (2012-2025) satellite imagery for the land use land cover classification as well as modeling soil salinity).[35,36] All Landsat images are 30 m spatial resolution. Images were limited to no more than 50 percent cloud cover to maximize the number of images available while reducing reflectance errors due to potential presence of clouds.

Higher resolution harmonized Sentinel-2 Multi Spectral Instrument (MSI) Level-2A [37] satellite image composites for February 2024 were used as reference images to identfy aquaculture locations, utilized for training and validating the classification model on Landsat 8 OLI composites. Similarly, Google Earth Archive high resolution satellite imagery was used to extract land-use and land-cover observations for 13 December 2011 to train Landsat 5 TM satellite imagery for the corresponding dates. Google Earth engine (Gorelick et al. 2017) is used for data extraction and ASU Sol supercomputer infrastructure for data processing (Jennewein et al. 2023).

Geolocated field observation points for active aquaculture and dry aquaculture were collected in the month of February 2024 from district Jagatsinghpur in Odisha. Active aquaculture ponds are those that are currently inundated and used for production, whereas, dry or abandoned aquaculture are ponds that have been drained, fallowed, or permanently abandoned but still retain their physical footprnt. These geolocations were used as reference points, to identify 274 active aquaculture points, 220 dry or abandoned aquaculture locations, and 391 non-aquaculture or other uses for Feb 2024 period on a high-resolution Sentinel-2 reference image. Additionally, 238 aquaculture points, 180 dry aquaculture points, and 405 other land uses were identified from Google Earth Archive imagery (December 2011) to train the supervised land use land cover classification in the Landsat 5 period.

Image processing and classification

To conduct a supervised land-use and land-cover classification of aquaculture, the following steps were taken to prepare and process the satellite data and field observations.

First, image composites were created for the various Landsat satellite image collections, to filter out clouds, haze, and high reflectance noise while retaining meaningful surface reflectance without overemphasizing dark shadows. Alternate pixel reducers were tested to create these composites, including geomedian, median, and 30th and 70th percentile (See S3 for accuracy comparisons). Since the geomedian and 30th percentile resulted in similar and high accuracy amongst these, the geomedian reducer was used for the current purpose. Geomedian reducer was preferred over percentile for three reasons (Roberts, Mueller, and Mcintyre 2017). (1) It enables stacking multiple images of varying quality to a high-quality pixel composite while reducing spatial noise. (2) In contrast to a median reducer, a geomedian also helps maintain spectral relationship between bands and consistency across scene boundaries. (3) It is a recommended method to maintain spectral relationships between bands when further analysis on the composite image is required, such as applying a machine learning algorithm. Geomedian reducer was also adopted to qualify using the machine-learning based model developed by Perikamana et al. (2025) to harmonize Landsat 5 TM images with Landsat 8 OLI in the subsequent steps for salinity modelling.

As is required by the Landsat dataset, pixel values were scaled to floating points (0-1) using published gain and offset values [45,46]. Some pixels (0.4 percent of all pixels) had negative surface reflectance after conversion, which is likely due to cloud cover or presence of water [47]. These values were adjusted to 0 to not get included in the analysis further.

Second, perennial surface water pixels were masked in the Landsat composites using the JRC Global Surface Water Mapping Layers, v1.4 [48]. A threshold of 90 percent occurrence of water was used, to not mix perennial surface water with aquaculture ponds, which do not operate throughout the year.

Finally, to identify aquaculture land uses, two separate supervised classification models were trained for the two satellite sensors - Landsat 8 OLI (2013-2025) and Landsat 5 TM (1990-2012). In total, observation data for three classes was used - active aquaculture (274 points for Landsat 8 OLI, 238 for Landsat 5 TM) dry aquaculture (220 for OLI and 180 for TM) and other land uses (391 for OLI and 405 for TM). Of these, 80% points were used for training, and 20% left for validation.

For the classification, a random forest classifier was employed, with 65 trees for Landsat 8 classifier and 40 trees for the Landsat 5 classifier (see S4 for hyperparameter tuning results to minimize classification error) and employed 6 spectral bands (Blue, Green, Red, Near Infrared or NIR, and the two Short Wave Infrared bands SWIR1 and SWIR2) and three indices popularly used for surface water and salinity detection (Normalized Difference Vegetation Index, Normalized Difference Water Index, and Normalized Difference Salinity Index). The supervised classification algorithm was implemented in Google Earth Engine.

Annual aggregation and variable generation

Once classified, area under aquaculture, dry aquaculture and other uses is calculated for each year between 1990-2025 for each village boundary. Since this database integrates data using two distinct Landsat satellite sources (Landsat 5 and 8), a discontinuity assessment was undertaken to ensure there are no residual differences due to satellite sensor differences (See S21). Aquaculture land use change showed no statistically significant discontinuity over the two periods, and therefore, the absolute measure of percentage area is used without normalization.

S3: Comparing accuracy outputs with alternate pixel reducers

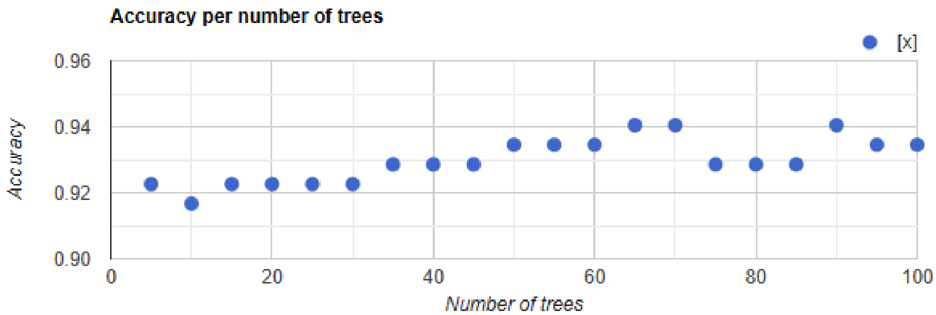
Table S1: Comparing accuracy outputs with alternate pixel reducers. The accuracy results with the 30th percentile (listed in the previous section in more detail) are the best among the various reducers tested, closely followed by the geometric median. For the CoSal-SA model, we have employed the geomedian. We preferred geomedian reducer over percentile for three reasons (Roberts, Mueller, and Mcintyre 2017): (1) it enables stacking multiple images of varying quality to a high-quality pixel composite while reducing spatial noise, (2) in contrast to a median reducer, a geomedian also helps maintain spectral relationship between bands and consistency across scene boundaries, and (3) it is a recommended method to maintain spectral relationships between bands when further analysis on the composite image is required, such as applying a machine learning algorithm. We have also adopted geomedian reducer to qualify using the machine-learning based model developed by Perikamana et al. (Perikamana, K.K., Parath, N., & Balakrishnan, K. 2025) to harmonise Landsat 5 TM images with Landsat 8 OLI (Jennewein et al. 2023).

Accuracy measure	**Percentile (30)**	**Geometric median (adopted)**	**Median/Percentile (50)**	**Percentile (70)**
Overall accuracy	0.96	0.94	0.92	0.90
Kappa	0.93	0.91	0.88	0.84
Producers accuracy	0: [0.95] 1: [0.95] 2: [0.98]	0: [0.95] 1: [0.93] 2: [0.94]	0: [0.97] 1: [0.85] 2: [0.90]	0: [0.95] 1: [0.80] 2: [0.90]
Consumers accuracy	0: [0.96] 1: [0.91] 2: [1]	0: [0.95] 1: [0.93] 2: [0.94]	0: [0.92] 1: [0.85] 2: [0.98]	0: [0.91] 1: [0.86] 2: [0.91]

S4: Remote-sensing-derived supervised land use classification of aquaculture using Landsat 5 (1990-2012) and Landsat 8 (2013-2025) satellite images.

Purpose: To create a random forest classifier to identify active aquaculture ponds, dry aquaculture ponds, and other land uses on a Landsat Satellite imagery. We use field observations and high-resolution Sentinel 2 imagery for reference data. This supplementary includes the accuracy results of the supervised classification.
Code: (Google Earth Engine) (Gorelick et al. 2017) - 1_LULCClassification_Code.txt

Landsat 8
Hyperparameter tuning for Random Forest Classifier
Selected 65 trees



Confusion matrix:
[68,1,0],[5,49,1],[3,1,40]]
0: [68,1,0]
1: [5,49,1]
2: [3,1,40]

Overall Accuracy:
0.9345238095238095

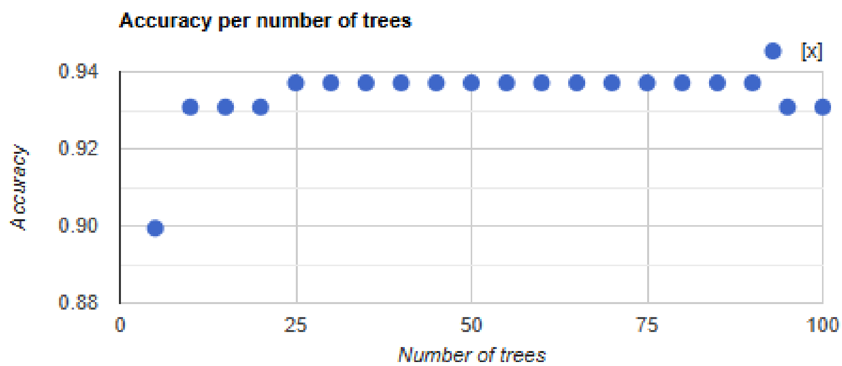
Producers Accuracy:
0: [0.9855072463768116]
1: [0.8909090909090909]
2: [0.9090909090909091]

Consumers Accuracy:
0: 0.8947368421052632
1: 0.9607843137254902
2: 0.975609756097561

Kappa:
0.8994066735615917

Figure S2: Accuracy results of remote-sensing-derived supervised land use classification of aquaculture using Landsat 8 (2013-2025) satellite images.

Landsat 5
Hyperparameter tuning for Random Forest Classifier
Selected 40 trees



Confusion matrix:
[[72,0,0],[4,29,5],[0,1,48]]
0: [72,0,0]
1: [4,29,5]
2: [0,1,48]

Overall Accuracy:
0.9371069182389937

Producers Accuracy:
[[1],[0.7631578947368421],[0.9795918367346939]]
0: [1]
1: [0.7631578947368421]
2: [0.9795918367346939]

Consumers Accuracy:
0: 0.9473684210526315
1: 0.9666666666666667
2: 0.9056603773584906

Kappa:
0.9010701841712295

Figure S3: Accuracy Results of remote-sensing-derived supervised land use classification of aquaculture using Landsat 5 (1990-2012) satellite images.

S5: Soil Salinity Modelling

Data sources

Geolocated primary soil samples (n=113) were collected distributed relatively evenly across 200 sq. km. of a predetermined sampling area were collected in the dry month of February in 2024 to avoid cloud cover, a common deterrent in remote sensing studies especially in tropical contexts. Landsat image composites generated for the aquaculture land use land cover classification were also used for modeling soil salinity.[35,36] Detail description of the coastal salinity model employed here, CoSal, including the sampling area, data collection and testing methodology are described in Jain et al.(Jain et al. 2025)

Data processing and variable generation

Unlike the aquaculture land-use and land-cover classification where observations from the older time periods could be captured using higher resolution images, the same was not possible for soil samples. Hence, before applying the CoSal model for the distinct Landsat periods, the two Landsat sensors (TM and OLI) needed to be harmonized.

Harmonization between Landsat 5 TM and Landsat 8 OLI is complicated by the fact that there is no temporal overlap between these sensors. Both Landsat 5TM and Landsat 8 OLI have temporal overlap with Landsat 7 ETM+, however. During their period of temporal overlap, both Landsat 5 and 7 and Landsat 7 and 8 orbited the earth with a gap of approximately 8 days.

Several researchers have demonstrated linear regression-based approaches for harmonizing across Landsat sensors and across Landsat and Sentinel sensors (e.g.: Roy et al., (2016); Shang and Zhu, (2019)). Roy et al. (2016), provide regression coefficients for harmonizing Landsat 7 ETM+ with Landsat 8 OLI data. This approach is based on OLS and RMA regression of each band of Landsat 7 ETM+ with corresponding band from Landsat 8 OLI data. Such an approach, however, is unable to account for non-linearities in band-to-band relationships across sensors. Since each band pair is treated separately, it also cannot leverage information from other bands in the harmonization process. Nevertheless, the Roy et al. (2016) approach is often used to harmonize Landsat 5TM with Landsat 8 OLI also (Suharyadi et al. 2022).

Instead of a linear regression-based approach, this Atlas employs Random Forest (RF) models to harmonize Landsat 5 TM data with Landsat 8 OLI data since RF models can account for non-linear relationships. A two-step harmonization process is undertaken, using two separate sets of RF models. The first set of RF models take six Landsat 5TM bands as input and predict corresponding Landsat 7 ETM+ bands as output. The second set of RF models take these predicted Landsat 5 ETM+ outputs and predict the corresponding Landsat 8 OLI bands (See S6 for the details on the two-step harmonization that uses a random-forest model along with the accuracy details (Perikamana, K.K., Parath, N., & Balakrishnan, K. 2025)). This step is done with a high accuracy between 0.99 to 0.94 for the six bands between L5 TM to L7 ETM+, and 0.99 to 0.91 for the six bands between L8 OLI to L7 ETM+. Arizona State University's Sol Supercomputing was used for conducting the harmonization.(Jennewein et al. 2023)

Following the harmonization, all surface water is masked. Perennial surface water is masked using the JRC layer as before, and the water-based land use of aquaculture is masked using the previous step output of aquaculture land use land cover classification. This is done to avoid confounding of surface reflectance of soil salinity with water. Once masked, the CoSal Ensemble Model (Jain et al. 2025) is applied on all the harmonized Landsat composites for the 31,088 villages. This is done using an R code (v. 4.3.1).

Annual aggregation and variable generation

Once classified, areas identified with salinity levels greater than 1,900 μS/cm—soil salinity threshold at which rice productivity is found to be impacted (see S8 for more details)—are aggregated for all 31,088 villages for each year between 1990-2025. Salinity estimates show variations in a discontinuity analysis (see S19) and therefore are normalized for the two time periods (1990-2012 and 2013-2025) as the z-score (i.e. number of standard deviations away from the mean).

S6: Summary of Landsat 5 and Landsat 8 harmonization

Cite: Perikamana, K.K., Parath, N., & Balakrishnan, K. (2026). *An improved method for harmonizing Landsat 5-TM and Landsat 8-OLI data.* Manuscript in preparation.

Several researchers have demonstrated linear regression-based approaches for harmonizing across Landsat sensors and across Landsat and Sentinel sensors (eg: Roy et al., (2016); Shang and Zhu, (2019)). Roy et al. (2016), provides regression coefficients for harmonizing Landsat 7 ETM+ with Landsat 8 OLI data. This is based on OLS and RMA regression of each band of Landsat 7 ETM+ with corresponding band from Landsat 8 OLI data. But such an approach is unable to account for non-linearities in band-to-band relationships across sensors. Since each band pair is treated separately, it also cannot leverage information from other bands in the harmonization process. Nevertheless, the Roy et al. (2016) approach has been used to harmonize Landsat 5TM with Landsat 8 OLI also (Suharyadi et al. 2022).

Instead of a linear regression-based approach, we use Random Forest (RF) models to harmonize Landsat 5 TM data with Landsat 8 OLI data since RF models can account for non-linear relationships. Harmonization between Landsat 5 TM and Landsat 8 OLI is complicated by the fact that there is no temporal overlap between these sensors. But both Landsat 5TM and Landsat 8 OLI have temporal overlap with Landsat 7 ETM+. During their period of temporal overlap, both Landsat 5 and 7 and Landsat 7 and 8 orbited the earth with a gap of approximately 8 days.

Based on this, we do a two-step harmonization using two separate sets of RF models. The first set of RF models take six Landsat 5TM bands as input and predict corresponding Landsat 7 ETM+ bands as output. The second set of RF models take these predicted Landsat 5 ETM+ outputs and predict the corresponding Landsat 8 OLI bands.

In each harmonization step, each RF model takes all the bands from the input sensor and predicts one band of the output sensor. In this way, we use all the available information across input sensor bands rather than relying on information only from one band like in the case of Roy et al. (2016). Therefore, a total of 12 RF models were prepared such that the six Landsat 5 TM bands can be converted into corresponding six Landsat 7 ETM+ bands and further converted to the corresponding six Landsat 8 OLI bands.

Our original focus while training the RF models was to analyze urban areas in India. Therefore, our training data consisted of Landsat scenes across the largest 100 cities in India based on contiguous built-up area.

Based on cloud cover and quality assessment band thresholds, the best 25 Landsat 5 and 7 pairs which were 8 days apart were selected. These scenes were from the period from 1 January 1999 to 1 May 2003. This ensured that the Landsat 7 ETM+ scenes do not have the scanline corrector error which starts from 31 May 2003.

Similarly, 25 best Landsat 7 ETM+ and Landsat 8 OLI scenes were selected from the period from 1 January 2013 to 1 January 2017. The Landsat 7 ETM+ data from this period has the scanline corrector error. The cells with unusable data were therefore set to ‘No Data’.

Each Landsat 7 ETM+ scene was reprojected to the UTM zone in which majority of the scene fell. After this each corresponding Landsat 5 TM scene and Landsat 8 OLI scene were reprojected with the Landsat 7 ETM+ scene as reference so that cell alignment is ensured.

For each harmonization step, we trained and validated the model with 100million cells from 22 pairs of scenes and used 3 scenes for testing.

The accuracy metrics (average across three test scenes) for the L5 TM to L7 ETM+ harmonization is given in the table below:

Metric	Band 1	Band 2	Band 3	Band 4	Band 5	Band 6
RMSE	163.20	158.77	164.99	171.23	167.12	171.40
R^2	0.94	0.97	0.98	0.98	0.99	0.99
SSIM	0.90	0.93	0.94	0.94	0.94	0.95

The accuracy metrics (average across three test scenes) for the L7 ETM+ to L8 OLI harmonization step is given in the table below:

Metric	Band 1	Band 2	Band 3	Band 4	Band 5	Band 6
RMSE	145.01	146.47	152.34	172.08	167.78	166.26
R^2	0.91	0.94	0.97	0.99	0.99	0.99
SSIM	0.94	0.93	0.92	0.90	0.91	0.94

We used Arizona State University's Sol Supercomputing for conducting the harmonization (Jennewein et al. 2023).

Key references

Roy, D. P., V. Kovalskyy, H. K. Zhang, E. F. Vermote, L. Yan, S. S. Kumar, and A. Egorov. 2016. "Characterization of Landsat-7 to Landsat-8 Reflective Wavelength and Normalized Difference Vegetation Index Continuity." *Remote Sensing of Environment* 185 (Iss 1): 57–70.

Shang, Rong, and Zhe Zhu. 2019. "Harmonizing Landsat 8 and Sentinel-2: A Time-Series-Based Reflectance Adjustment Approach." *Remote Sensing of Environment* 235 (111439): 111439.

Suharyadi, R., Deha Agus Umarhadi, Disyacitta Awanda, and Wirastuti Widyatmanti. 2022. "Exploring Built-up Indices and Machine Learning Regressions for Multi-Temporal Building Density Monitoring Based on Landsat Series." *Sensors (Basel, Switzerland)* 22 (13): 4716.

S7: Study areas for soil sampling and aquaculture land use training data

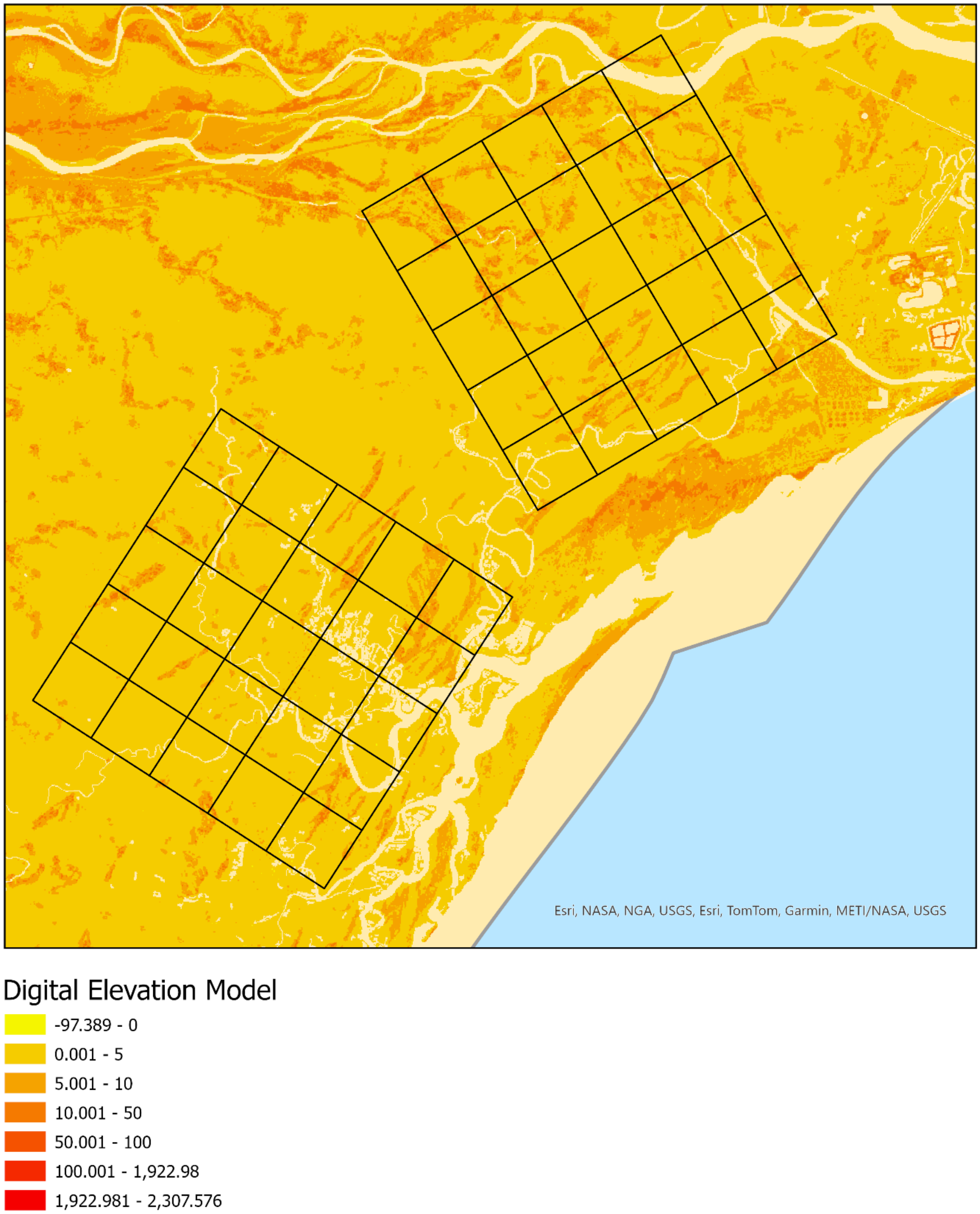


Figure S4: Digital Elevation Model (Source: Copernicus DEM GLO-30. Global 30m Digital Elevation Model) of the area shows similar conditions, such that the entire area is below 10m from the mean sea level. Both sub-areas are intersected by one continuous perennial river, as well as other large and small perennial water bodies. This makes both areas conducive to aquaculture.

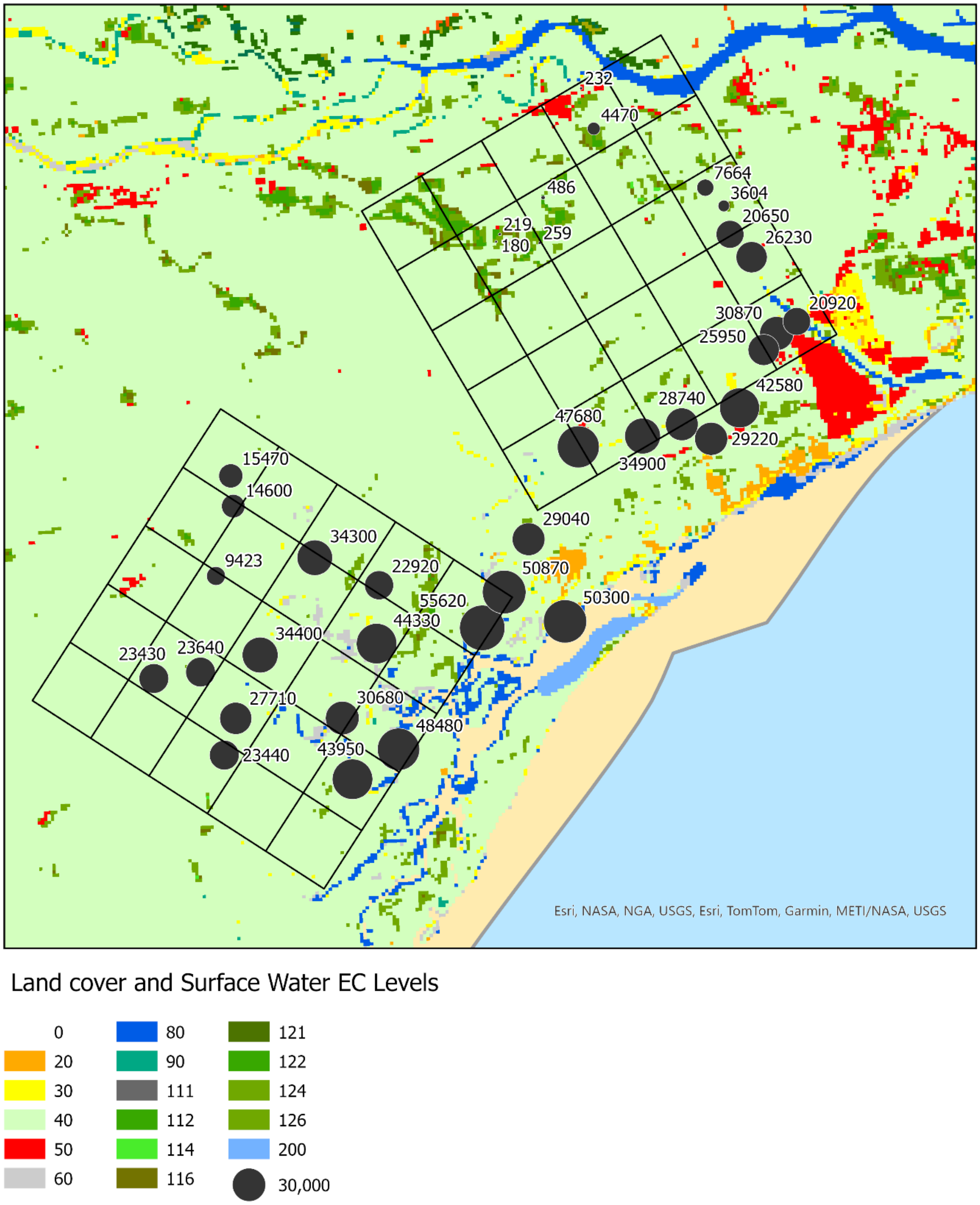


Figure S5: Land Cover layer. (Source: European Space Agency - Climate Change Initiative) shows the various land uses in the area. The two regions are dominantly classified agriculture and open fields, with some urban and forested areas. The map also shows the EC of surface water samples we collected in both regions, showing similar water quality.

Class Table (discrete_classification)

Value	Color	Description
0	#282828	Unknown. No or not enough satellite data available.
20	#ffbb22	Shrubs. Woody perennial plants with persistent and woody stems and without any defined main stem being less than 5 m tall. The shrub foliage can be either evergreen or deciduous.
30	#ffff4c	Herbaceous vegetation. Plants without persistent stem or shoots above ground and lacking definite firm structure. Tree and shrub cover is less than 10 %.
40	#f096ff	Cultivated and managed vegetation / agriculture. Lands covered with temporary crops followed by harvest and a bare soil period (e.g., single and multiple cropping systems). Note that perennial woody crops will be classified as the appropriate forest or shrub land cover type.
50	#fa0000	Urban / built up. Land covered by buildings and other man-made structures.
60	#b4b4b4	Bare / sparse vegetation. Lands with exposed soil, sand, or rocks and never has more than 10 % vegetated cover during any time of the year.
70	#f0f0f0	Snow and ice. Lands under snow or ice cover throughout the year.
80	#0032c8	Permanent water bodies. Lakes, reservoirs, and rivers. Can be either fresh or salt-water bodies.
90	#0096a0	Herbaceous wetland. Lands with a permanent mixture of water and herbaceous or woody vegetation. The vegetation can be present in either salt, brackish, or fresh water.
100	#fae6a0	Moss and lichen.
111	#58481f	Closed forest, evergreen needle leaf. Tree canopy >70 %, almost all needle leaf trees remain green all year. Canopy is never without green foliage.
112	#009900	Closed forest, evergreen broad leaf. Tree canopy >70 %, almost all broadleaf trees remain green year round. Canopy is never without green foliage.
113	#70663e	Closed forest, deciduous needle leaf. Tree canopy >70 %, consists of seasonal needle leaf tree communities with an annual cycle of leaf-on and leaf-off periods.
114	#00cc00	Closed forest, deciduous broad leaf. Tree canopy >70 %, consists of seasonal broadleaf tree communities with an annual cycle of leaf-on and leaf-off periods.
115	#4e751f	Closed forest, mixed.
116	#007800	Closed forest, not matching any of the other definitions.
121	#666000	Open forest, evergreen needle leaf. Top layer- trees 15-70 % and second layer- mixed of shrubs and grassland, almost all needle leaf trees remain green all year. Canopy is never without green foliage.
122	#8db400	Open forest, evergreen broad leaf. Top layer- trees 15-70 % and second layer- mixed of shrubs and grassland, almost all broadleaf trees remain green year-round. Canopy is never without green foliage.
123	#8d7400	Open forest, deciduous needle leaf. Top layer- trees 15-70 % and second layer- mixed of shrubs and grassland, consists of seasonal needle leaf tree communities with an annual cycle of leaf-on and leaf-off periods.

124	#a0dc00	Open forest, deciduous broad leaf. Top layer- trees 15-70 % and second layer- mixed of shrubs and grassland, consists of seasonal broadleaf tree communities with an annual cycle of leaf-on and leaf-off periods.
125	#929900	Open forest, mixed.
126	#648c00	Open forest, not matching any of the other definitions.
200	#000080	Oceans, seas. Can be either fresh or salt-water bodies.

S8: Surface Reflectance measures of 28 Parameters and Indices

(1) 6 Surface Reflectance bands – Blue, Green, Red, NIR, SWIR1, SWIR2

(2) 11 Normalized Band Combinations
Blue and Red: NBR <- (Blue-Red)/(Blue+Red)
Blue and green: NBG <- (Blue-Green)/(Blue+Green)
Blue and NIR: NBNIR <- (Blue-NIR)/(Blue+NIR)
Blue and SWIR1: NBSWIR1 <- (Blue-SWIR1)/(Blue+SWIR1)
Blue and SWIR2: NBSWIR2 <- (Blue-SWIR2)/(Blue+SWIR2)
Red and SWIR1: NRSWIR1 <- (Red-SWIR1)/(Red+SWIR1)
Red and SWIR2: NRSWIR2 <- (Red-SWIR2)/(Red+SWIR2)
Green and SWIR1: NGSWIR1 <- (Green-SWIR1)/(Green+SWIR1)
Green and SWIR2: NGSWIR2 <- (Green-SWIR2)/(Green+SWIR2)
NIR and SWIR1: NNIRSWIR1 <- (NIR-SWIR1)/(NIR+SWIR1)
NIR and SWIR2: NNIRSWIR2 <- (NIR-SWIR2)/(NIR+SWIR2)

(3) 11 Widely used salinity and other indices
# Salinity Index 1 = sqrt(green^2+red^2)
SI1 <- sqrt((Green)^2 + (Red)^2)
Sources: (Douaoui, Nicolas, and Walter 2006; Nguyen et al. 2020)

# Salinity Index 2 = sqrt(green x red)
SI2 <- sqrt(Green * Red)
Sources: (Douaoui, Nicolas, and Walter 2006; Nguyen et al. 2020)

# Salinity Index 3 = sqrt(blue x red)
SI3 <- sqrt(Blue * Red)
Source: (Khan et al. 2001; Nguyen et al. 2020)

# Salinity index 4 = red x NIR / green
SI4 <- (Red * NIR / Green)
Source: (Abbas and Khan 2007; Abbas et al. 2013; Nguyen et al. 2020)

# Salinity index 5 = blue/red
SI5 <- (Blue / Red)
Source: (Abbas and Khan 2007; Abbas et al. 2013; Nguyen et al. 2020)

# Soil Adjusted Vegetation Index (SAVI) = ((1.5)x NIR) - (red/0.5) + NIR + Red
SAVI <- (1.5 * NIR) - (0.5 * Red) + NIR + Red
Source: (Alhammadi and Glenn 2008; Nguyen et al. 2020)

# Vegetation Soil Salinity Index (VSSI) = (2 x green) - 5 x (red + NIR)
VSSI <- (2 * Green) - 5 * (Red + NIR)
Source: (Dehni and Lounis 2012; Nguyen et al. 2020)

# Normalised Difference Salinity Index 1 (NDSI)
NDSI1 <- (SWIR1-SWIR2)/(SWIR1+SWIR2)
Source: (Al-Khaier et al. 2003; Henrich et al. 2012)

# Normalised Difference Salinity Index 2 (NDSI2)
NDSI2 <- (Red-NIR)/(Red+NIR)
Source: (Khan et al. 2001; Nguyen et al. 2020)

# Normalised Difference Vegetation Index (NDVI) or Green-Red Vegetation Index (GRVI)
NDVI <- (Red-Green)/(Red+Green)
Source: (Li et al. 2023; Wang et al. 2007)

# NDWI
NDWI <- (Green-NIR)/(Green+NIR)
Source: (Yousefian et al. 2019; McFEETERS 1996)

S9: Regression model fit results with EC

Table S2: Regression coefficients of univariate relationships between EC and various parameters starting with the highest explanatory powers

Band/Index	Fit	β_0	SE (β_0)	β_1	SE(β_1)	β2	SE (β2)	R2	Adj R2	AIC	BIC	RMSE
NDWI	Quadratic	0.277***	(0.038)	2.275***	(0.347)	1.342***	(0.347)	0.419	0.405	65.0	74.6	0.34
NDSI2	Quadratic	0.277***	(0.040)	2.010***	(0.361)	1.477***	(0.361)	0.374	0.359	71.2	80.9	0.35
NRSWIR1	Quadratic	0.277***	(0.041)	2.162***	(0.373)	0.894**	(0.373)	0.329	0.312	76.9	86.6	0.37
NDWI	Linear	1.641***	(0.229)	3.333***	(0.551)			0.311	0.303	77.2	84.4	0.37
NBNIR	Quadratic	0.277***	(0.042)	1.921***	(0.379)	1.206***	(0.379)	0.309	0.292	79.4	89.0	0.37
NIR	Quadratic	0.277***	(0.042)	-2.051***	(0.379)	0.954**	(0.379)	0.308	0.290	79.6	89.2	0.37
NRSWIR1	Linear	1.909***	(0.293)	3.977***	(0.707)			0.281	0.272	80.7	87.9	0.38
NIR	Linear	2.202***	(0.370)	-8.908***	(1.701)			0.253	0.244	83.9	91.1	0.39
NRSWIR1	Log-Linear	10.873***	(0.802)	9.884***	(1.934)			0.244	0.235	1379.4	1386.7	1.04
NDSI2	Linear	1.091***	(0.165)	2.083***	(0.408)			0.243	0.234	85.0	92.2	0.39
SAVI	Quadratic	0.277***	(0.044)	-1.853***	(0.399)	0.691*	(0.399)	0.235	0.216	87.8	97.5	0.39
NDWI	Log-Linear	9.905***	(0.649)	7.548***	(1.560)			0.224	0.215	1381.5	1388.8	1.05
NBNIR	Linear	1.886***	(0.338)	2.698***	(0.561)			0.222	0.212	87.3	94.5	0.39
SAVI	Linear	2.155***	(0.411)	-3.194***	(0.696)			0.207	0.197	88.9	96.2	0.40
NBSWIR1	Quadratic	0.277***	(0.045)	1.764***	(0.406)	0.554	(0.406)	0.206	0.186	91.0	100.7	0.40

NGSWIR2	Quadratic	0.277***	(0.045)	0.550	(0.410)	-1.705***	(0.410)	0.193	0.173	92.3	102.0	0.40
NBSWIR1	Linear	2.790***	(0.584)	4.106***	(0.951)			0.187	0.177	90.9	98.1	0.40
NRSWIR2	Quadratic	0.277***	(0.045)	1.359***	(0.412)	-1.110***	(0.412)	0.185	0.165	93.1	102.8	0.40
NIR	Log-Linear	11.090***	(1.038)	-19.769***	(4.769)			0.175	0.165	1386.6	1393.9	1.08
NBSWIR1	Log-Linear	13.292***	(1.569)	10.578***	(2.556)			0.175	0.164	1386.7	1393.9	1.09

S10: A note on the thresholds for soil salinity for rice productivity

Rice salinity tolerance studies show productivity losses at relatively low salinity levels. Laboratory studies have demonstrated that rice productivity drops at 1900 μS/cm under varying conditions of acidity and soil nutrients(Grattan et al. 2002), with measurable stress responses evident even at early seedling stages when exposed to 1800, 3200, and 4600 μS/cm(Zeng, Shannon, and Lesch 2001).

While South Asian soil salinity studies commonly use categorical classifications—slightly saline (2000-4000 μS/cm), moderately saline (4000-8000 μS/cm), strongly saline (8000-16,000 μS/cm) and very strongly saline (>16,000 μS/cm) (Dagar 2005; Sarkar et al. 2023; Krishnamurthy et al. 2022)—these are not rice-specific thresholds. Evidence indicates that Indian rice varieties show measurable stress responses well before the commonly cited 3 dS/m threshold(Kalaiyarasi et al. 2019). Although some South Asian rice cultivars demonstrate resilience to higher salinity levels(Kalaiyarasi et al. 2019; Joseph 2013), their adoption remains limited across the region(Singh et al. 2016), and this tolerance is both stage-specific and genotype-dependent(Shrivastava and Kumar 2015).

The quantitative basis for our 1900 μS/cm threshold is supported by research showing approximately 12% yield loss in rice with every dS/m rise in soil EC(Gao, Chao, and Lin 2007). This translates to ~23% yield loss at 1.9 dS/m versus ~24% at 2.0 dS/m—a difference of only ~1% additional yield loss. However, this 100 μS/cm difference could be crucial for enabling preventive management interventions before significant economic losses occur.

Traditional rice varieties remain widely used but are less productive in saline-affected areas(Mandal 2023; Mahata et al. 2010). For smallholder farmers with limited access to saline-tolerant varieties or inputs, early detection tools like CoSal can provide critical warnings for better adaptation and planning. Our conservative threshold approach prioritizes early detection to enable timely preventive action, consistent with precision agriculture principles for saline soil management.

We test result sensitivity against the 2000 μS/cm threshold commonly used in South Asian soil salinity studies. The CoSal framework can be adapted to predict categorical salinity values when supported by appropriate field data and research objectives

S11: Sensitivity Analysis and 100 Iteration average results of individual machine learning and regression models under varying model specifications

Table S3: Test results of a variety of linear and non-linear machine learning and regression models with EC as the dependent variable and all indices.

EC	**Continuous**	
Variables	All	
Train-Test	70-30	
Row Labels	**Average of R2_Train**	**Average of R2_Test**
Bagging	0.883958695	0.390130017
RandomForest	0.886629656	0.38153062
Quadratic_correlated	0.493063218	0.075613038
Stepwise	0.667250327	-0.4812077
OLS	0.684766378	-0.881249685
ANN	0.899084686	-1.589536008
Quadratic_all	0.883570262	-66.78533693
Log-Linear	0.480977049	-9.34509E+19

EC	**Continuous**	
Variables	All	
Train-Test	70-30	
Random Forest maxnodes = 6 (reduced depth to avoid overfitting in small dataset)		
Row Labels	**Average of R2_Train**	**Average of R2_Test**
Bagging	0.740572046	0.40654547
RandomForest	0.761469885	0.402998112
Quadratic_correlated	0.485572096	0.075818915
Stepwise	0.65584359	-0.519069914
OLS	0.674089184	-0.821922951
ANN	0.889559236	-4.73293199
Quadratic_all	0.875804963	-69.9171603
Log-Linear	0.465514807	-2.5877E+14

EC	**Continuous**	
Variables	Pure bands and highly correlated	

Train-Test	80-20	
Parameter adjustments for small data (maxnodes = 6)		
Row Labels	**Average of R2_Train**	**Average of R2_Test**
RandomForest	0.729574484	0.444587276
Bagging	0.745901662	0.432524444
Lasso	0.385876678	0.301727259
Ridge	0.336485852	0.273195735
OLS	0.519148023	0.270856197
Quadratic_correlated	0.546587506	0.247986491
Stepwise	0.504797542	0.235174696
Quadratic_all	0.606914335	-0.234295277
Log-Linear	-0.113285846	-6.731768748
ANN	-6.586192881	-179.2320997

EC	**Continuous**	
Variables	PCA (3)	
Train-Test	80-20 (random)	
Row Labels	**Average of R2_Train**	**Average of R2_Test**
Bagging	0.454671094	0.191897586
Ridge	0.271008882	0.167371934
Quantile Regression	0.294190471	0.156754325
OLS	0.339315496	0.151301889
Lasso	0.336095118	0.145802083
Log-Linear	0.226656671	0.145261048
RandomForest	0.362814088	0.140630829
Stepwise	0.332848696	0.127901676
ANN	-5.770963851	-30.57203946

A. Testing variable selection with random sampling (Binary 1900 or 2000 μs/cm threshold)

EC	**Binary (1900 μs/cm)**	
Variables	All	
Train-Test	80-20 (random)	

Row Labels	Average of Accuracy Train	Average of Accuracy Test
Random Forest	0.992857143	0.847272727
ANN	0.954166667	0.787272727
Logistic	0.956785714	0.762727273

EC	Binary (2000 µs/cm)	
Variables	All	
Train-Test	80-20 (random)	
Row Labels	**Average of Accuracy Train**	**Average of Accuracy Test**
ANN	0.969518	0.850476
Logistic	0.989398	0.782381
Random Forest	1	0.919048

EC	Binary (1900 µs/cm)	
Variables	Lasso	
Train-Test	80-20 (random)	
Row Labels	**Average of Accuracy Train**	**Average of Accuracy Test**
Logistic	0.889166667	0.848181818
Random Forest	0.992857143	0.841363636
ANN	0.960595238	0.789090909

EC	Binary (1900 µs/cm)	
Variables	Pure bands and highly correlated	
Train-Test	80-20 (random)	
Row Labels	**Average of Accuracy Train**	**Average of Accuracy Test**
Random Forest	0.992857143	0.838181818
Logistic	0.901071429	0.833181818
ANN	0.96	0.775909091

EC	Binary (2000 µs/cm)	
Variables	Pure bands and highly correlated	
Train-Test	80-20 (random)	

Row Labels	Average of Accuracy Train	Average of Accuracy Test
Random Forest	0.97253012	0.848095238
Logistic	0.932289157	0.896666667
ANN	1	0.907619048

EC	**Binary (1900 µs/cm)**	
Variables	Lasso Variables	
Train-Test	80-20 (random)	
Row Labels	**Average of Accuracy Train**	**Average of Accuracy Test**
ANN	0.965421687	0.810952381
Logistic	0.913614458	0.864761905
Random Forest	0.992650602	0.88047619

EC	**Binary (2000 µs/cm)**	
Variables	Lasso Variables	
Train-Test	80-20 (random)	
Row Labels	**Average of Accuracy Train**	**Average of Accuracy Test**
ANN	0.974698795	0.843333333
Logistic	0.936024096	0.915238095
Random Forest	1	0.917619048

EC	**Categorical**	
Variables	All	
Train-Test	80-20 (random)	
Row Labels	**Average of Accuracy Train**	**Average of Accuracy Test**
Random Forest	0.995364489	0.76482743
ANN	0.899229683	0.671691396

EC	**Categorical**	
Variables	Pure bands and highly correlated	
Train-Test	80-20 (random)	
Row Labels	**Average of Accuracy Train**	**Average of Accuracy Test**

Random Forest	0.995541963	0.74184358
ANN	0.886974731	0.667993751

EC	**Categorical**	
Variables	Lasso variables	
Train-Test	80-20 (random)	
Row Labels	**Average of Accuracy Train**	**Average of Accuracy Test**
Random Forest	0.947499565	0.730831604
ANN	0.780972881	0.709749143

Table S4: Testing Variable selection with stratified sampling

EC	**Binary (1900 µs/cm)**	
Variables	All	
Train-Test	80-20 (stratified)	
Row Labels	**Average of Accuracy Train**	**Average of Accuracy Test**
Random Forest	0.992352941	0.844285714
Logistic	0.947294118	0.754761905
ANN	0.954705882	0.798571429

EC	**Binary (2000 µs/cm)**	
Variables	All	
Train-Test	80-20 (stratified)	
Row Labels	**Average of Accuracy Train**	**Average of Accuracy Test**
Random Forest	1	0.9235
Logistic	0.991309524	0.784
ANN	0.97047619	0.871

EC	**Binary (1900 µs/cm)**	
Variables	Lasso	
Train-Test	80-20 (stratified)	
Row Labels	**Average of Accuracy Train**	**Average of Accuracy Test**
ANN	0.956785714	0.8055

Logistic	0.914285714	0.8625
Random Forest	0.992142857	0.871

EC	**Binary (2000 µs/cm)**	
Variables	Lasso	
Train-Test	80-20 (stratified)	
Row Labels	**Average of Accuracy Train**	**Average of Accuracy Test**
ANN	0.968571429	0.8515
Logistic	0.933690476	0.901
Random Forest	1	0.9185

EC	**Binary (1900 µs/cm)**	
Variables	Pure bands and highly correlated	
Train-Test	80-20 (stratified)	
Row Labels	**Average of Accuracy Train**	**Average of Accuracy Test**
Logistic	0.900941176	0.830952381
Random Forest	0.992352941	0.826190476
ANN	0.960117647	0.798095238

EC	**Binary (2000 µs/cm)**	
Variables	Pure bands and highly correlated	
Train-Test	80-20 (stratified)	
Row Labels	**Average of Accuracy Train**	**Average of Accuracy Test**
ANN	0.972142857	0.856
Logistic	0.931547619	0.8905
Random Forest	1	0.914

EC	**Categorical**	
Variables	All	
Train-Test	80-20 (stratified)	
Row Labels	**Average of Accuracy Train**	**Average of Accuracy Test**
Random Forest	0.99173913	0.78800641

ANN	0.900927668	0.686840798

EC	**Categorical**	
Variables	Pure bands and highly correlated	
Train-Test	80-20 (stratified)	
Row Labels	**Average of Accuracy Train**	**Average of Accuracy Test**
Random Forest	0.992463768	0.76102381
ANN	0.881891211	0.710176005

EC	**Categorical**	
Variables	Lasso variables	
Train-Test	80-20 (stratified)	
Row Labels	**Average of Accuracy Train**	**Average of Accuracy Test**
Random Forest	0.931594203	0.752554487
ANN	0.781774522	0.707827381

S12: LASSO Regression Results

LASSO regressions are regressions that penalize adding more variables into the model and decrease coefficient values of variables that contribute less to the model. Two LASSO penalized regressions were run in R with the package ‘glmnet’ to reduce dimensionality of predictors. The first regression was used to determine the best lambda value for penalizing adding more predictors. The second regression was used to select predictors to retain in the model.

Step 1: Initial LASSO regression to determine lambda value

Code:
```
cvmodel <- cv.glmnet(x, y, alpha=1, family='binomial')
```

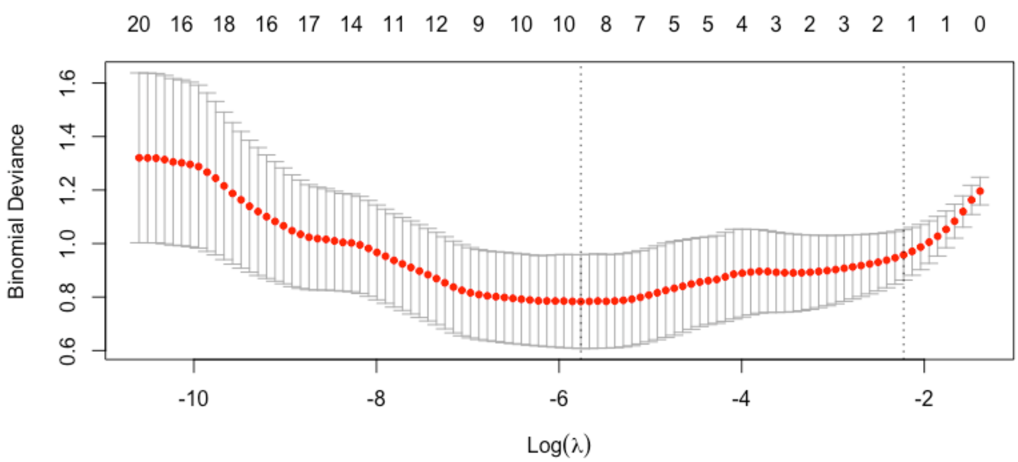


Figure S6: Cross-validation error plot (1900 threshold EC)

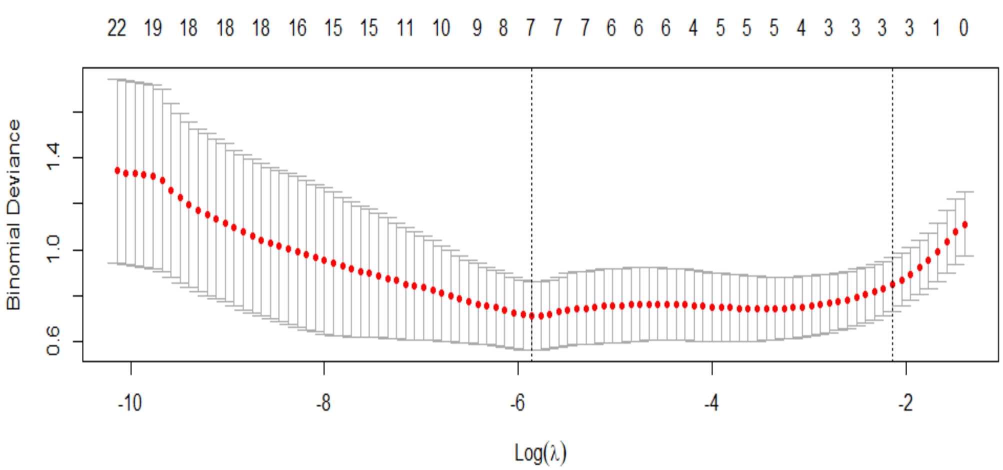


Figure S7: Cross-validation error plot (2000 threshold EC)

(1) Best lambda value with the lowest error (for 1900 EC): 0.00256729

Best lambda value with the lowest error (for 2000 EC): 0.002594856

Step 2: Final LASSO regression to select variables

Code:

```
bestlasso <- glmnet(x, y, alpha=1, lambda=best_lambda, family='binomial')
```

Table S5: LASSO regression Coefficients (for 1900 EC)

Table 1: LASSO regression coefficients (EC threshold= 1900)	
Predictor	**Coefficient values**
(Intercept)	12.746899
Blue	
Green	
Red	14.440849
NIR	
SWIR1	
SWIR2	-34.093366
NDVI	
NDWI	
NDSI1	
NDSI2	
SI1	
SI2	
SI3	
SI4	-9.024859
SI5	
SAVI	
VSSI	-7.410690
NBR	
NBG	
NBNIR	16.084767
NBSWIR1	

NBSWIR2	-11.205443
NRSWIR1	
NRSWIR2	-5.315038
NGSWIR1	-1.603849
NGSWIR2	
NNIRSWIR1	
NNIRSWIR2	

Table S6: LASSO regression Coefficients (for 2000 EC)

Table 1: LASSO regression coefficients (EC threshold= 2000)	
Predictor	**Coefficient values**
(Intercept)	12.552773
Blue	
Green	
Red	2.385307
NIR	
SWIR1	-2.601638
SWIR2	-21.685650
NDVI	
NDWI	
NDSI1	
NDSI2	
SI1	
SI2	
SI3	
SI4	-10.001931
SI5	
SAVI	
VSSI	-11.635416
NBR	
NBG	

NBNIR	20.251038
NBSWIR1	
NBSWIR2	-12.939521
NRSWIR1	
NRSWIR2	
NGSWIR1	
NGSWIR2	
NNIRSWIR1	
NNIRSWIR2	

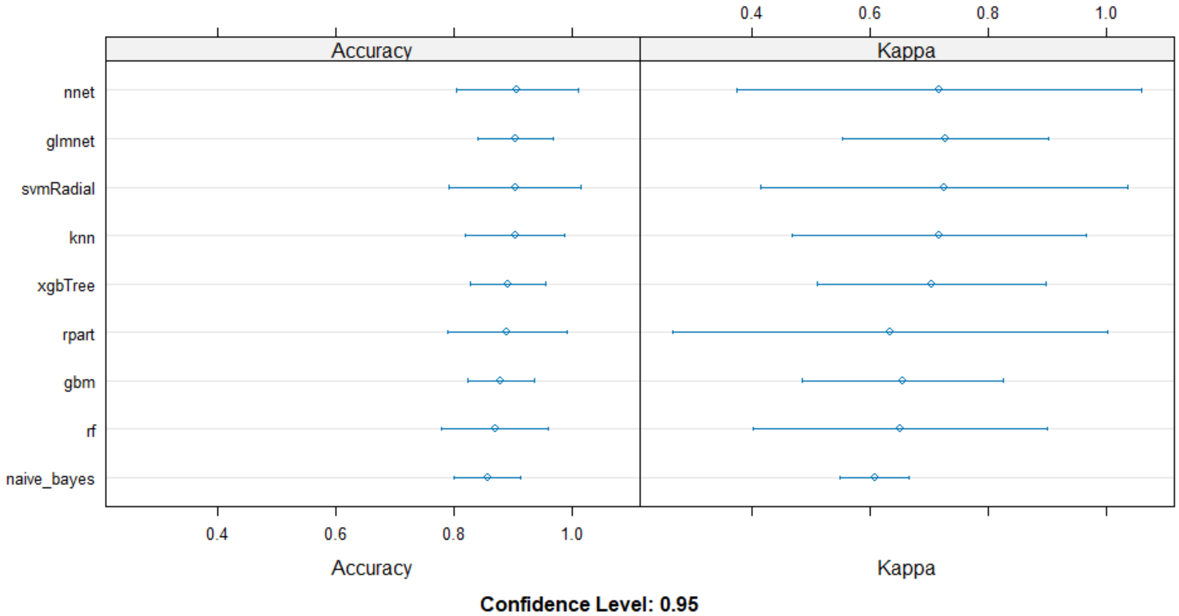
(3) Retained predictors (with non-zero coefficient values for EC threshold =1900): Red, SWIR2, SI4, VSSI, NBNIR, NBSWIR2, NRSWIR2, NGSWIR1.

Retained predictors (with non-zero coefficient values for EC threshold =2000): Red, SWIR1, SWIR2, SI4, VSSI, NBNIR, NBSWIR2.

S13: Machine learning model tests (1900 EC threshold)

Individual model mean accuracy and kappa

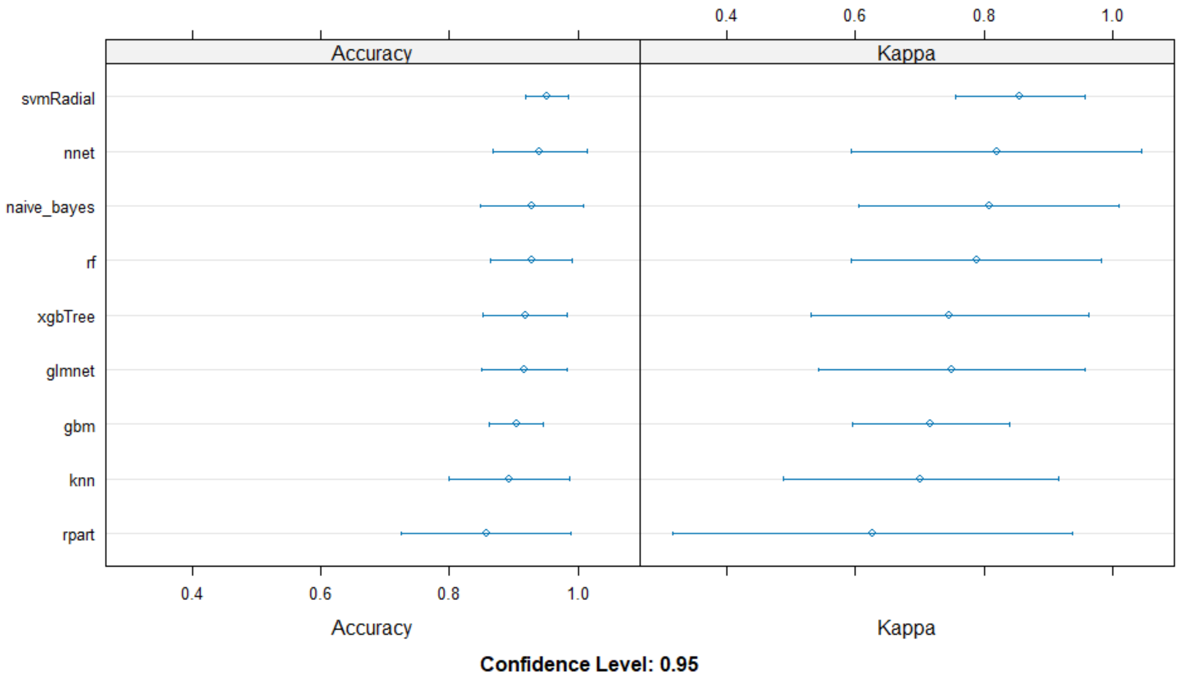
```
Models: rf, rpart, nnet, svmRadial, gbm, naive_bayes, xgbTree, knn, glm
net
Number of resamples: 5

Accuracy
                 Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
NA's
rf          0.7777778 0.8125000 0.8823529 0.8702614 0.9375000 0.9411765
0
rpart       0.8125000 0.8125000 0.8888889 0.8910131 0.9411765 1.0000000
0
nnet        0.7777778 0.8888889 0.9375000 0.9083333 0.9375000 1.0000000
0
svmRadial   0.7647059 0.8750000 0.9411765 0.9044118 0.9411765 1.0000000
0
gbm         0.8125000 0.8823529 0.8823529 0.8801471 0.8823529 0.9411765
0
naive_bayes 0.7777778 0.8750000 0.8750000 0.8583333 0.8750000 0.8888889
0
xgbTree     0.8125000 0.8823529 0.8888889 0.8924837 0.9375000 0.9411765
0
knn         0.8235294 0.8750000 0.8823529 0.9044118 0.9411765 1.0000000
0
glmnet      0.8235294 0.8823529 0.9375000 0.9050654 0.9375000 0.9444444
0

Kappa
                 Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
NA's
rf          0.3684211 0.5384615 0.6792453 0.6506669 0.8210526 0.8461538
0
rpart       0.3333333 0.3333333 0.6842105 0.6343860 0.8210526 1.0000000
0
nnet        0.2653061 0.6842105 0.8181818 0.7171761 0.8181818 1.0000000
0
svmRadial   0.3584906 0.6000000 0.8210526 0.7258201 0.8495575 1.0000000
0
gbm         0.4545455 0.6046512 0.6792453 0.6552322 0.7166667 0.8210526
0
naive_bayes 0.5555556 0.6000000 0.6000000 0.6079532 0.6000000 0.6842105
0
xgbTree     0.4545455 0.6842105 0.7166667 0.7046324 0.8181818 0.8495575
0
knn         0.4848485 0.6000000 0.6792453 0.7170293 0.8210526 1.0000000
0
glmnet      0.5486726 0.6046512 0.8181818 0.7284293 0.8181818 0.8524590
0
Mean accuracy: 0.8905
Models above threshold: rpart, nnet, svmRadial, xgbTree, knn, glmnet
```




```
The following models were ensembled: rf, rpart, nnet, svmRadial, gbm, n
aive_bayes, xgbTree, knn, glmnet

Model Importance:
rf     rpart  nnet    svmRadial  gbm    naive_bayes  xgbTree  knn    gl
mnet
0.1606 0.0495 0.1374  0.1067     0.1095 0.1052       0.1398   0.0810 0.
1102


Model accuracy:
     model_name   metric     value         sd
 1:    ensemble      ROC 0.8393590 0.10699331
 2:          rf Accuracy 0.8702614 0.07346050
 3:       rpart Accuracy 0.8910131 0.08174294
 4:        nnet Accuracy 0.9083333 0.08295632
 5:   svmRadial Accuracy 0.9044118 0.08975409
 6:         gbm Accuracy 0.8801471 0.04559416
 7: naive_bayes Accuracy 0.8583333 0.04543174
 8:     xgbTree Accuracy 0.8924837 0.05222643
 9:         knn Accuracy 0.9044118 0.06779077
10:      glmnet Accuracy 0.9050654 0.05200646

Best threshold: 0.661
Mean difference: 0.1355085
```

Machine learning model tests (2000 EC threshold)

```
Models: rf, rpart, nnet, svmRadial, gbm, naive_bayes, xgbTree, knn, glm
net
Number of resamples: 5

Accuracy
                 Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
NA's
rf          0.8750000 0.8823529 0.9411765 0.9279412 0.9411765 1.0000000
0
rpart       0.7058824 0.8235294 0.8750000 0.8573529 0.8823529 1.0000000
0
nnet        0.8823529 0.8823529 0.9375000 0.9404412 1.0000000 1.0000000
0
svmRadial   0.9375000 0.9411765 0.9411765 0.9522059 0.9411765 1.0000000
0
gbm         0.8750000 0.8823529 0.8823529 0.9044118 0.9411765 0.9411765
0
naive_bayes 0.8235294 0.9375000 0.9411765 0.9286765 0.9411765 1.0000000
0
xgbTree     0.8823529 0.8823529 0.8823529 0.9176471 0.9411765 1.0000000
0
knn         0.7647059 0.8823529 0.9375000 0.8933824 0.9411765 0.9411765
0
glmnet      0.8750000 0.8823529 0.8823529 0.9161765 0.9411765 1.0000000
0

Kappa
                 Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
NA's
rf          0.6046512 0.6666667 0.8210526 0.7883856 0.8495575 1.0000000
0
rpart       0.3511450 0.4631579 0.6046512 0.6266480 0.7142857 1.0000000
0
nnet        0.6046512 0.6730769 0.8181818 0.8191820 1.0000000 1.0000000
0
svmRadial   0.8181818 0.8210526 0.8210526 0.8562679 0.8210526 1.0000000
0
gbm         0.6046512 0.6666667 0.6730769 0.7173000 0.8210526 0.8210526
0
naive_bayes 0.5486726 0.8181818 0.8210526 0.8074929 0.8495575 1.0000000
0
xgbTree     0.6046512 0.6046512 0.6730769 0.7463874 0.8495575 1.0000000
0
knn         0.4426230 0.6046512 0.8181818 0.7015122 0.8210526 0.8210526
0
glmnet      0.6000000 0.6046512 0.7213115 0.7494031 0.8210526 1.0000000
0
Mean accuracy: 0.9154
Models above threshold: rf, nnet, svmRadial, naive_bayes, xgbTree,
glmnet
```




```
The following models were ensembled: rf, rpart, nnet, svmRadial, gbm, n
aive_bayes, xgbTree, knn, glmnet

Model Importance:
rf       rpart    nnet    svmRadial gbm     naive_bayes xgbTree knn    glmn
et
0.1012   0.0612   0.1405 0.3107     0.0978 0.1397      0.0622  0.0081 0.07
87


Model accuracy:
     model_name   metric     value         sd
 1:    ensemble      ROC 0.8807692 0.12785208
 2:          rf Accuracy 0.9279412 0.05104869
 3:       rpart Accuracy 0.8573529 0.10650349
 4:        nnet Accuracy 0.9404412 0.05884650
 5:   svmRadial Accuracy 0.9522059 0.02676511
 6:         gbm Accuracy 0.9044118 0.03369541
 7: naive_bayes Accuracy 0.9286765 0.06429095
 8:     xgbTree Accuracy 0.9176471 0.05261336
 9:         knn Accuracy 0.8933824 0.07614822
10:      glmnet Accuracy 0.9161765 0.05393271

Best threshold: 0.624
Mean difference = 0.05155791
```

S14: Threshold Test Results

The sensitivity of the final ensemble model to threshold values for converting predicted probabilities to binary values was tested. 81 possible threshold values were tested ranging from 0.1 to 0.9 at 0.01 increments. In each iteration, the threshold was used to convert predicted probabilities to binary values and compute accuracy for the training set and the internal testing set.

For 1900 EC Threshold: Mean accuracy for the training set was 0.91 (±0.025) and mean accuracy for the testing set was 0.82 (±0.056). Differences between the training and testing sets were low for all threshold values, with a mean value of 0.090 (±0.045). These results suggest low possibility of overfitting and relatively low sensitivity of model performance to changing threshold values. The final threshold used is 0.661, owing to high overall training and testing accuracies and low differences between these accuracies.

Table S7: Accuracy results for different threshold values. (EC= 1900 EC)

Threshold	Train accuracy	Test accuracy	Difference in accuracy
0.1	**0.892857**	0.678571	0.45
0.11	**0.916667**	0.702381	0.5
0.12	**0.916667**	0.75	0.5
0.13	**0.928571**	0.761905	0.5
0.14	**0.928571**	0.785714	0.5
0.15	**0.940476**	0.821429	0.5
0.16	**0.964286**	0.833333	0.5
0.17	**0.964286**	0.892857	0.55
0.18	**0.964286**	0.928571	0.5
0.19	**0.964286**	0.952381	0.55
0.2	**0.952381**	0.952381	0.55
0.21	**0.952381**	0.964286	0.55
0.22	**0.952381**	0.964286	0.6
0.23	**0.952381**	0.964286	0.6
0.24	**0.940476**	0.964286	0.6
0.25	**0.940476**	0.964286	0.6
0.26	**0.928571**	0.964286	0.6
0.27	**0.904762**	0.964286	0.7
0.28	**0.904762**	0.964286	0.75
0.29	**0.904762**	0.964286	0.75
0.3	**0.916667**	0.964286	0.75
0.31	**0.928571**	0.940476	0.75
0.32	**0.928571**	0.940476	0.75
0.33	**0.928571**	0.928571	0.75
0.34	**0.928571**	0.952381	0.75
0.35	**0.928571**	0.940476	0.75
0.36	**0.928571**	0.940476	0.75
0.37	**0.928571**	0.940476	0.8
0.38	**0.928571**	0.928571	0.85
0.39	**0.928571**	0.928571	0.85
0.4	**0.928571**	0.928571	0.85
0.41	**0.928571**	0.928571	0.85
0.42	**0.928571**	0.928571	0.85
0.43	**0.916667**	0.928571	0.85
0.44	**0.916667**	0.928571	0.85

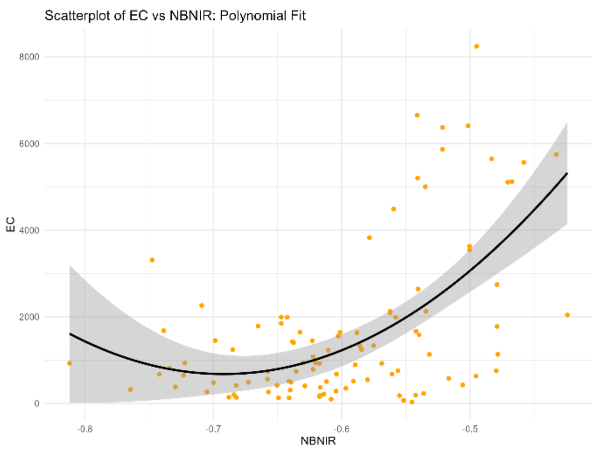
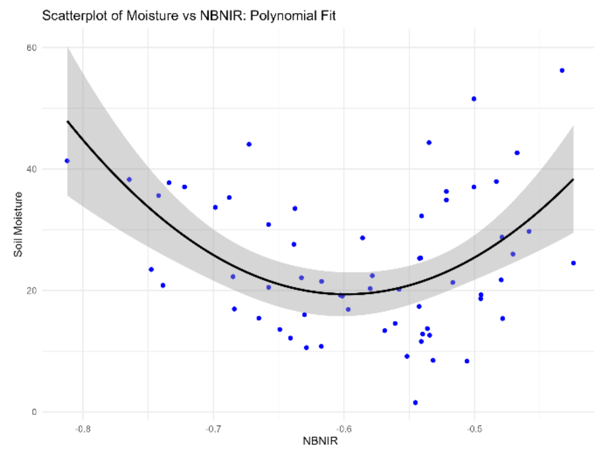
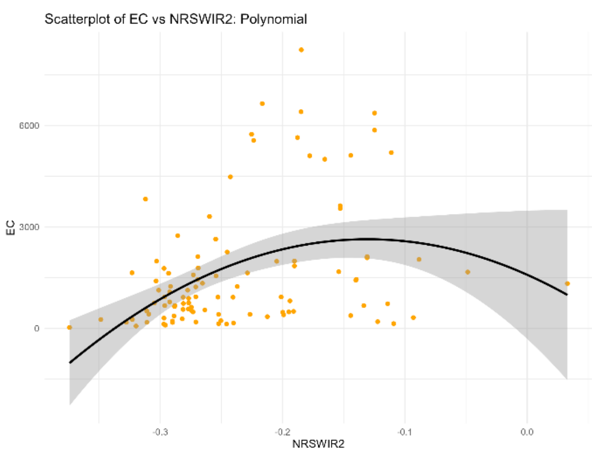
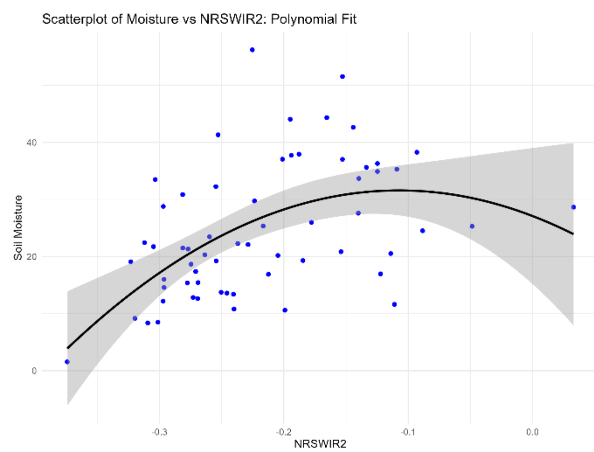
0.45	**0.916667**	0.928571	0.85
0.46	**0.916667**	0.928571	0.85
0.47	**0.916667**	0.916667	0.85
0.48	**0.916667**	0.916667	0.85
0.49	**0.916667**	0.916667	0.85
0.5	**0.916667**	0.916667	0.85
0.51	**0.916667**	0.916667	0.85
0.52	**0.928571**	0.916667	0.85
0.53	**0.928571**	0.916667	0.85
0.54	**0.928571**	0.916667	0.85
0.55	**0.928571**	0.916667	0.85
0.56	**0.928571**	0.916667	0.85
0.57	**0.928571**	0.916667	0.85
0.58	**0.928571**	0.916667	0.85
0.59	**0.928571**	0.916667	0.85
0.6	**0.928571**	0.916667	0.85
0.61	**0.916667**	0.916667	0.9
0.62	**0.916667**	0.916667	0.9
0.63	**0.916667**	0.916667	0.9
0.64	**0.916667**	0.916667	0.9
0.65	**0.916667**	0.916667	0.9
0.66	**0.916667**	0.916667	0.9
0.67	**0.916667**	0.916667	0.9
0.68	**0.916667**	0.916667	0.9
0.69	**0.916667**	0.916667	0.9
0.7	**0.916667**	0.916667	0.9
0.71	**0.916667**	0.916667	0.9
0.72	**0.916667**	0.916667	0.85
0.73	**0.916667**	0.916667	0.85
0.74	**0.916667**	0.916667	0.85
0.75	**0.916667**	0.916667	0.85
0.76	**0.916667**	0.916667	0.85
0.77	**0.916667**	0.916667	0.85
0.78	**0.916667**	0.916667	0.85
0.79	**0.916667**	0.916667	0.85
0.8	**0.916667**	0.916667	0.85
0.81	**0.916667**	0.916667	0.85
0.82	**0.916667**	0.916667	0.85
0.83	**0.916667**	0.916667	0.85
0.84	**0.916667**	0.916667	0.85
0.85	**0.916667**	0.916667	0.85
0.86	**0.916667**	0.916667	0.85
0.87	**0.916667**	0.916667	0.85
0.88	**0.916667**	0.916667	0.85
0.89	**0.916667**	0.916667	0.85
0.9	**0.916667**	0.916667	0.85

For 2000 EC Threshold: Mean accuracy for the training set was 0.95 (±0.025) and mean accuracy for the testing set was 0.82 (±0.056). Differences between the training and testing sets were low for all threshold values, with a mean value of 0.090 (±0.045). These results suggest low possibility of overfitting and relatively low sensitivity of model performance to changing threshold values. The final threshold used is 0.624, owing to high overall training and testing accuracies and low differences between these accuracies.

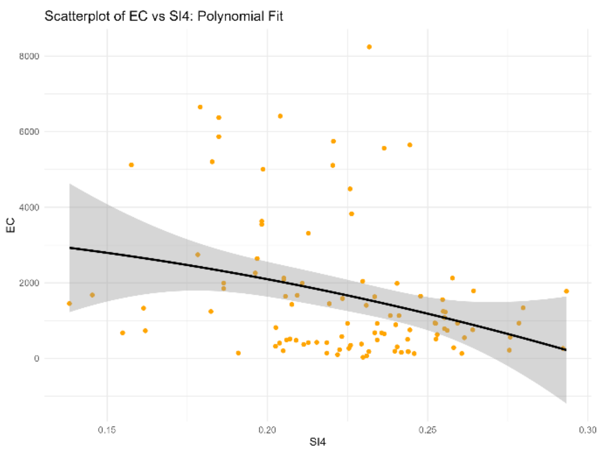
Table S8: Accuracy results for different threshold values. (EC= 2000 EC)			
Threshold	**Train accuracy**	**Test accuracy**	**Difference in accuracy**
0.1	0.880952	0.85	0.030952
0.11	0.892857	0.85	0.042857
0.12	0.892857	0.85	0.042857
0.13	0.892857	0.85	0.042857
0.14	0.904762	0.85	0.054762
0.15	0.892857	0.85	0.042857
0.16	0.880952	0.85	0.030952
0.17	0.880952	0.85	0.030952
0.18	0.892857	0.85	0.042857
0.19	0.892857	0.85	0.042857
0.2	0.892857	0.85	0.042857
0.21	0.892857	0.85	0.042857
0.22	0.904762	0.85	0.054762
0.23	0.904762	0.85	0.054762
0.24	0.916667	0.85	0.066667
0.25	0.916667	0.85	0.066667
0.26	0.940476	0.85	0.090476
0.27	0.940476	0.9	0.040476
0.28	0.940476	0.9	0.040476
0.29	0.940476	0.9	0.040476
0.3	0.940476	0.9	0.040476
0.31	0.952381	0.9	0.052381
0.32	0.952381	0.9	0.052381
0.33	0.940476	0.9	0.040476
0.34	0.940476	0.9	0.040476
0.35	0.952381	0.9	0.052381
0.36	0.952381	0.9	0.052381
0.37	0.952381	0.9	0.052381
0.38	0.952381	0.9	0.052381
0.39	0.952381	0.9	0.052381
0.4	0.952381	0.9	0.052381
0.41	0.952381	0.9	0.052381
0.42	0.964286	0.9	0.064286
0.43	0.964286	0.9	0.064286
0.44	0.964286	0.9	0.064286
0.45	0.964286	0.9	0.064286
0.46	0.964286	0.9	0.064286
0.47	0.964286	0.9	0.064286
0.48	0.952381	0.9	0.052381
0.49	0.952381	0.9	0.052381
0.5	0.952381	0.9	0.052381
0.51	0.952381	0.9	0.052381
0.52	0.952381	0.9	0.052381
0.53	0.952381	0.9	0.052381
0.54	0.952381	0.9	0.052381
0.55	0.952381	0.9	0.052381
0.56	0.952381	0.9	0.052381

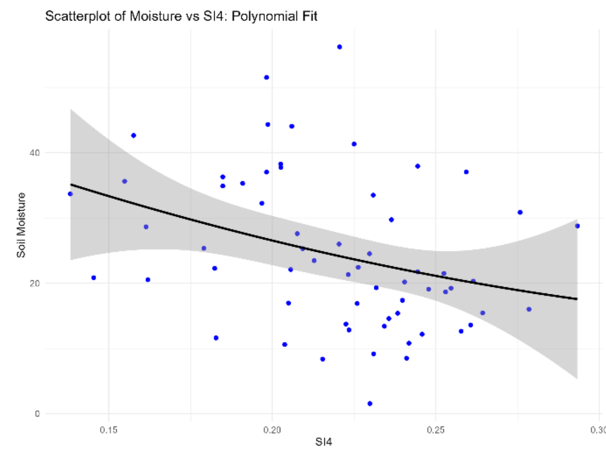
0.57	0.952381	0.9	0.052381
0.58	0.952381	0.9	0.052381
0.59	0.952381	0.9	0.052381
0.6	0.952381	0.9	0.052381
0.61	0.952381	0.9	0.052381
0.62	0.952381	0.9	0.052381
0.63	0.952381	0.9	0.052381
0.64	0.952381	0.9	0.052381
0.65	0.952381	0.9	0.052381
0.66	0.952381	0.9	0.052381
0.67	0.952381	0.9	0.052381
0.68	0.952381	0.9	0.052381
0.69	0.952381	0.9	0.052381
0.7	0.952381	0.9	0.052381
0.71	0.952381	0.9	0.052381
0.72	0.952381	0.9	0.052381
0.73	0.952381	0.9	0.052381
0.74	0.952381	0.9	0.052381
0.75	0.952381	0.9	0.052381
0.76	0.952381	0.9	0.052381
0.77	0.952381	0.9	0.052381
0.78	0.952381	0.9	0.052381
0.79	0.952381	0.9	0.052381
0.8	0.952381	0.9	0.052381
0.81	0.952381	0.9	0.052381
0.82	0.952381	0.9	0.052381
0.83	0.952381	0.9	0.052381
0.84	0.952381	0.9	0.052381
0.85	0.952381	0.9	0.052381
0.86	0.952381	0.9	0.052381
0.87	0.952381	0.9	0.052381
0.88	0.952381	0.9	0.052381
0.89	0.952381	0.9	0.052381
0.9	0.952381	0.9	0.052381

S15: Relationship between predictor variables and Electrical Conductivity and Soil Moisture

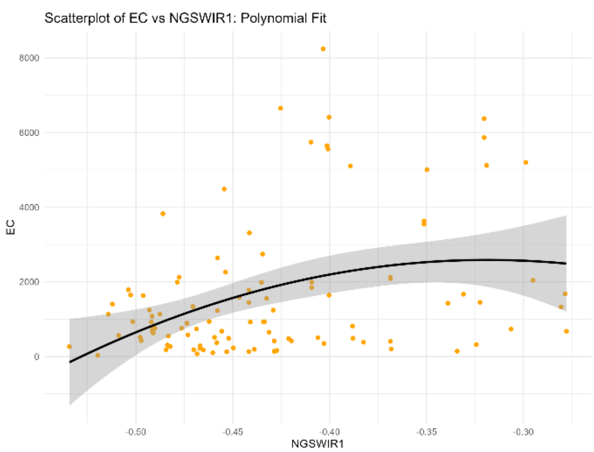
Relationship of predictor with EC	**Relationship of predictor with soil moisture**
Normalized Blue and NIR (NBNIR)  	
Normalized Red and SWIR2 (NRSWIR2)  	

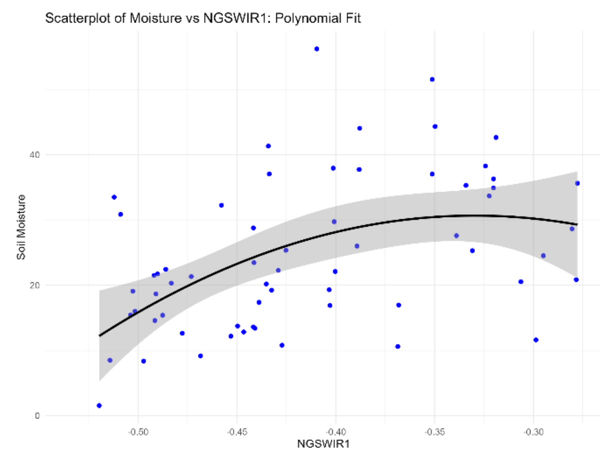
Salinity Index 4 (SI4)



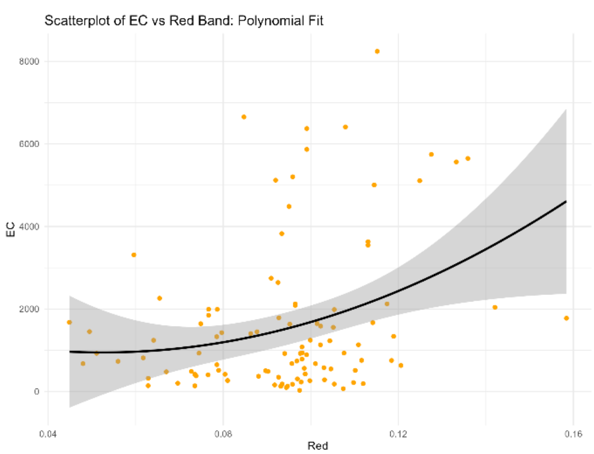


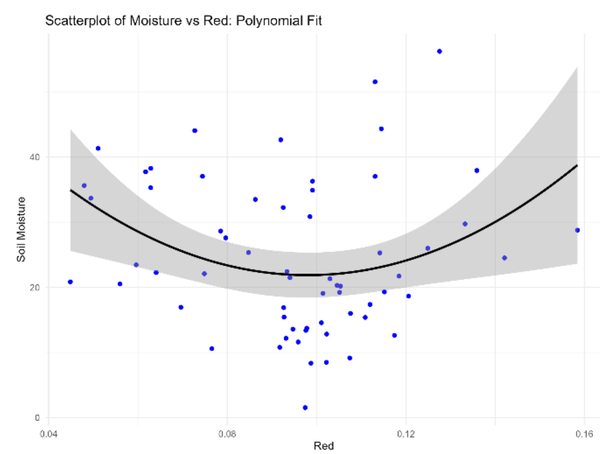
Normalized Green and SWIR1 (NGSWIR1)



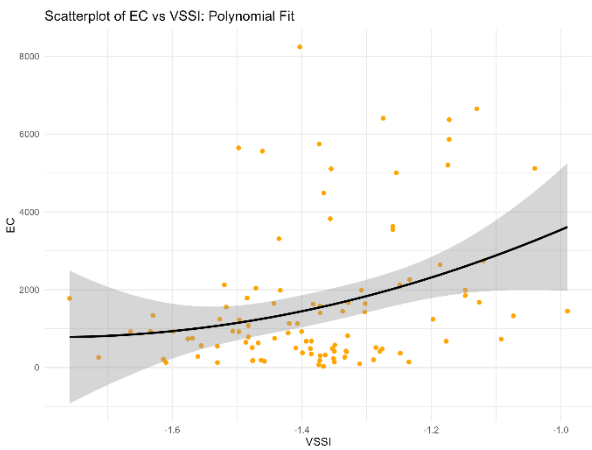


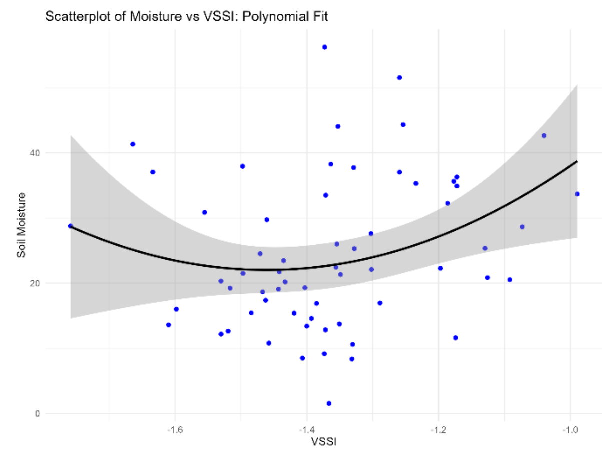
Red



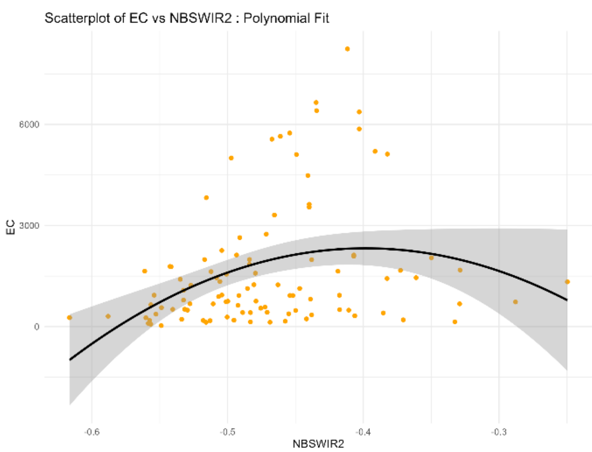


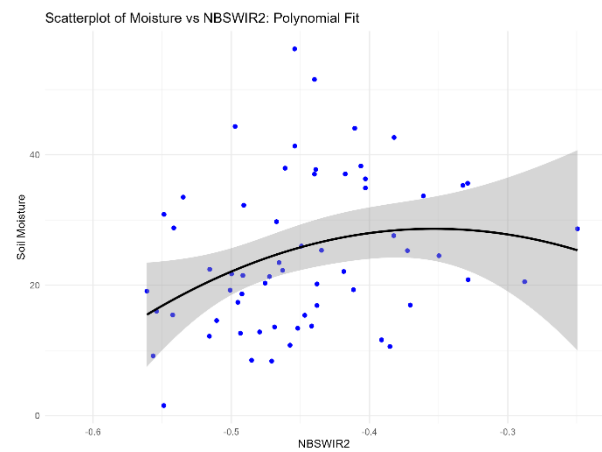
Vegetation Soil Salinity Index (VSSI)



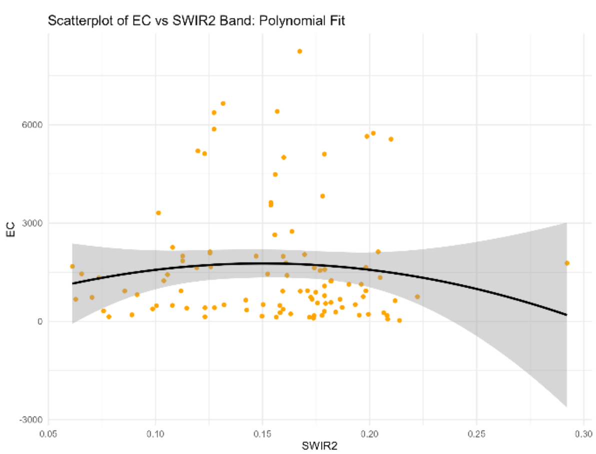


Normalized Blue and SWIR2 (NBSWIR2)





SWIR2



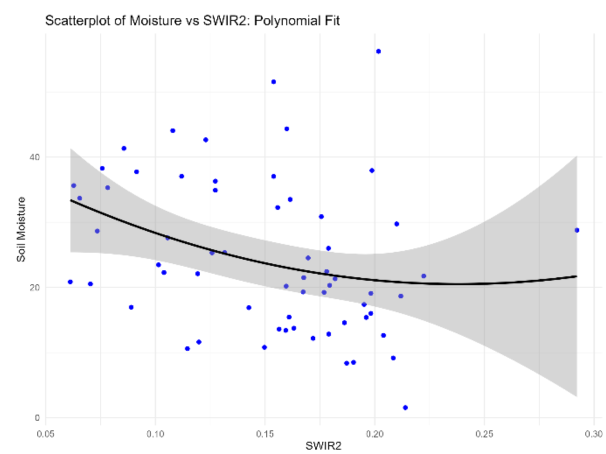


Figure S8: Relationships between indices, featuring high in the CoSal-SA model, with EC and moisture. Different indices detect salinity and moisture in varying ways: Indices, such as VSSI, increase for higher levels of EC as well as moisture; SWIR2 and Red only detect water or salinity respectively; other indices such as NRSWIR2, have a non-linear relationship with salinity but a positive relationship with moisture.

S16: Model Validation metrics

For each model as they use a different subset of data—training, and internal and external testing data sets—we calculated the confusion matrix of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), and derived the four metrics to evaluate the model comprehensively:

1. Accuracy = correct classifications / total classifications = (TP + TN)/(TP+TN+FP+FN)

   Accuracy involves all four components of the confusion matrix, and usually serves as a coarse measure of overall model quality. In case of imbalanced data, such as ours where there are fewer high salinity points (30 out of 106) than there are for lower salinity (70 out of 106), there is a possibility of getting a relatively high accuracy even if all the high salinity points are incorrectly classified as low (accuracy would still be 70% because TN=70 even when FN=30). We, therefore, also consider the following three metrics in tandem.

2. Precision = correctly classified positives / all classified as positive = TP / (TP+FP)

   Precision improves when false positives decrease, that is, high saline data is not classified as low.

3. Recall = correctly classified positives / all actual positives = TP / (TP+FN)

   Recall, or sensitivity or true positive rate, improves when false negatives reduce (that is fewer low saline points are classified as high).

4. F1-score = 2 * (precision * recall) / (precision + recall)

   F1 score is the harmonic mean of precision and recall, and higher score suggests a more balanced model.

S17: Cross-validation of AquaClim-Atlas aquaculture land cover classification with Clark University CGA data (Alemohammad 2024)

The *AquaClim-Atlas* land-use outputs were compared with the Clark University Center for Geospatial Analytics (Clark CGA) aquaculture maps for 1999, 2014, 2018, 2020, and 2022 (**Error! Reference source not found.**). The Clark CGA maps, based on Landsat 5, 8, or 9 imagery and occasionally Sentinel 1/2 data, employ a Mahalanobis Typicality classifier at 15 m resolution [59]. Differences between the two datasets likely arise from methodological distinctions: the CGA approach merges panchromatic and visible bands, which improves pond-boundary delineation but can introduce noise in mixed-pixel environments (e.g., mangrove–pond mosaics).

A confusion matrix reveals the that the overall accuracy ranges between 81% - 89.5%; precision ranges between 76.9% and 85.7%, recall between 83.3% to 88.2% and F1 Score ranges between 80.8% to 86.1% for the 4 corresponding years (Note: Year 1999 has missing Landsat satellite data for most regions on the east coast of India, and is therefore excluded in AquaClim-Atlas).

The minor differences could be owing to (1) missing smaller ponds adjacent to perennial water bodies in Clark CGA, (2) perennial water bodies margin error in AquaClim due to JRC water mask layer (3) temporal aggregation (AquaClim-Atlas only captures active ponds in dry season). Overall, however, these comparisons across representative coastal sections show broadly consistent spatial patterns, validating the reliability of the *AquaClim-Atlas* aquaculture layer.

Raster values for AquaClim and Clark CGA database for raster calculation

AquaClim -> Clark |	Aqua (2)	Inactive Aqua (1)	Non-Aqua (0)
Aqua (10)	8	9	10
Non-Aqua (20)	18	19	20

Since Clark CGA data does not capture the complete extents of the districts, we have taken a random 5km x 5km grid and calculated the accuracies using a confusion matrix. Also, since CGA data does not differentiate between active and inactive aquaculture, for our confusion matrix, we assume inactive aquaculture and aquaculture overlap as aquaculture and inactive aquaculture and non-aquaculture overlap as non-aquaculture. Following are the confusion matrix results for 2014, 2018, 2020, and 2022. Note: Year 1999 is dropped in AquaClim-Atlas due to unavailable or poor-quality Landsat imagery for the entire coast of India, and hence a comparison is not made.

Results - 2014 (Extent = 5 x 5 km grid in Jagatsinghpur District)			
AquaClim -> Clark |	Aqua (2)	Inactive Aqua (1)	Non-Aqua (0)
Aqua (10)	9.80%	13.20%	4.60%
Non-Aqua (20)	5.90%	11.70%	54.60%
TP =	23.0%		
TN =	66.3%		
FP =	5.9%		
FN =	4.6%		

Accuracy =	89.5%	TP+TN/TP+TN+FP+FN	
Precision =	79.6%	TP/(TP+FP)	
Recall =	83.3%	TP/(TP+FN)	
F1-Score =	81.4%	(2 * (precision * recall) / (precision + recall)	

Results - 2018 (Extent = 5 x 5 km grid in Jagatsinghpur District)			
AquaClim -> Clark |	Aqua (2)	Inactive Aqua (1)	Non-Aqua (0)
Aqua (10)	21.10%	19.30%	5.40%
Non-Aqua (20)	7.60%	5.00%	41.60%
TP =	40.4%		
TN =	46.6%		
FP =	7.6%		
FN =	5.4%		
Accuracy =	87.0%	TP+TN/TP+TN+FP+FN	
Precision =	84.2%	TP/(TP+FP)	
Recall =	88.2%	TP/(TP+FN)	
F1-Score =	86.1%	(2 * (precision * recall) / (precision + recall)	

Results - 2020 (Extent = 5 x 5 km grid in Jagatsinghpur District)			
AquaClim -> Clark |	Aqua (2)	Inactive Aqua (1)	Non-Aqua (0)
Aqua (10)	27.20%	11.90%	7.70%
Non-Aqua (20)	6.50%	2.70%	44.00%
TP =	39.1%		
TN =	46.7%		
FP =	6.5%		
FN =	7.7%		
Accuracy =	85.8%	TP+TN/TP+TN+FP+FN	
Precision =	85.7%	TP/(TP+FP)	
Recall =	83.5%	TP/(TP+FN)	
F1-Score =	84.6%	(2 * (precision * recall) / (precision + recall)	

Results - 2022 (Extent = 5 x 5 km grid in Jagatsinghpur District)			
AquaClim -> Clark |	Aqua (2)	Inactive Aqua (1)	Non-Aqua (0)
Aqua (10)	33.10%	6.80%	7.00%
Non-Aqua (20)	12.00%	1.60%	39.60%

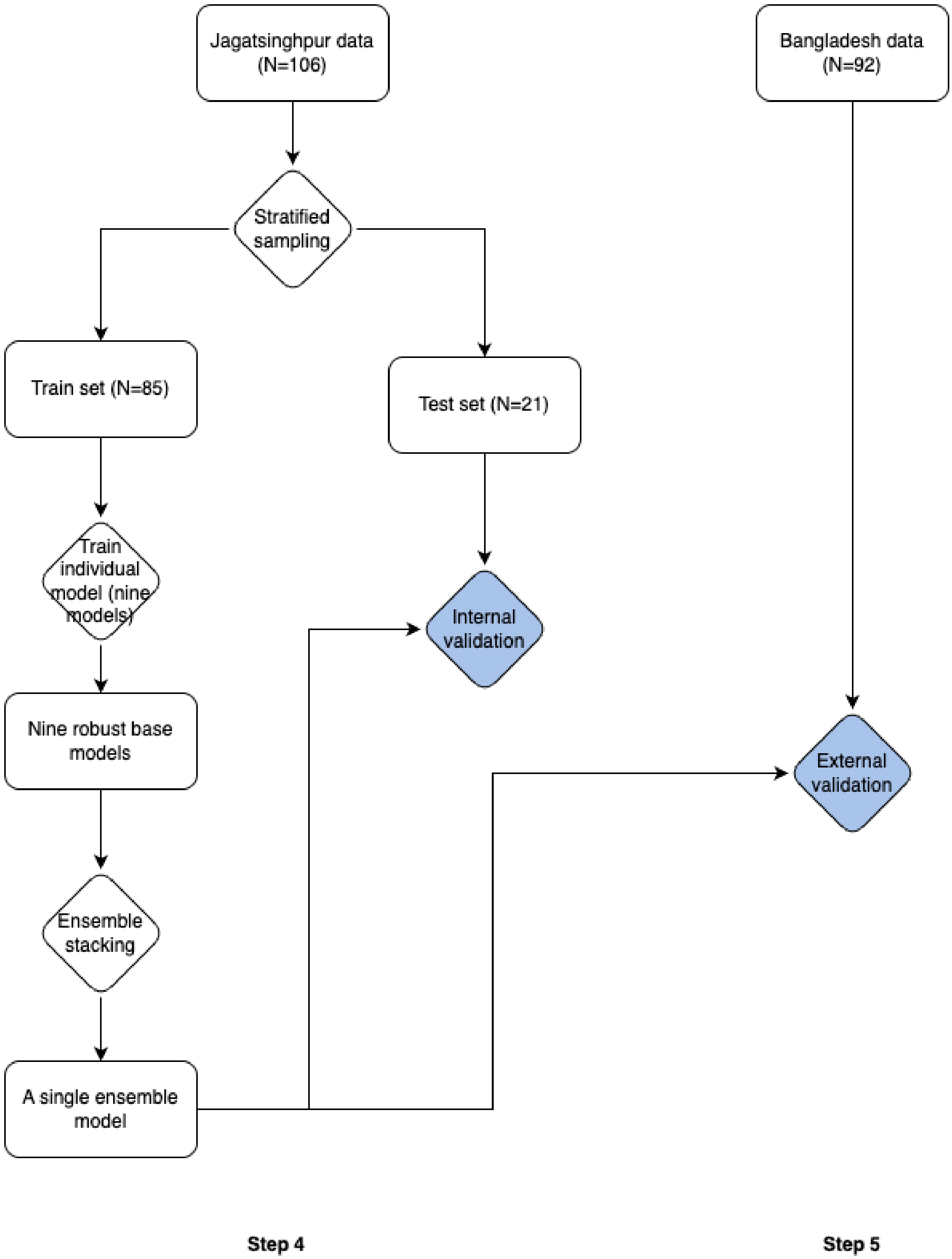
TP =	39.9%		
TN =	41.2%		
FP =	12.0%		
FN =	7.0%		
Accuracy =	81.0%	TP+TN/TP+TN+FP+FN	
Precision =	76.9%	TP/(TP+FP)	
Recall =	85.1%	TP/(TP+FN)	
F1-Score =	80.8%	(2 * (precision * recall) / (precision + recall)	

S18: CoSal Validation process

We included two validation steps in the analysis (see flowchart below). We performed internal validation directly using a held out set (referred to as the test set in the paper) from the Jagatsinghpur dataset, and external validation on the Bangladesh data set. We clarify how these were performed as follows.

Steps to ensure model robustness and stability (train test split and cross validation)
In the model building process (Step 4), we trained the nine models individually, and then applied ensemble stacking to improve model robustness and performance (see the left part of the flowchart below). First, we applied a single stratified random sampling to derive a train set (80%, N=85) and a test set (20%, N=21).
Second, using the train set, we trained individual models for the nine selected models (e.g., random forest, neural networks, etc.) to (1) understand individual model performance, and (2) tune hyperparameters to create relatively more stable and robust base models for ensemble stacking. For each of the nine models, we ran 5-fold cross validations to evaluate combinations of hyperparameters, selected the best combination, and fit one model using the best parameters. Specifically, the cross-validation process randomly partitioned the test set into five folds. Repeated five times for each combination of hyperparameters, each model was trained on data from the four folds and validated on the other. The best combination of hyperparameters for each model was used to train a model for each of the nine so that there were nine trained models at the end of this process.
Third, we applied ensemble stacking on the nine base models from the prior step to derive a single final model. The ensemble process was based on 5-fold cross validation and a random forest (RF) meta-classifier. Repeated five times, the RF meta-classifier was trained on predictions from the nine base models of the four folds and validated on the other one fold. The entire cross-validation process was again performed on all combinations of parameters, so that the best parameters for the RF meta-classifier were chosen to reduce overfitting and improve performance. Finally, we fit a single model using the determined parameters from the CV process on the entire test set with the RF meta-classifier.
Internal validation process
We applied the final ensemble model on the held out test set (N=21) for internal validation.

External validation process
In step 5, we applied the final ensemble model on the Bangladesh dataset (N=92) for external validation.



CoSal Validation Process. Internal and external validation processes for the CoSal Model.

Internal validation: Each model within the CoSal-SA ensemble was evaluated using confusion matrices for true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Overall accuracy was calculated as (TP + TN) / (TP + TN + FP + FN). Since this metric involves all components of the confusion matrix, it usually serves as a coarse measure of overall model quality. In case of imbalanced data, such as these salinity data (fewer points above the salinity risk threshold (36/106 samples above 1,900 μS/cm electrical conductivity) than there are for below the threshold), a relatively high accuracy could result even if all the high salinity points are incorrectly classified as low (e.g. accuracy would still be 70% because TN=70 even when FN=30). Therefore, three additional metrics are also considered to gain better insights into the robustness of the model predictions.

Precision quantifies the share of correctly classified positives [TP / (TP + FP)], recall measures sensitivity [TP / (TP + FN)], and F1-score represents their harmonic mean (2 * (precision * recall) / (precision + recall). The ensemble model achieved a training accuracy of 91.7% and internal testing accuracy of 90%. The highest accuracy among individual base models was 0.84, while the ensemble achieved 0.90 with precision = 0.80, recall = 0.80, and F1-score = 0.80 (see Appendix 28). Ensemble modeling mitigated variance and bias, improving generalization relative to single-model approaches—particularly important given the limited training data typical for soil measurements.

External validation: Model performance is evaluated using both internal validation with field data from India and external validation using independent soil observations from Bangladesh, enabling assessment of transferability across space and time. Soil data collected in 2016 from Satkhira district in Bangladesh (Morshed et al., 2021; Sarkar, Rudra, Nur, et al., 2023) is used to test its external validity across time and space for South Asian coastal geographies.

With the Bangladesh sample, we found a testing accuracy of 64%, precision of 75%, recall of 60% and F1 Score of 67.7.3%. The accuracy drops on external data is common and expected since there are variations in data collection methods and other differences in the geography. The model maintains high precision, suggesting a reasonable reliability in positive predictions of salinity and the maintained F1-score suggests our model is robust.
While external validation using data from Bangladesh demonstrates the transferability of the framework, model performance declines across contexts due to differences in environmental conditions, sampling methods, and reflectance characteristics.

S19: Surge events

Environmental shock impact of storm surges using SURGEDAT data base that documents all surge events from 1880-2015 (Keim 2015)

Table S9: SURGEDAT database for Coastal India (Keim 2015)

Year	Storm_Name	Storm_Date	Country	Location	Surge_tide
1737	Unnamed	October	IN	Sunderbans (storm crossed coast at mouth of Hooghly River)	12
1864	Unnamed	Oct	IN	Calcutta and Surroundings	12
1876	Unnamed	Oct 27- Nov 1	BD	Bangladesh	13.7
1885	False Point Cyclone	Sept	IN	North Orissa	7
1952	Unnamed	November	IN	South Coromandal Coast and Northern Shores of Palk Bay	3
1964	Rameswaram	Dec	IN	India	5
1969	Unnamed	10-Oct	BD	Bangladesh	7.3
1971	Unnamed	30-Sep	BD	Bangladesh	5
1971	Unnamed	30-Oct	IN	India	5
1974	Unnamed	15-Aug	BD	Bangladesh	7.1
1975	Porbandar Cyclone	October	IN	Just north of Porbandar India	2.8
1976	Unnamed	25-Nov	IN	Visakhapatam	0.22
1977	Andhra Cyclone	19-Nov	IN	Just southwest of Machilipatnam (near Divi Island)	5
1978	Unnamed	4-Nov	IN	Chennai	0.2
1981	Unnamed	10-Dec	BD	Bangladesh	2.2
1981	Unnamed	7-Aug	IN	Paradip	0.43
1982	Orissa Cyclone	3-Jun	IN	Dharma Port	3.5
1984	Unnamed	1-Aug	IN	Paradip	0.4
1986	Unnamed	12-Aug	IN	Visakhapatam	1.8
1989	Unnamed	29-Nov	BD	Bangladesh	3.3
1989	Kavali Cyclone	2-9 Nov	IN	Near Kottapatnam	4
1990	Andhra or Divi Cyclone	9-May	IN	Andhra Pradesh just SW of Krishna River	4.5
1992	Tuticorin Cyclone	11-17 Nov	IN	East of Tuticorin Tamil Nadu	1.5
1993	Karaikal Cyclone	2-4 Dec	IN	North of Karaikal	1
1996	Kakinada Cyclone	1-7 Nov	IN	Near Kakinada	4.1
1998	Kandla	June	IN	India	2.5
1999	Orissa Cyclone	Oct	IN	Near Paradip India	7.5
2001	Unnamed	May	PK	South of Kori Creek	1
2010	Laila	May 17-21	IN		0
2010	Jal	Nov 4-7	IN		0
2011	Thane	Dec 25-30	IN	Cuddalore Puducherry and Villuparam	1

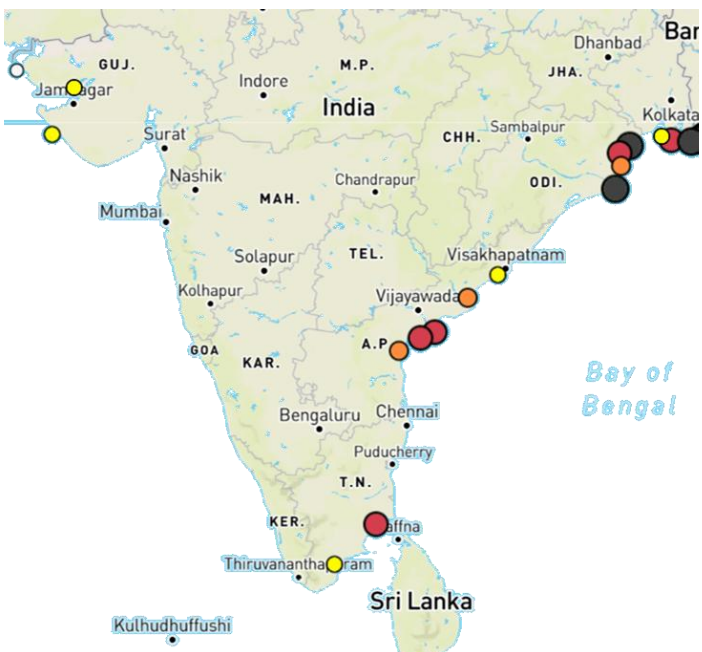
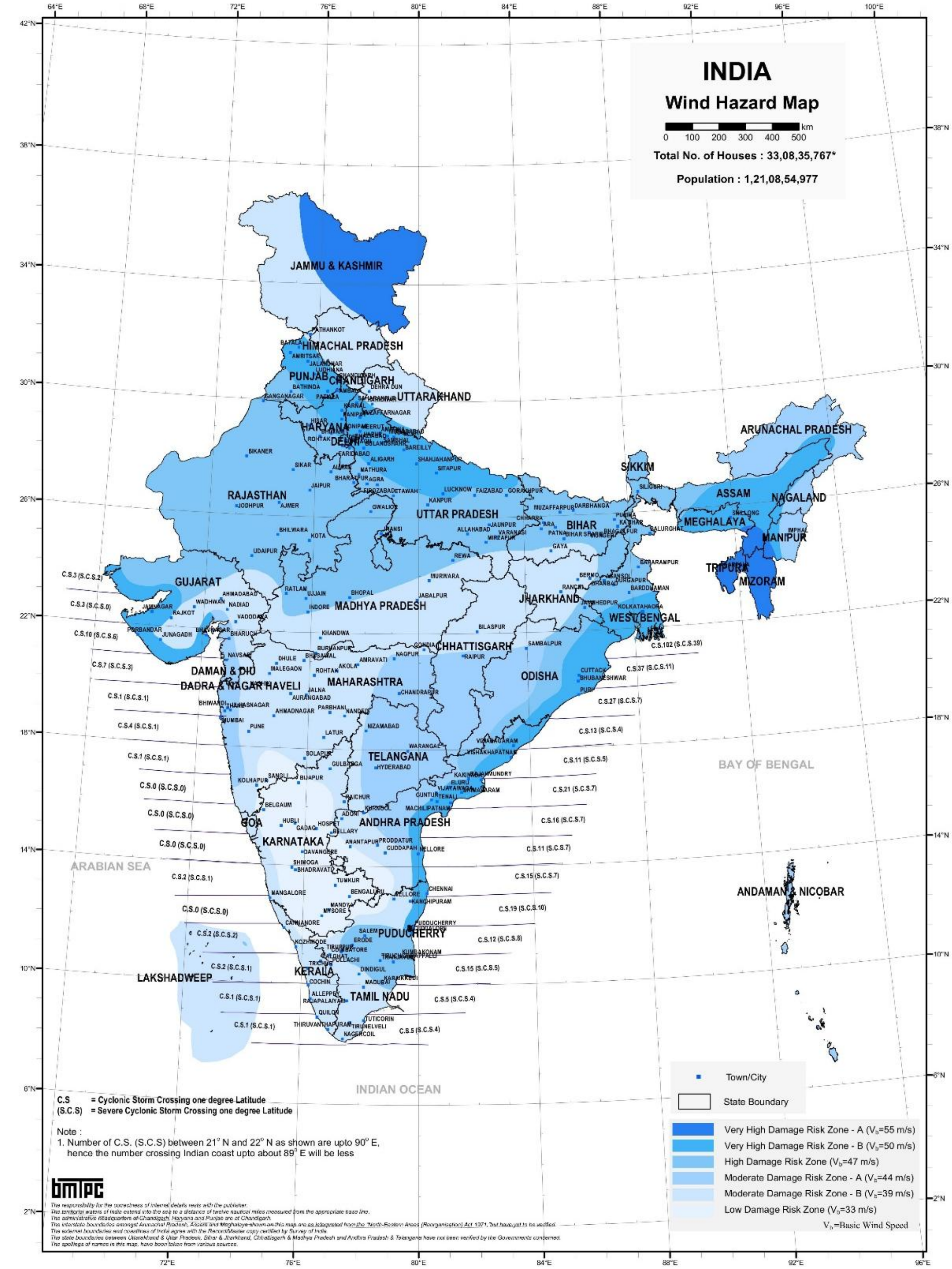


Figure S9: Locations of historical storm surges recorded from 1880-2015

S20: Cyclonic Storm Hazard Exposure (BMTPC 2019)

The western coastline of India is equally exposed to cyclonic storm and associated surge risk; any part of the coastline being hit by a storm surge or not in the past is a random event.



BMTPC : Vulnerability Atlas- 3rd Edition; Peer Group, MoHUA; Map is Based on digitised data of SOI, GOI; Basic Wind Speed Map National Building Code: 2016; Cyclone Data, 1891-2015, IMD, GOI. Houses/Population as per Census 2011; *Houses including vacant & locked houses. Disclaimer: The maps are solely for thematic presentation.

Figure S10: Cyclone exposure map of India. This shows equal exposure of "Very High Damage Risk" to the western coastline of the country, with the exception of southern part of Tamil Nadu, that has "High Damage Risk".

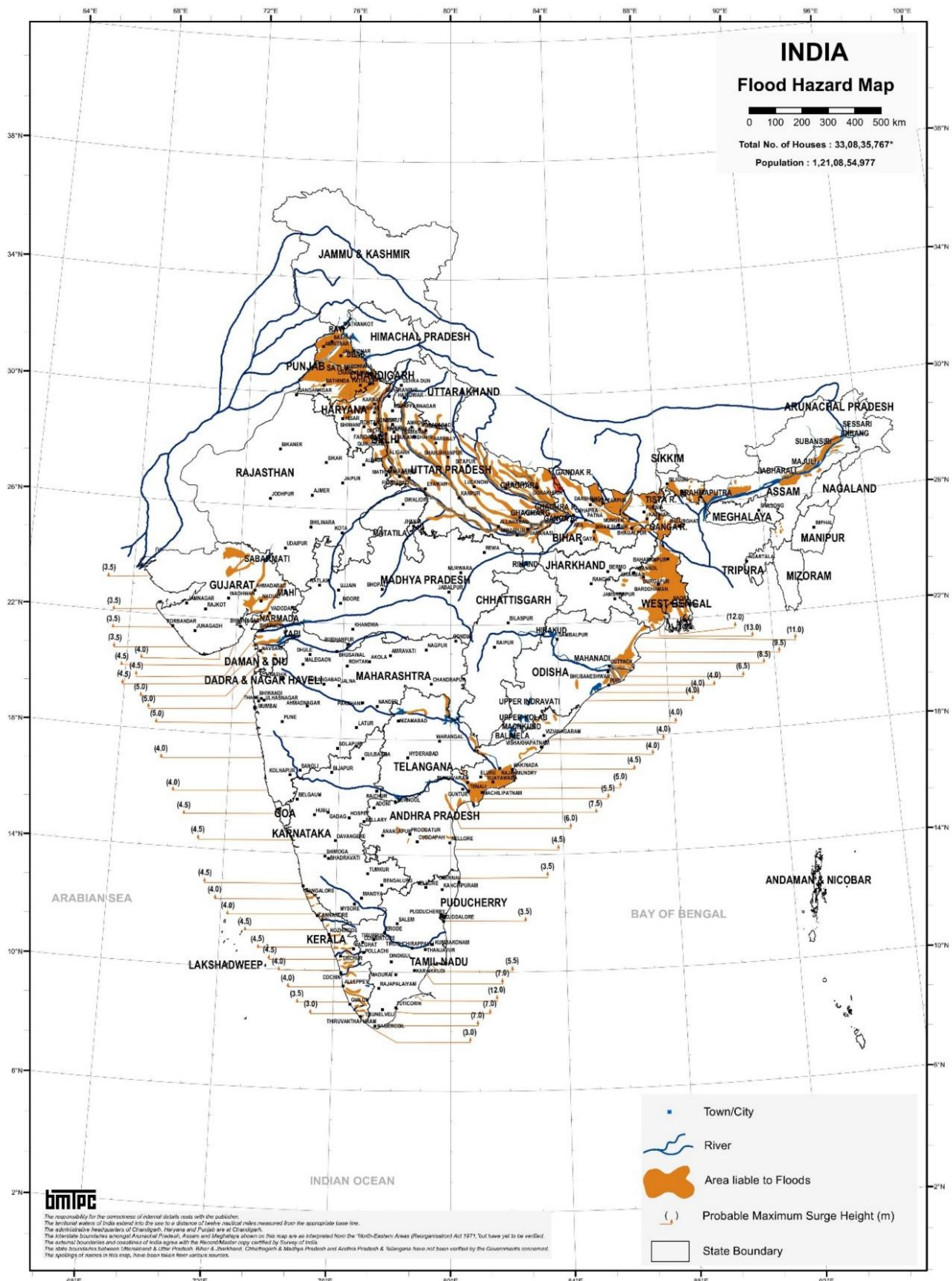


Figure S11: Flood and Surge Height Exposure Map of India. This shows that while the entire coast is equally susceptible to surges, the surge heights they are exposed to vary. These surge height variations are accounted in the surge effect modelling based on land surface elevation and surge heights of the events. Source: BMTPC, 2019 (BMTPC 2019).

S21: Discontinuity analysis for sensor differences in two time periods

Since this database integrates aquaculture and salinity data using two distinct Landsat satellite sources (Landsat 5 and 8), a discontinuity assessment was undertaken to ensure there are no residual differences due to satellite sensor differences.

1. **Aquaculture**

Table S10: Discontinuity analysis for aquaculture. A statistically significant but very small upward trend over time in aquaculture land share (0.023 percentage point per year). The coefficient for the sensor break (post-2012) is small and not statistically significant. This implies there is no evidence that aquaculture values jump discontinuously at 2013.
Aquaculture does not have discontinuity at 2013, and need not be normalized.

```
Equation:
> lm_discont <- lm(Aqua_perc ~ Year + I(Year >= 2013), data = aqua_salinity_surge)
> summary(lm_discont)
```

```
Residuals:
   Min     1Q Median     3Q    Max
-1.527 -1.258 -0.903 -0.582 98.431

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        -4.566e+01  1.714e+00 -26.630   <2e-16 ***
Year                2.322e-02  8.567e-04  27.107   <2e-16 ***
I(Year >= 2013)TRUE  1.551e-01  1.825e-02   8.502   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.629 on 886146 degrees of freedom
  (233019 observations deleted due to missingness)
Multiple R-squared:  0.004578,   Adjusted R-squared:  0.004575
F-statistic:  2038 on 2 and 886146 DF,  p-value: < 2.2e-16
```

2. **Salinity**

Table S11: Discontinuity analysis for salinity measure. This captures salinity percentage (Saline_perc) as a function of the year (to capture the trend) and a binary dummy (I(Year >= 2013)) that captures the effect of the Landsat sensor switch. There's a strong structural break in the salinity values starting in 2013, likely due to the switch from Landsat 5 to Landsat 8 despite the harmonization. Salinity declines by about 0.1 percentage point per year on average, all else equal. After 2013, salinity values drop by 4.32 percentage points on average, even after controlling for the time trend. This suggests that the Landsat 5 and 8 models are not on the same scale. We therefore, **normalize salinity** by sensor period. This will remove level differences between periods, but will retain within-period variation. Z = (x-mean)/SD. This is a standardized measure centered at 0. Values are also -ve here because of z-score standardization, which centers the data around 0 within each sensor period.
All interpretations to be read as: ***A one standard deviation increase in salinity (relative to the Landsat period's mean) is associated with a x percentage point increase/decrease in aquaculture land area percentage, holding storm and year constant***.

```
Equation:
lm_discont <- lm(Saline_perc ~ Year + I(Year >= 2013), data = aqua_salinity_surge)
```

```
summary(lm_discont)
```

```
Results:

Residuals:

   Min     1Q Median     3Q    Max

-9.843 -7.168 -2.589 -0.807 94.943

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept)        211.087856   4.434209   47.60   <2e-16 ***

Year               -0.101128   0.002216  -45.64   <2e-16 ***

I(Year >= 2013)TRUE  -4.321616   0.047249  -91.47   <2e-16 ***

---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.89 on 876418 degrees of freedom

 (242747 observations deleted due to missingness)

Multiple R-squared:  0.06397,     Adjusted R-squared:  0.06397

F-statistic: 2.995e+04 on 2 and 876418 DF,  p-value: < 2.2e-16
```

```
Normalization:
aqua_salinity_surge <- aqua_salinity_surge %>%
  mutate(period = ifelse(Year <= 2012, "L5", "L8")) %>%
  group_by(period) %>%
  mutate(Saline_perc_norm = scale(Saline_perc)) %>%
  ungroup()

str(aqua_salinity_surge$Saline_perc_norm)
aqua_salinity_surge$Saline_perc_norm <-
as.numeric(aqua_salinity_surge$Saline_perc_norm)
str(aqua_salinity_surge$Saline_perc_norm)
```

```
summary(aqua_salinity_surge$Saline_perc_norm)
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  -0.60   -0.51   -0.33    0.00   -0.08   12.20  242747
```

```
> mean_salinity_perc_13_25

[1] 2.5775

> sd_salinity_perc_13_25

[1] 7.766171
```

S22: Soil Data

Table S12: Primary Soil Data details. The table provides all the field observations including soil type, land use land cover and moisture, and lab results corresponding to the collected samples, including EC and pH, and moisture. The table also indicates the Grid Cell number corresponding to Figure 2.2 grids and aquaculture area in each grid at the time of data collection (February 2024).

Note: Geolocations are available in the GitHub repository

ID	Name	Soil Type	Land cover Land Use	Colour	Soil moisture	pH	EC	Aquaculture in Grid	Grid Cell ID
1	C-S01_20240213	Clayey	Fallow rice, recently planted moong dal	5Y 5/1 | Grey	Moist	6.04	488	104	C15
2	C-S02_20240213	Clayey	Fallow rice field	5Y 4/2 | Greyish olive	Matte	6.61	678.2	447	C14
3	C-S03_20240213	Clayey	Former agricultural land	5Y 4/2 | Greyish olive	Matte	6.72	504.4	112	C24
4	C-S04_20240213	Clayey	Fallow rice field	5Y 4/1 | Grey	Matte	6.52	370	165	C35
5	C-S05_20240213	Clayey	No agricultural use, no surrounding built-up	5Y 4/4 | Olive	Dry	5.6	515.2	165	C35
6	C-S06_20240213	Clayey	Fallow rice, recently planted moong dal	2.5Y 5/1 | Yellowish grey	Moist	5.85	891.1	300	C44
7	C-S07_20240213	Clayey	Fallow rice field	2.5Y 5/2 | Dark greyish yellow	Moist	5.48	98.1	300	C44
8	C-S08_20240213	Clayey	Fallow rice field	2.5Y 5/2 | Dark greyish yellow	Matte	5.31	177.3	242	C43
9	C-S09_20240213	Clayey	Fallow rice field	5Y 4/1 | Grey	Matte	5.79	283.9	242	C43
10	C-S10_20240214	Sandy clay	Other food crop	5Y 4/1 | Grey	Moist	6.09	139.5	104	C15
11	C-S11_20240214	Sandy clay	Other food crop	5Y 4/1 | Grey	Moist	6.01	417.2	104	C15

12	C-S12_20240214	Clayey	Fallow rice, recently planted moong dal	7.5Y 5/1 | Grey	Matte	6.05	478.7	112	C24
13	C-S13_20240214	Clayey	Fallow rice, recently planted moong dal	7.5Y 5/1 | Grey	Matte	6.07	181.8	209	C34
14	C-S14_20240214	Clayey	Fallow rice, recently planted moong dal	5Y 4/1 | Grey	Slightly dry	6.42	513	209	C34
15	C-S15_20240214	Clayey	Fallow rice, recently planted moong dal	2.5Y 4/1 | Yellowish grey	Moist	4.65	1080	409	C33
16	C-S16_20240214	Clayey	Fallow rice, recently planted moong dal	2.5Y 4/1 | Yellowish grey	Moist	4.86	786	409	C33
17	C-S17_20240214	Clayey	Fallow rice and grazing field	2.5Y 5/2 | Dark greyish yellow	Moist	5.37	380.6	409	C33
18	C-S18_20240214	Clayey	Fallow rice, recently planted moong dal	7.5Y 5/1 | Grey	Matte	5.82	1229	518	C23
19	C-S19_20240214	Clayey	Fallow rice, recently planted moong dal	7.5Y 5/1 | Grey	Matte	5.5	418.4	194	C22
20	C-S20_20240214	Clayey	Fallow rice field	7.5Y 5/1 | Grey	Matte	4.89	739.2	194	C22
21	C-S21_20240215	Clayey	Fallow rice field	2.5 Y5/4 | Grey	Very dry	5.16	937.2	1273	C13
22	C-S22_20240215	Clayey	Fallow rice field	2.5 Y 5/3 | Yellowish grey	Dry	5.54	1633	447	C14
23	C-S23_20240215	Clayey	Fallow rice, recently planted moong dal	2.5 Y 4/2 | Dark greyish yellow	Slightly moist	5.59	1448	1273	C13
24	C-S24_20240215	Clayey	Fallow rice field	10 YR 5/1 | Brownish grey	Moist	6.12	1338	518	C23
25	C-S25_2	Clayey	Fallow rice field	2.5Y 4/1 |	Moist	5.56	1849	2234	C12

	0240215			Yellowish grey					
26	C-S26_20240215	Clayey	Fallow rice field	2.5Y 4/1 | Yellowish grey	Moist	5.67	1993	2234	C12
27	C-S27_20240215	Clayey	Abandoned rice field turned wetland	2.5Y 3/3 | Dark olive brown	Wet	5.22	2260	2234	C12
28	C-S28_20240215	Clayey	Fallow rice field	2.5Y 5/2 | Dark greyish yellow	Matte	5.33	1991	3039	C11
29	C-S29_20240216	Clayey	Fallow rice field	2.5Y 5/1 | Yellowish grey	Matte	5.58	217.7	154	C25
30	C-S30_20240216	Clayey	Fallow rice field	2.5Y 5/2 | Dark greyish yellow	Matte	6.44	306.6	154	C25
31	C-S31_20240216	Clayey	Genda Phool'/ Marigold garden	2.5Y 4/4 | Brown	Slightly dry	5.66	650.7	103	C45
32	C-S32_20240216	Clayey	Fallow rice, recently planted moong dal	2.5 Y 5/3 | Yellowish grey	Dry | 2.5	5.46	263.5	103	C45
33	C-S33_20240216	Clayey	Fallow rice, recently planted moong dal	7.5 R 5/2 | Greyish red	Wet | >10	5.68	268.5	116	C55
34	C-S34_20240216	Clayey	Fallow rice, recently planted moong dal	7.5 R 5/2 | Greyish red	Wet | >10	5.69	264.4	116	C55
35	C-S35_20240216	Clayey	Fallow rice, recently planted moong dal	10R 5/1 | Reddish grey	Wet | >10	5.85	181.8	116	C55
36	C-S36_20240304	Sandy	Fallow rice, recently planted moong dal	2.5 Y 5/3 | Yellowish grey	9.17	5.87	71.1	72	C54
37	C-S37_20240304	Clayey	Fallow rice, recently planted moong dal	10YR 5/2 | Greyish yellow brown	12.18	6.12	127.8	72	C54
38	C-S38_20240304	Clayey	Fallow rice field	5YR 5/2 | Greyish brown	19.09	5.56	1647	278	C53

39	C-S39_20240304	Clayey	Fallow rice field	10YR 5/3 | Dull yellowish brown	16.89	6.1	348.2	278	C53
40	C-S40_20240304	Clayey	Rice field	7.5 YR 5/2 | Greyish brown	33.68	6.49	1452	278	C53
41	C-S41_20240304	Clayey	Fallow rice field	10YR 5/2 | Greyish yellow brown	22.1	7.13	1643	151	C52
42	C-S42_20240304	Clayey	Fallow rice field	10R 5/1 | Reddish grey	19.24	5.99	1556	151	C52
43	C-S43_20240304	Clayey	Abandoned field, no agricultural use	2.5YR 5/2 | Greyish red	22.44	7.63	3826	26	C51
44	C-S44_20240304	Clayey loamy	Fallow rice field	5Y 5/6 | Olive	15.44	6.06	1787	296	C42
45	C-S45_20240304	Clayey	Fallow rice field	2.5YR 5/3 | Dull reddish brown	30.86	4.85	562.8	296	C42
46	C-S46_20240304	Clayey	Fallow rice field	5YR 5/2 | Greyish brown	16.96	5.71	201	1495	C41
47	C-S47_20240305	Clayey sand	Fallow rice field	7.5 YR 4/3 | Brown	16.01	5	932.6	128	C32
48	C-S48_20240305	Clayey sand	Rice field	5Y 5/4 | Olive	38.27	5.02	321.3	128	C32
49	C-S49_20240305	Clayey sand	Rice field	5Y 5/4 | Olive	37.74	5.23	816.2	887	C31
50	C-S50_20240305	Clayey sand	Fallow rice field	7.5 YR 5/4 | Dull brown	21.5	4.97	922.4	1495	C41
51	C-S51_20240305	Clayey	Fallow rice field	2.5Y 4/6 | Olive brown	37.05	5.36	930.8	887	C31

52	C-S52_20240305	Clayey	Rice field	5Y 5/3 | Greyish olive	41.33	4.39	926.6	1296	C21
53	C-S53_20240305	Sandy	Forest cover	5Y 8/4 | Pale yellow	1.55	5.38	29.3	1296	C21
54	C-S54_20240305	Clayey sand	Fallow rice field	2.5Y 4/6 | Olive brown	33.5	4.81	1404	3039	C11
55	C-S55_20240305	Clayey sand	Fallow rice field	7.5 YR 4/6 | Brown	23.46	5.1	3312	26	C51
56	T-S01_20240217	Clayey	Fallow rice field	7.5 YR 5/1| Brownish grey		4.7	1245	1088	T35
57	T-S02_20240217	Clayey	Fallow rice field	7.5 YR 5/1| Brownish grey	Dry	5.27	756.4	1618	T25
58	T-S03_20240217	Clayey	Drying aquaculture pond	5Y 5/1 | Grey	Wet	6.64	2125	1618	T25
59	T-S04_20240217	Clayey	Fallow rice field	5Y 5/1 | Grey	Wet > 10	7.32	2745	2980	T24
60	T-S05_20240217	Clayey	Fallow rice field	7.5 YR 5/1| Brownish grey	Wet > 10	5.02	4485	3265	T23
61	T-S06_20240217	Clayey	Drying aquaculture pond	5Y 5/1 | Grey	Wet	6.77	2089	1618	T25
62	T-S07_20240217	Clayey	Drying aquaculture pond	7.5 YR 5/1| Brownish grey	Wet > 10	5.3	6410	3265	T23
63	T-S08_20240218	Clayey	Fallow rice field	2.5Y 5/6 | Yellowish grey	13.4	5.68	924.4	2150	T55
64	T-S09_20240218	Clayey sand	Fallow rice, recently planted moong dal	2.5Y 5/6 | Yellowish grey	10.8	6.24	160.2	2150	T55
65	T-S10_2	Clayey	Fallow rice, recently	2.5Y 5/4 |	13.6	6.37	130.8	1173	T44

	0240218		planted moong dal	Yellowish grey					
66	T-S11_20240218	Clayey	Fallow rice, recently planted moong dal	2.5Y 5/4 | Yellowish grey	10.6	6.33	405.1	3363	T54
67	T-S12_20240218	Clayey	Scrapped dry aquaculture pond	2.5Y 7/2 | Greyish yellow	36.3	5.69	6372	3363	T54
68	T-S13_20240218	Clayey sand	Fallow rice field	5Y 4/1 | Grey	8.5	6.9	1132	1113	T53
69	T-S14_20240218	Clayey	Fallow rice field	10R 4/1 | Dark reddish grey	25.3	5.9	1669	5707	T42
70	T-S15_20240218	Clayey	Fallow rice field	5Y 6/4 | Olive yellow	27.6	4.93	1428	1113	T53
71	T-S16_20240218	Clayey sand	Rice field	7.YR 5/1 | Brownish grey	35.3	5.29	142.4	479	T52
72	T-S17_20240218	Clayey	Drying aquaculture pond	2.5Y 5/3 | Yellowish grey	19.3	3.89	8242	479	T52
73	T-S18_20240218	Clayey	Scrapped dry aquaculture pond	2.5Y 7/2 | Greyish yellow	34.9	5.4	5867	3363	T54
74	T-S19_20240219	Clayey	Abandoned aquaculture pond	2.5Y 6/2 | Greyish yellow	60.16	7.65	8127	3763	T33
75	T-S20_20240219	Clayey	Fallow rice field	10YR 5/1 | Brownish grey	21.32	6.53	578.3	1088	T35
76	T-S21_20240219	Clayey	Fallow rice field	10YR 5/2 | Greyish yellow brown	25.35	6.33	6652	3651	T34
77	T-S22_20240219	Clayey	Fallow rice field	2.5Y 5/4 | Yellowish grey	21.74	5.4	753.6	3651	T34
78	T-S23_20240219	Clayey	Drying aquaculture pond	10YR 6/1 | Brownish grey	44.7	6.72	2228	3763	T33

79	T-S24_20240219	Clayey	Scrapped dry aquaculture pond	10YR 5/1 | Brownish grey	39.39	7.28	3634	6056	T32
80	T-S25_20240219	Clayey	Scrapped dry aquaculture pond	10YR 5/1 | Brownish grey	51.54	7.29	3630	6056	T32
81	T-S26_20240219	Clayey	No agricultural use, no surrounding built-up	2.5Y 6/1 | Greyish yellow	37.03	7.07	3548	6056	T32
82	T-S27_20240219	Clayey	Drying aquaculture pond	2.5Y 6/2 | Greyish yellow	24.52	8.53	2042	6056	T32
83	T-S28_20240219	Clayey	Dried aquaculture pond	2.5 Y 5/3 | Yellowish grey				5202	T31
84	T-S29_20240219	Clayey sand	River floodplain	10YR 6/3 | Dull yellow orange				5202	T31
85	T-S30_20240219	Clayey	Fallow rice field	2.5Y 6/4 | Dull yellow	12.64	4.83	2125	3758	T43
86	T-S31_20240220	Clayey	Fallow rice field	2.5YR 5/2 | Greyish red				1618	T25
87	T-S32_20240220	Sandy clay	Fallow rice field	2.5YR 5/2 | Greyish red				2487	T15
88	T-S33_20240220	Sandy clay	Fallow rice field	10YR 5/2 | Greyish yellow brown	14.58	4.83	675.5	2487	T15
89	T-S34_20240220	Clayey	Fallow field	7.5 YR 5/1| Brownish grey	32.26	5.7	2642	3022	T14
90	T-S35_20240220	Clayey	Drying aquaculture pond	2.5YR 5/1 | Reddish grey				3022	T14
91	T-S36_20240220	Clayey	Drying aquaculture pond	7.5YR 6/1 | Brownish grey	44.34	6.68	5005	2269	T13

92	T-S37_20240220	Clayey	Fallow rice field	10YR 6/1 | Brownish grey				2269	T13
93	T-S38_20240220	Clayey	Drying aquaculture pond	10YR 5/1 | Brownish grey	56.22	6.9	5745	3265	T23
94	T-S39_20240220	Sandy clay	Fallow rice field	10YR 6/2 | Greyish yellow brown				4348	T22
95	T-S40_20240220	Clayey	Scrapped dry aquaculture pond	2.5 Y 6/3 | Dull yellow	29.74	4.68	5563	4348	T22
96	T-S41_20240220	Clayey	Drying aquaculture pond	5Y 5/6 | Olive with black streaks	36.71	7.12	6542	2955	T12
97	T-S42_20240220	Clayey	Drying aquaculture pond	10YR 5/2 | Greyish yellow brown	37.94	6.81	5648	2955	T12
98	T-S43_20240220	Clayey	Fallow rice field	10YR 5/1 | Brownish grey				3389	T21
99	T-S44_20240220	Clayey sand	Rice field	10YR 5/2 | Greyish yellow brown	20.53	5.09	732.1	3389	T21
100	T-S45_20240306	Sandy clay	Fallow rice, recently planted moong dal	2.5YR 4/4 | Dull reddish brown	17.38	5.34	190.8	1884	T45
101	T-S46_20240306	Clayey sand	Fallow rice field	2.5YR 4/6 | Reddish brown	20.32	4.66	547.2	1884	T45
102	T-S47_20240306	Clayey	Fallow rice field	10YR 6/4 | Dull yellow orange	13.74	5.74	229.8	1173	T44
103	T-S48_20240306	Clayey sand	Fallow rice field	7.5YR 5/6 | Bright brown	8.37	5.48	425.5	3758	T43

104	T-S49_20240306	Clayey sand	Abandoned field, no agricultural use	2.5Y 5/6 | Yellowish grey	28.78	5.58	1777	3758	T43
105	T-S50_20240306	Clayey sand	Drying aquaculture pond	2.5Y 4/6 | Olive brown	42.65	6.47	5120	5707	T42
106	T-S51_20240306	Sandy	Rice field	5G 5/1 | Greenish grey	20.85	5.11	1678	2954	T41
107	T-S52_20240306	Sandy clay	Rice field	10Y 5/2 | Olive grey	35.63	6.31	676.1	2954	T41
108	T-S53_20240306	Sandy clay	Rice field	10Y 5/2 | Olive grey	44.06	6.82	488.2	2561	T51
109	T-S54_20240306	Sandy clay	Drying aquaculture pond	5Y 6/3 | Olive yellow	35.84	8.25	3040	2561	T51
110	T-S55_20240306	Clayey	Fallow rice field	2.5Y 4/6 | Olive brown	28.3	4.98	4530	2980	T24
111	T-S56_20240307	Clayey sand	Drying aquaculture pond	10YR 4/4 | Brown	43.62	4.58	3521	3389	T21
112	T-S57_20240307	Sandy	Abandoned field	10Y 5/2 | Olive grey	22.28	6.1	1242	4923	T11
113	T-S58_20240307	Sandy clay	Drying aquaculture pond	2.5Y 4/4 | Olive brown	28.65	7.13	1329	4923	T11

S23: IRB Approval

ASU Knowledge Enterprise Development

EXEMPTION GRANTED

Billie Turner

CLAS-SS: Geographical Sciences and Urban Planning, School of (SGSUP)

-

Billie.L.Turner@asu.edu

Dear Billie Turner:

On 1/17/2024 the ASU IRB reviewed the following protocol:

Type of Review:	Initial Study
Title:	Aquaculture land transformations in climate-exposed coastal India
Investigator:	Billie Turner
IRB ID:	STUDY00019294
Funding:	None
Grant Title:	None
Grant ID:	None

Documents Reviewed:	• CITI Certification External Consultant 1, Category: Off-site authorizations (school permission, other IRB approvals, Tribal permission etc); • CITI Certification Garima Jain, Category: Other; • Consent_FGD, Category: Consent Form; • Consent_HHSurvey, Category: Consent Form; • Consent_Interview, Category: Consent Form; • FGD_Questionnaire, Category: Measures (Survey questions/Interview questions /interview guides/focus group questions); • HH Survey_Type2, Category: Measures (Survey questions/Interview questions /interview guides/focus group questions); • HH_Survey_TYPE1, Category: Measures (Survey questions/Interview questions /interview guides/focus group questions); • Information Security and Management Ethics Certification for S Ramesh, Category: Off-site authorizations (school permission, other IRB approvals, Tribal permission etc); • KeyInformantInterview_Questionnaire, Category: Measures (Survey questions/Interview questions /interview guides/focus group questions); • Modification Response Letter, Category: Other; • Project Flyer, Category: Recruitment Materials; • Research ethics training Certificate for External Consultant S Ramesh, Category: Off-site authorizations (school permission, other IRB approvals, Tribal permission etc); • Social Behavior Protocol, Category: IRB Protocol;

The IRB determined that the protocol is considered exempt pursuant to Federal Regulations 45CFR46 (2)(ii) Tests, surveys, interviews, or observation (low risk), (4) Secondary research on data or specimens (no consent required) on 1/16/2024.

In conducting this protocol you are required to follow the requirements listed in the INVESTIGATOR MANUAL (HRP-103).

If any changes are made to the study, the IRB must be notified at research.integrity@asu.edu to determine if additional reviews/approvals are required. Changes may include but not limited to revisions to data collection, survey and/or interview questions, and vulnerable populations, etc.

Sincerely,

IRB Administrator cc: Garima Jain

Garima Jain Dylan Connor Hallie Eakin

S24: Supplementary References

Abbas, Akhtar, and Shahbaz Khan. 2007. "Using Remote Sensing Techniques for Appraisal of Irrigated Soil Salinity," 2632–38.

Abbas, Akhtar, Shahbaz Khan, Nisar Hussain, Munir A. Hanjra, and Saud Akbar. 2013. "Characterizing Soil Salinity in Irrigated Agriculture Using a Remote Sensing Approach." *Physics and Chemistry of the Earth (2002)* 55–57 (January): 43–52.

Alemohammad, Hamed. 2024. "Aquaculture Data Download." Geospatial Analytics, Clark University. November 19, 2024. https://www.clarku.edu/centers/geospatial-analytics/aquaculture-data-download/.

Alhammadi, M. S., and E. P. Glenn. 2008. "Detecting Date Palm Trees Health and Vegetation Greenness Change on the Eastern Coast of the United Arab Emirates Using SAVI." *International Journal of Remote Sensing* 29 (6): 1745–65.

Al-Khaier, Andrew D. Richardson, Shane P. Duigan, and Graeme P. Berlyn. 2003. "Index: Normalized Difference Salinity Index." Index Database. 2003. https://www.indexdatabase.de/db/i-single.php?id=57.

BMTPC. 2019. "Flood Hazard Map." BMTPC. 2019. https://vai.bmtpc.org/flood.html.

Buchhorn, M., B. Smets, L. Bertels, B. D. Roo, M. Lesiv, N-E Tsendbazar, M. Herold, and S. Fritz. 2020. "*Copernicus Global Land Service: Land Cover 100m: Collection 3: Epoch 2015: Globe* (Version V3.0.1)." *Copernicus*. https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_Landcover_100m_Proba-V-C3_Global.

Census of India. 2001. "India Village and Town Boundaries 2001." 2001. https://sedac.ciesin.columbia.edu/downloads/metadata/india-india-village-level-geospatial-socio-econ-1991-2001.html.

Copernicus. 2013. "Copernicus DEM GLO-30: Global 30m Digital Elevation Model." Google for Developers. 2013. https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_DEM_GLO30.

———. 2015. "Harmonized Sentinel-2 MSI: MultiSpectral Instrument, Level-2A (SR)." *European Union/ESA*. https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S2_SR_HARMONIZED.

Dagar, J. C. 2005. "Salinity Research in India: An Overview." *Bulletin of the National Institute of Ecology* 15: 69–80.

Dehni, Abdellatif, and Mourad Lounis. 2012. "Remote Sensing Techniques for Salt Affected Soil Mapping: Application to the Oran Region of Algeria." *Procedia Engineering* 33: 188–98.

Douaoui, Abd El Kader, Hervé Nicolas, and Christian Walter. 2006. "Detecting Salinity Hazards within a Semiarid Context by Means of Combining Soil and Remote-Sensing Data." *Geoderma* 134 (1–2): 217–30.

Earth Resources Observation and Science (EROS) Center. 2020. "Landsat 4-5 Thematic Mapper Level-2, Collection 2." *U.S. Geological Survey*. https://doi.org/10.5066/P9IAXOVV.

Earth Resources Observation and Science (EROS) Center. 2020. "Landsat 8-9 Operational Land Imager / Thermal Infrared Sensor Level-2, Collection 2." *U.S. Geological Survey*. https://doi.org/10.5066/P9OGBGM6.

European Commission. 2016. "Global Human Settlement." *Global Human Settlement*. https://human-settlement.emergency.copernicus.eu/download.php?ds=pop.

Gao, Ji-Ping, Dai-Yin Chao, and Hong-Xuan Lin. 2007. "Understanding Abiotic Stress Tolerance Mechanisms: Recent Studies on Stress Response in Rice." *Journal of Integrative Plant Biology* 49 (6): 742–50.

Gorelick, N., M. Hancher, M. Dixon, and S. Ilyushchenko. 2017. “Google Earth Engine: Planetary-Scale Geospatial Analysis for Everyone.” *Remote Sensing of Environment*. https://www.sciencedirect.com/science/article/pii/S0034425717302900.

Grattan, Stephen R., Linghe Zeng, Michael C. Shannon, and Stacy R. Roberts. 2002. “Rice Is More Sensitive to Salinity than Previously Thought.” *California Agriculture* 56 (6): 189–95.

Henrich, Verena, Gunther Krauss, Christian Götze, and Christopher Sandow. 2012. “IDB - Index DataBase.” Index DataBase. 2012. https://www.indexdatabase.de/.

IMD, Climate Monitoring and Prediction Group. 2024. “Yearly Gridded Rainfall (0.25 x 0.25) Data NetCDF File.” IMD. 2024. https://imdpune.gov.in/cmpg/Griddata/Rainfall_25_NetCDF.html.

Jain, Garima, Wenxin Yang, Showmitra Sarkar, and Dylan S. Connor Amy Frazier. 2025. “CoSal: A Remote Sensing and Machine Learning Framework for Mapping Coastal Soil Salinity Trends around Aquaculture in South Asia.” *Res. Sq.* https://doi.org/10.21203/rs.3.rs-6173117/v1.

Jennewein, Douglas M., Johnathan Lee, Chris Kurtz, William Dizon, Ian Shaeffer, Alan Chapman, Alejandro Chiquete, et al. 2023. “The Sol Supercomputer at Arizona State University.” In *Practice and Experience in Advanced Research Computing*. New York, NY, USA: ACM. https://doi.org/10.1145/3569951.3597573.

Joseph, E. Abhilash. 2013. “A Study on the Effect of Salinity Stress on the Growth and Yield of Some Native RiceCultivars of Kerala State of India.” *Agriculture Forestry and Fisheries* 2 (3): 141.

Kalaiyarasi, R., V. Kuralarasan, Jeena George, N. M. Praveen, and V. Manikandan. 2019. “Salinity Tolerance Screening in Local Rice Varieties of Tamil Nadu and Kerala.” *International Journal for Chemical Studies* 7 (4): 1667–71.

Keim, Barry. 2015. “SURGEDAT: The World’s Storm Surge Information Center.” *SURGEDAT: Global Peak Surge Map*. https://surge.climate.lsu.edu/data.html#GOM_DB.

Khan, N., V. V. Rastoskuev, E. Shalina, and Yohei Sato. 2001. “Mapping Salt-Affected Soils Using Remote Sensing Indicators-A Simple Approach with the Use of GIS IDRISI -.” https://acrs-aars.org/proceeding/ACRS2001/Papers/AGS-05.pdf.

Krishnamurthy, Saraswathipura L., Byragondlu M. Lokeshkumar, Suman Rathor, Arvinder S. Warraich, Satyendra Yadav, Raj K. Gautam, Rakesh Kumar Singh, and Parbodh C. Sharma. 2022. “Development of Salt-Tolerant Rice Varieties to Enhancing Productivity in Salt-Affected Environments.” In *The 2nd International Laayoune Forum on Biosaline Agriculture*, 59:30. Basel Switzerland: MDPI.

Li, Wangchao, Rui Chen, Dujuan Ma, Changjing Wang, Yajie Yang, Cong Wang, Huai Chen, and Gaofei Yin. 2023. “Tracking Autumn Photosynthetic Phenology on Tibetan Plateau Grassland with the Green–Red Vegetation Index.” *Agricultural and Forest Meteorology* 339 (109573): 109573.

Mahata, K. R., D. P. Singh, S. Saha, A. M. Ismail, and S. M. Haefele. 2010. “Improving Rice Productivity in the Coastal Saline Soils of the Mahanadi Delta of India through Integrated Nutrient Management.” In *Tropical Deltas and Coastal Zones: Food Production, Communities and Environment at the Land and Water Interface*, 239–48. UK: CABI.

Mandal, T. 2023. “Intervention of Soil Salinity in Agriculture of Indian Sundarbans: A Review.” *International Journal of Bioresource Science* 10 (1): 117–27.

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): 1425–32.

Morshed, Md Manjur, Showmitra Kumar Sarkar, Md Rashed Uz Zzaman, and Md Mazharul Islam. 2021. "Application of Remote Sensing for Salinity Based Coastal Land Use Zoning in Bangladesh." *Spatial Information Research* 29 (3): 353–64.

Nguyen, Kim-Anh, Yuei-An Liou, Ha-Phuong Tran, Phi-Phung Hoang, and Thanh-Hung Nguyen. 2020. "Soil Salinity Assessment by Using Near-Infrared Channel and Vegetation Soil Salinity Index Derived from Landsat 8 OLI Data: A Case Study in the Tra Vinh Province, Mekong Delta, Vietnam." *Progress in Earth and Planetary Science* 7 (1): 1–16.

Pekel, Jean-François, Andrew Cottam, Noel Gorelick, and Alan S. Belward. 2016. "High-Resolution Mapping of Global Surface Water and Its Long-Term Changes." *Nature* 540 (7633): 418–22.

Perikamana, K.K., Parath, N., & Balakrishnan, K. 2025. "*An Improved Method for Harmonizing Landsat 5-TM and Landsat 8-OLI Data.*" *(Manuscript in Preparation)*.

Roberts, Dale, Norman Mueller, and Alexis Mcintyre. 2017. "High-Dimensional Pixel Composites from Earth Observation Time Series." *IEEE Transactions on Geoscience and Remote Sensing: A Publication of the IEEE Geoscience and Remote Sensing Society* 55 (11): 6254–64.

Ronald Eastman, J., Stefano C. Crema, Florencia Sangermano, Sean Cunningham, Xiao Xiao, Zhuoyue Zhou, Po Hu, Caroline Johnson, Bernadette Arakwiye, and Nelson Crone. 2025. "A Baseline Mapping of Aquaculture and Coastal Habitats in Thailand, Cambodia and Vietnam." 1. Clark University. https://s45055.pcdn.co/centers/geospatial-analytics/www-content/blogs.dir/7/files/sites/354/2024/12/Aquaculture-and-Coastal-Habitats-Report-No1.docx.pdf.

Roy, D. P., V. Kovalskyy, H. K. Zhang, E. F. Vermote, L. Yan, S. S. Kumar, and A. Egorov. 2016. "Characterization of Landsat-7 to Landsat-8 Reflective Wavelength and Normalized Difference Vegetation Index Continuity." *Remote Sensing of Environment* 185 (Iss 1): 57–70.

Sarkar, Showmitra Kumar, Rhyme Rubayet Rudra, Md Sadmin Nur, and Palash Chandra Das. 2023. "Partial Least-Squares Regression for Soil Salinity Mapping in Bangladesh." *Ecological Indicators* 154 (110825): 110825.

Sarkar, Showmitra Kumar, Rhyme Rubayet Rudra, Abid Reza Sohan, Palash Chandra Das, Khondaker Mohammed Mohiuddin Ekram, Swapan Talukdar, Atiqur Rahman, Edris Alam, Md Kamrul Islam, and Abu Reza Md Towfiqul Islam. 2023. "Coupling of Machine Learning and Remote Sensing for Soil Salinity Mapping in Coastal Area of Bangladesh." *Scientific Reports* 13 (1): 17056.

Shang, Rong, and Zhe Zhu. 2019. "Harmonizing Landsat 8 and Sentinel-2: A Time-Series-Based Reflectance Adjustment Approach." *Remote Sensing of Environment* 235 (111439): 111439.

Shrivastava, Pooja, and Rajesh Kumar. 2015. "Soil Salinity: A Serious Environmental Issue and Plant Growth Promoting Bacteria as One of the Tools for Its Alleviation." *Saudi Journal of Biological Sciences* 22 (2): 123–31.

Singh, Y. P., V. K. Mishra, Sudhanshu Singh, D. K. Sharma, D. Singh, U. S. Singh, R. K. Singh, S. M. Haefele, and A. M. Ismail. 2016. "Productivity of Sodic Soils Can Be Enhanced through the Use of Salt Tolerant Rice Varieties and Proper Agronomic Practices." *Field Crops Research* 190 (April): 82–90.

Suharyadi, R., Deha Agus Umarhadi, Disyacitta Awanda, and Wirastuti Widyatmanti. 2022. "Exploring Built-up Indices and Machine Learning Regressions for Multi-Temporal Building Density Monitoring Based on Landsat Series." *Sensors (Basel, Switzerland)* 22 (13): 4716.

USGS. 2022. “Why Are Negative Values Observed over Water in Some Landsat Surface Reflectance Products?” USGS. November 4, 2022. https://www.usgs.gov/faqs/why-are-negative-values-observed-over-water-some-landsat-surface-reflectance-products.

———. 2023. “How Do I Use a Scale Factor with Landsat Level-2 Science Products?” USGS. October 17, 2023. https://www.usgs.gov/faqs/how-do-i-use-a-scale-factor-landsat-level-2-science-products.

———. n.d. “Landsat Project Documents.” USGS. Accessed February 12, 2025. https://www.usgs.gov/landsat-missions/landsat-project-documents.

Wang, Fu-Min, Jing-Feng Huang, Yan-Lin Tang, and Xiu-Zhen Wang. 2007. “New Vegetation Index and Its Application in Estimating Leaf Area Index of Rice.” *Rice Science* 14 (3): 195–203.

Yousefian, F., M. Sahebi, M. Shokri, and M. Moradi. 2019. “A Novel Water Index (Swi) for Salty Water from Landsat 8 Oli/TIRs.” *ISPRS - International Archives of the Photogrammetry Remote Sensing and Spatial Information Sciences* XLII-4/W18 (October): 1097–1105.

Zeng, Linghe, Michael C. Shannon, and Scott M. Lesch. 2001. “Timing of Salinity Stress Affects Rice Growth and Yield Components.” *Agricultural Water Management* 48 (3): 191–206.