ENDSEM REPORT

ML meets sun: smart mapping for Solar plants

Name: J. Manish **Roll. No.: 25M0612**

Aim: To map Solar Suitability zones using Machine Learning and GIS

Introduction:

Solar energy is essential for meeting our increasing energy needs while also cutting down on our reliance on fossil fuels. To find the best spots for solar power plants, we need to consider a mix of geospatial, climatic, and topographic factors. Traditional methods that use GIS-based multi-criteria decision analysis often depend on set weights, which might not accurately reflect the complex, non-linear relationships among the various influencing factors.

In this study, we take a different approach by using a GIS-based supervised regression method to create a continuous solar suitability map. Instead of relying on existing solar plant locations as training labels, we utilize a physically derived solar suitability index as our target variable. This method emphasizes pinpointing potential areas for solar development rather than just mimicking past deployment trends.

Study Area: Telangana State, India.

Tools to be used: Google Earth Engine, QGIS/ArcGIS Pro, Python for Machine Learning.

Datasets:

i. Primary Datasets

Climatic and Solar Datasets

- Global Horizontal Irradiance (GHI) - ERA5-land
- Cloud cover – MODIS GEE Dataset
- Land Surface Temperature - ERA5, WorldClim
- Rainfall – CHIRPS

Topographic Datasets

- Digital Elevation Model (DEM) – USGS Earth Explorer ASTER GDEM

Land and Environmental Datasets

- Land Use / Land Cover (LULC) – ESRI Sentinel 2 Landcover Explorer
- Protected areas and water bodies (for exclusion): Source1 Source2

Infrastructure Datasets

- Road network - OSM Data by Geofabrik
- Settlement locations – Extracted from ESRI LULC Layer

ii. Derived Layers

From the primary datasets, several derivative layers are generated:

- **Slope** - derived from USGS-EE ASTER GDEM
- **Aspect** - derived from USGS-EE ASTER GDEM
- **Distance to roads –** derived from Road Network data (shapefile)
- **Exclusion masks** – for water bodies, forests, and urban areas

All datasets are projected to a common coordinate reference system of **WGS 1984 UTM Zone 44N** and resampled to a uniform spatial resolution of **100 meters.**

Methodology:

1. Data Pre-Processing:

The data were downloaded for the Telangana region. Each dataset had a different spatial resolution. The GHI data were relatively coarse, with a resolution of approximately 9 km, while the cloud fraction data had a resolution of 500 m. Rainfall data were collected from CHIRPS through Google Earth Engine (GEE) with a resolution of about 5 km. Temperature data at 2 m with a resolution of 4.5km were acquired from WorldClim and ERA5-Land. The DEM was downloaded for the available dates of the year 2015 at a resolution of 30 m. The land use/land cover (LULC) layer was acquired at a 10 m resolution. Protected area data were obtained as shapefiles from the two given sources. Settlement data were extracted from the ESRI LULC dataset, and a settlement proximity map was generated at a resolution of 100 m.

Road data were also used to generate a road proximity map at 100 m resolution, in line with the target resolution. Aspect and slope maps were prepared by filling the required DEM tiles, mosaicking them, and deriving the parameters using ArcGIS Pro.

Except for layers such as LULC, DEM, and shapefiles, all datasets were collected for the period 2014–2024 and converted into annual averages using both GEE scripts and Python for easier processing.

All layers were reclassified to a spatial resolution of 100 m. Nearest neighbour and bilinear interpolation methods were applied where appropriate. Since the data layers had different units, normalization was performed using Python. Direct normalization was applied to GHI and settlement proximity layers. Inverse normalization was applied to temperature, rainfall, railway proximity, road proximity, cloud cover, and slope layers.

Aspect encoding was carried out by assigning higher values to south-facing slopes (scaled from 1 to 0), with values decreasing toward the north on both eastern and western sides. LULC encoding was also performed on a scale of 1 to 0, where rangelands were assigned a value of 1, barren land 0.75, cropland, vegetation forests, and water bodies and built-up areas were assigned a value of 0.

Finally, all layers were reprojected to WGS 1984 UTM Zone 44N and resampled to a spatial resolution of 100 m. These all layers are aligned such that it exactly matches resampled GHI Layer to stack all the layers for Machine learning model using python code making sure all layers share same rows and columns.

2. Target Variable processing:

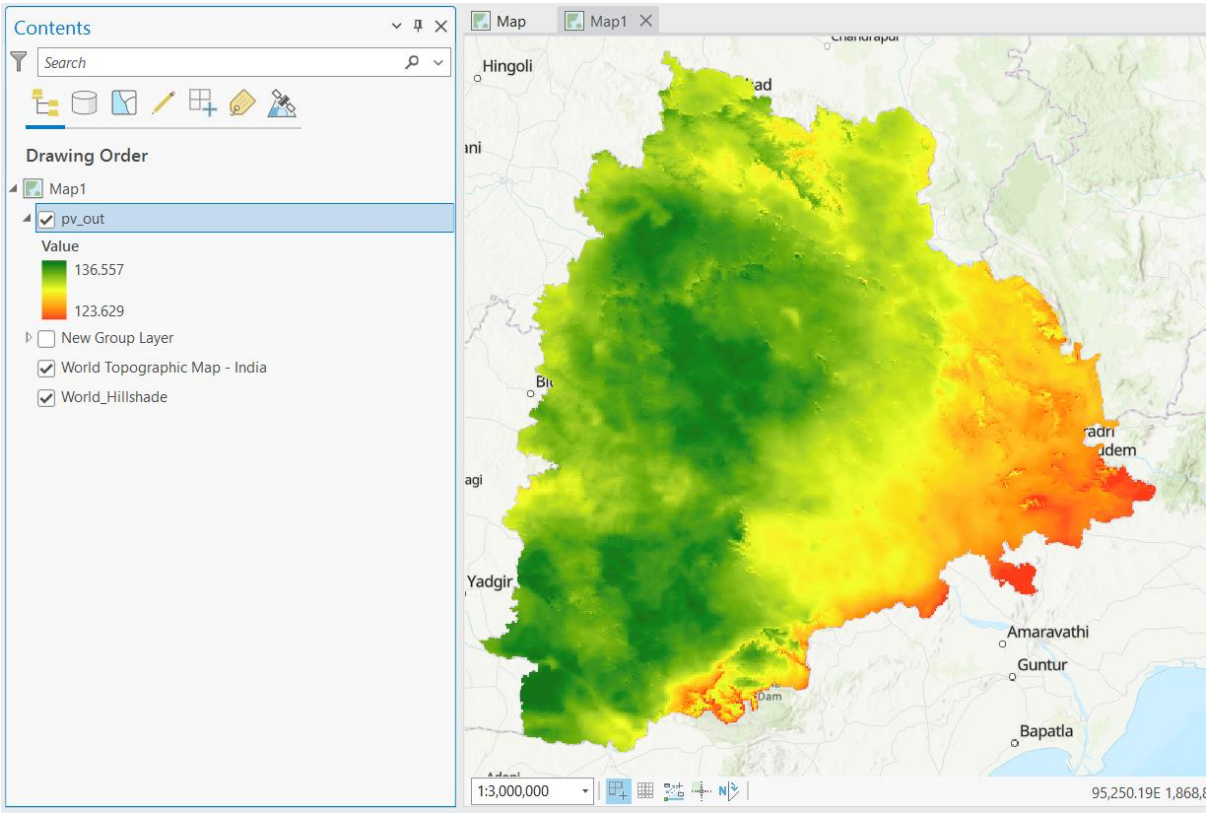
The Photovoltaic (PV_out) dataset used in this research comes from the Global Solar Atlas (GSA) and serves as the key variable for evaluating solar suitability. PV_out indicates the long-term average potential for photovoltaic power output under optimal system conditions, based on satellite data, atmospheric models, and ground validation. The Global Solar Atlas, developed by the World Bank Group in collaboration with Solargis, provides high-resolution global solar resource data, including solar irradiation, photovoltaic power potential, and various climatic factors. This dataset is widely utilized for renewable energy planning, policy formulation, and assessing the viability of solar projects, thanks to its scientific reliability and extensive global coverage.

For this study, we gathered and compiled PV_out data from January to December to create an annual average photovoltaic potential layer. This dataset reflects the long-term yearly average conditions from 1999 to 2018, ensuring that short-term weather variations don't distort the results. The PV_out layer is closely linked to Diffuse Horizontal Irradiation (DIF) and other radiation factors that influence solar panel efficiency, measured in kWh/m². I specifically downloaded data for India and clipped it to match the boundaries of Telangana state.

The original spatial resolution of the dataset (~946 m) was refined to 100 m using ArcGIS Pro to ensure it aligns with the resolution of other thematic layers in the suitability analysis. Preprocessing steps included standardizing projections, clipping rasters, resampling, and normalizing values. Within Telangana, PV_out values ranged from 123.629 to 136.557, indicating

moderate spatial variability in photovoltaic potential across the state. The processed raster is continuous, allowing for detailed pixel-level analysis of solar resource distribution.

The Solar Suitability Index (SSI) layer showed a clear link to the current patterns of solar installations, which really backs up the reliability of the dataset. The western regions of Telangana emerged as the most suitable and active areas, thanks to abundant sunlight and favourable weather conditions. As you head east, the suitability starts to taper off, probably because of increased cloud cover, higher humidity, and diverse terrain. Overall, the PV_out dataset from the Global Solar Atlas provides a strong, scientifically supported foundation for identifying the prime locations for solar energy development.

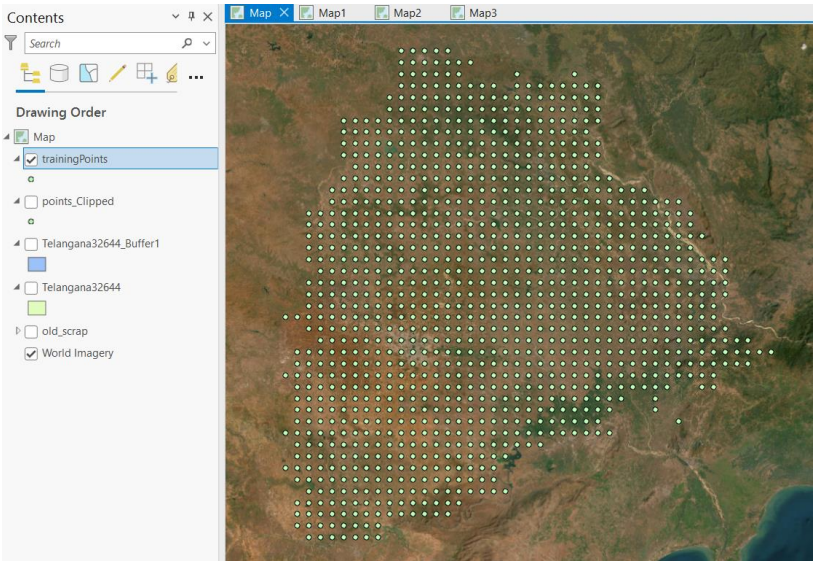


Photovoltaic Data download from Global Solar Atlas (GSA)

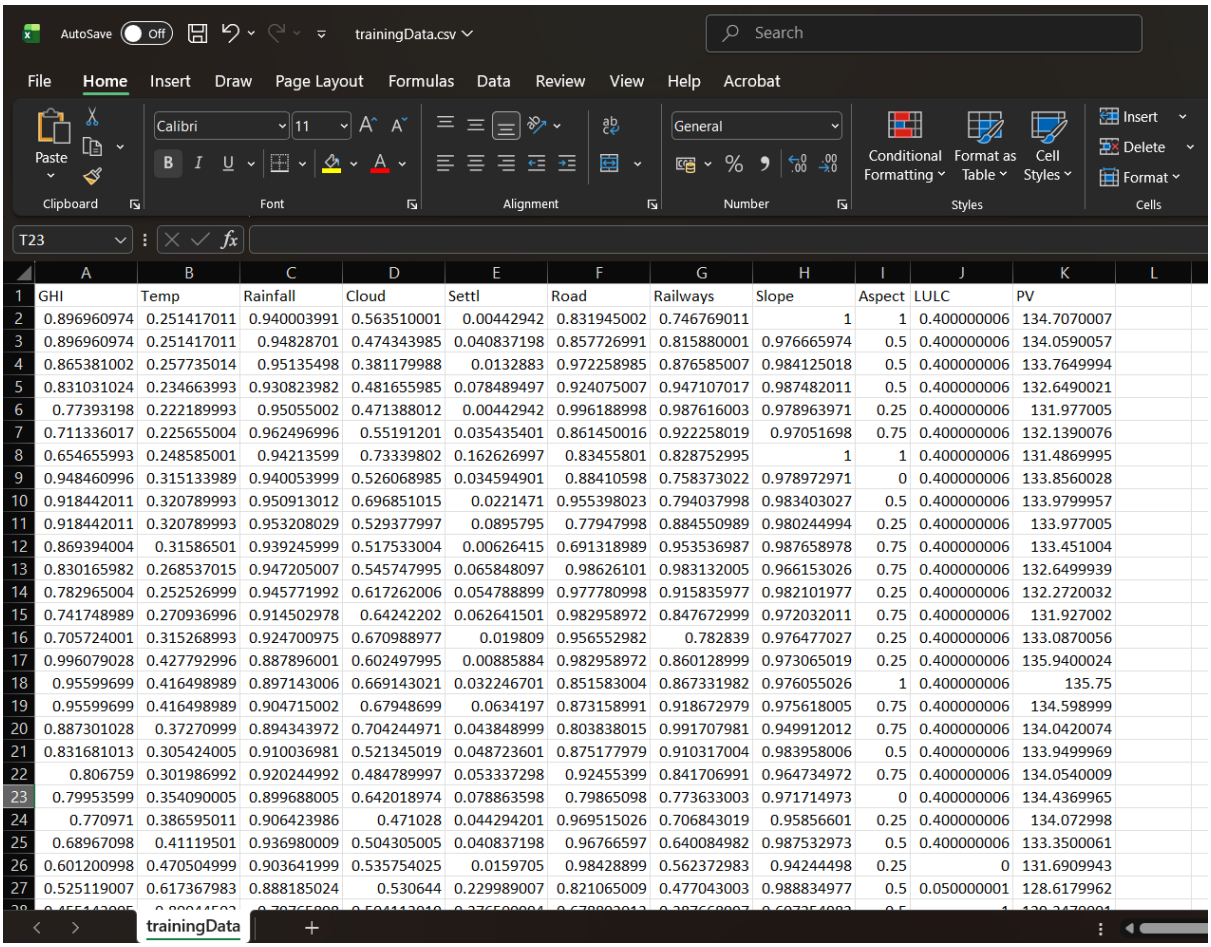
3. Training Data Generation:

To train the model, approximately 60–70% of the original GHI resolution was used for sampling points, as using a finer resolution than the GHI data would result in replication of the same GHI values. A total of 1,114 samples were selected for training and validation. For each sample, values from all 10 input (X) variables, along with the corresponding SSI values as the target variable (Y), were extracted, resulting in a total of 11 fields in the dataset.

This process was carried out in ArcGIS Pro. The point-based training dataset was then converted to a CSV file using the ArcGIS Pro geoprocessing toolbox and the Excel “Save As” option, to make it compatible with the machine learning model. The final CSV dataset therefore consists of 1,114 rows and 11 columns.



Training and Validation points



	A	B	C	D	E	F	G	H	I	J	K	L
1	GHI	Temp	Rainfall	Cloud	Settl	Road	Railways	Slope	Aspect	LULC	PV	
2	0.896960974	0.251417011	0.940003991	0.563510001	0.00442942	0.831945002	0.746769011	1	1	0.400000006	134.7070007	
3	0.896960974	0.251417011	0.94828701	0.474343985	0.040837198	0.857726991	0.815880001	0.976665974	0.5	0.400000006	134.0590057	
4	0.865381002	0.257735014	0.95135498	0.381179988	0.0132883	0.972258985	0.876585007	0.984125018	0.5	0.400000006	133.7649994	
5	0.831031024	0.234663993	0.930823982	0.481655985	0.078489497	0.924075007	0.947107017	0.987482011	0.5	0.400000006	132.6490021	
6	0.77393198	0.222189993	0.95055002	0.471388012	0.00442942	0.996188998	0.987616003	0.978963971	0.25	0.400000006	131.977005	
7	0.711336017	0.225655004	0.962496996	0.55191201	0.035435401	0.861450016	0.922258019	0.97051698	0.75	0.400000006	132.1390076	
8	0.654655993	0.248585001	0.94213599	0.73339802	0.162626997	0.83455801	0.828752995	1	1	0.400000006	131.4869995	
9	0.948460996	0.315133989	0.940053999	0.526068985	0.034594901	0.88410598	0.758373022	0.978972971	0	0.400000006	133.8560028	
10	0.918442011	0.320789993	0.950913012	0.696851015	0.0221471	0.955398023	0.794037998	0.983403027	0.5	0.400000006	133.9799957	
11	0.918442011	0.320789993	0.953208029	0.529377997	0.0895795	0.77947998	0.884550989	0.980244994	0.25	0.400000006	133.977005	
12	0.869394004	0.31586501	0.939245999	0.517533004	0.00626415	0.691318989	0.953536987	0.987658978	0.75	0.400000006	133.451004	
13	0.830165982	0.268537015	0.947205007	0.545747995	0.065848097	0.98626101	0.983132005	0.966153026	0.75	0.400000006	132.6499939	
14	0.782965004	0.252526999	0.945771992	0.617262006	0.054788899	0.977780998	0.915835977	0.982101977	0.25	0.400000006	132.2720032	
15	0.741748989	0.270936996	0.914502978	0.64242202	0.062641501	0.982958972	0.847672999	0.972032011	0.75	0.400000006	131.927002	
16	0.705724001	0.315268993	0.924700975	0.670988977	0.019809	0.956552982	0.782839	0.976477027	0.25	0.400000006	133.0870056	
17	0.996079028	0.427792996	0.887896001	0.602497995	0.00885884	0.982958972	0.860128999	0.973065019	0.25	0.400000006	135.9400024	
18	0.95599699	0.416498989	0.897143006	0.669143021	0.032246701	0.851583004	0.867331982	0.976055026	1	0.400000006	135.75	
19	0.95599699	0.416498989	0.904715002	0.67948699	0.0634197	0.873158991	0.918672979	0.975618005	0.75	0.400000006	134.598999	
20	0.887301028	0.37270999	0.894343972	0.704244971	0.043848999	0.803838015	0.991707981	0.949912012	0.75	0.400000006	134.0420074	
21	0.831681013	0.305424005	0.910036981	0.521345019	0.048723601	0.875177979	0.910317004	0.983958006	0.5	0.400000006	133.9499969	
22	0.806759	0.301986992	0.920244992	0.484789997	0.053337298	0.92455399	0.841706991	0.964734972	0.75	0.400000006	134.0540009	
23	0.79953599	0.354090005	0.899688005	0.642018974	0.078863598	0.79865098	0.773633003	0.971714973	0	0.400000006	134.4369965	
24	0.770971	0.386595011	0.906423986	0.471028	0.044294201	0.969515026	0.706843019	0.95856601	0.25	0.400000006	134.072998	
25	0.68967098	0.41119501	0.936980009	0.504305005	0.040837198	0.96766597	0.640084982	0.987532973	0.5	0.400000006	133.3500061	
26	0.601200998	0.470504999	0.903641999	0.535754025	0.0159705	0.98428899	0.562372983	0.94244498	0.25	0	131.6909943	
27	0.525119007	0.617367983	0.888185024	0.530644	0.229989007	0.821065009	0.477043003	0.988834977	0.5	0.050000001	128.6179962	

CSV consisting of 1114 rows x 11 columns

4. Regression Modelling:

A Random Forest algorithm was used for the regression modelling. The input variable set (X) consisted of all columns except the SSI variable, while SSI was used as the target variable (Y). The dataset was split into training and validation sets using a 70:30 ratio. The model achieved an R^2 value of 0.869 and an RMSE of 0.662 kWh/m².

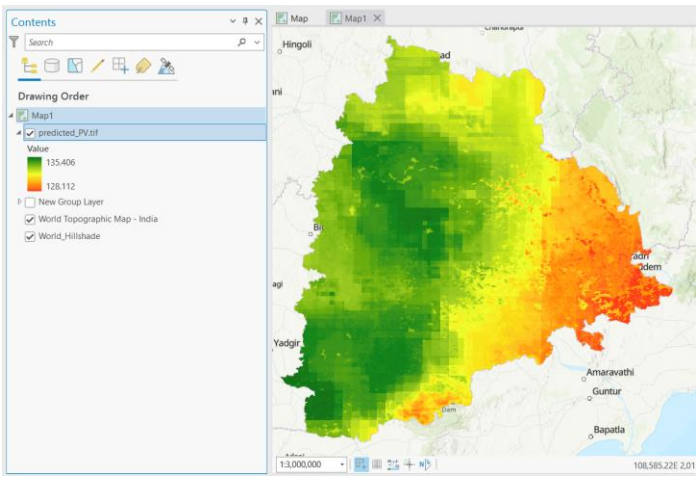
The trained model was saved, and feature importance was generated. The results indicated that GHI contributed the highest feature importance among all input variables. In addition, an XGBoost (XGB) model was also applied to the same dataset, which resulted in a slight improvement in both accuracy and RMSE with the values 0.84 and 0.737 kWh/m² respectively. Using the trained model, the PV values were predicted for the entire Telangana State.

5. Spatial Prediction:

The trained model was applied to the stacked input datasets (X), ensuring equal rows and columns through nearest neighbour resampling to avoid any processing or programming errors. The resulting output was largely consistent with the Photo-voltaic layer, although some key differences were observed.

The machine learning–generated layer contains a certain amount of noise, which is unavoidable due to the coarse spatial resolution of the GHI dataset. This noise is present in both the GIS AHP layer and the machine learning output, as both use the same underlying datasets. However, the noise appears slightly amplified in the Random Forest–generated suitability layer.

The predicted layer indicates that highly suitable areas are predominantly concentrated in the western regions, with suitability gradually decreasing toward the eastern regions. This spatial pattern closely matches the trends observed in the GIS-based suitability layer.

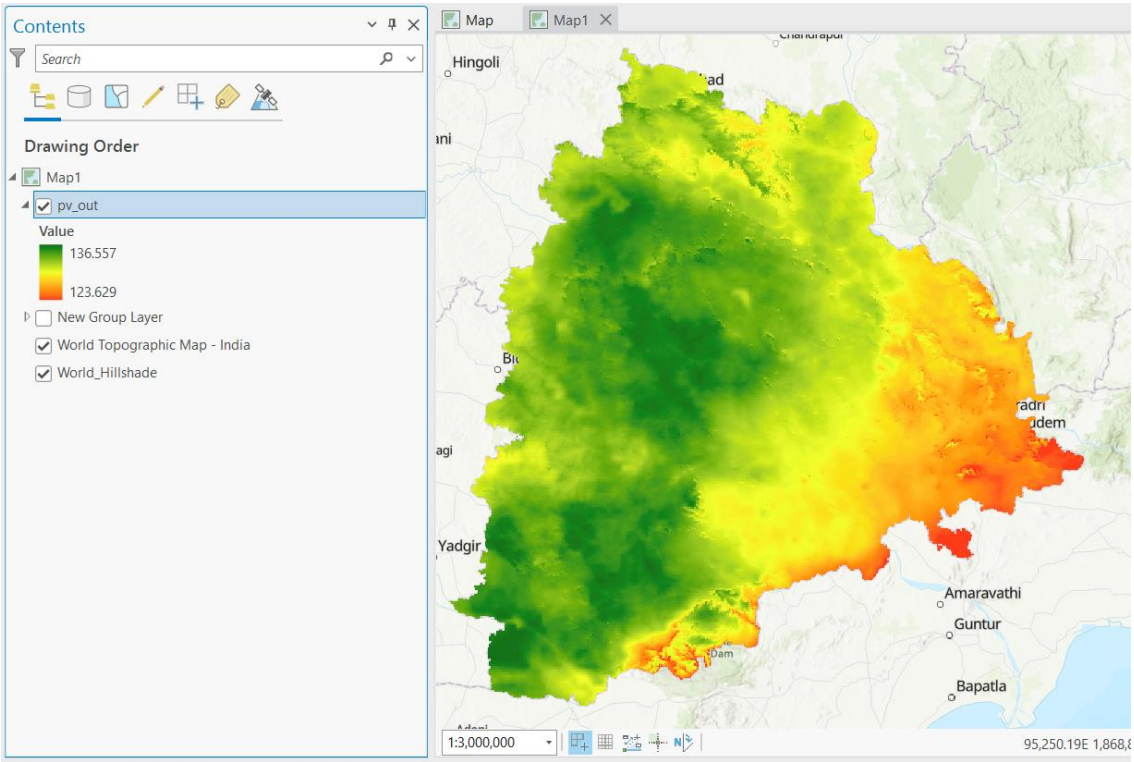


Predicted Solar Suitability (PV) Layer using Random Forest

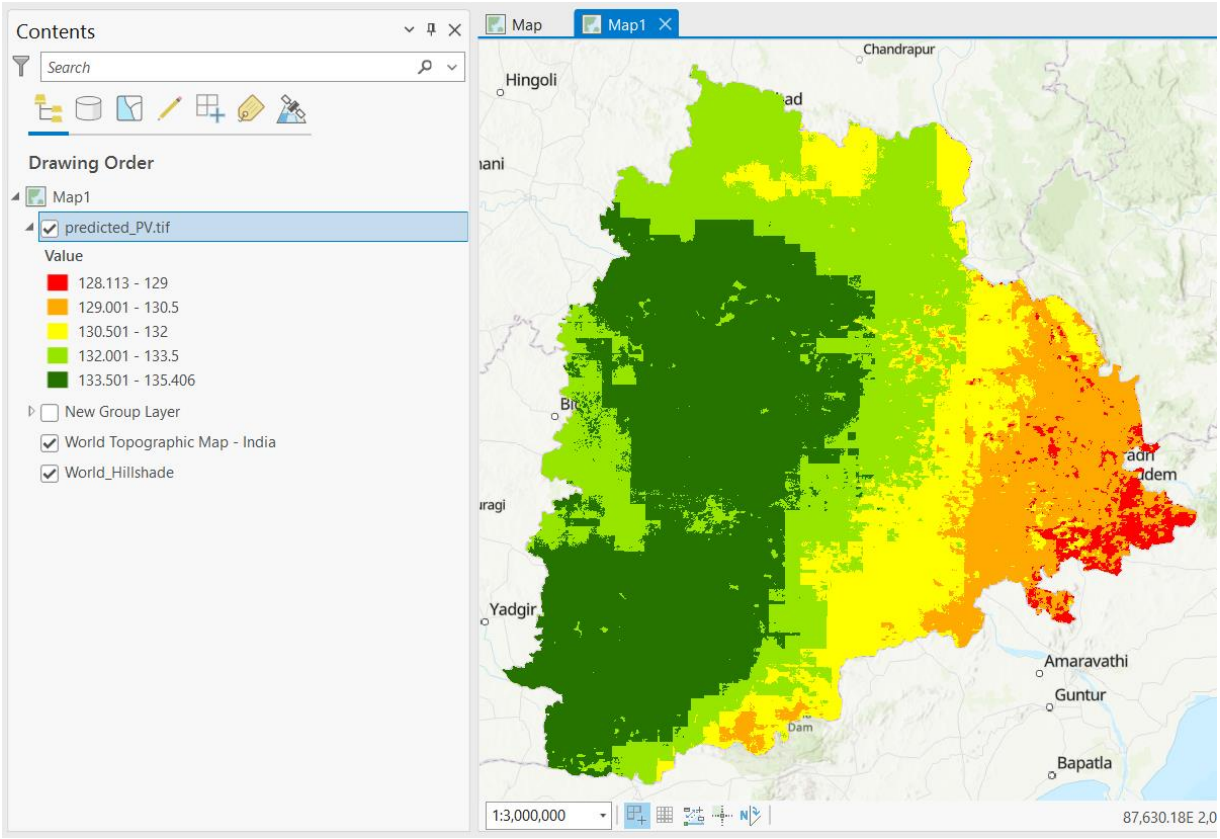
6. Post-Processing:

Finally, the predicted suitability layer was clipped using restriction areas, wildlife sanctuaries, and water bodies. The resulting layer was then classified into five categories—very low, low, medium, and high, very high suitability—highlighting areas that are suitable for setting up solar power plants (indicated by colours). PV suitability was classified into five ordinal classes based on predicted photovoltaic productivity values following thresholds commonly used in solar resource assessment

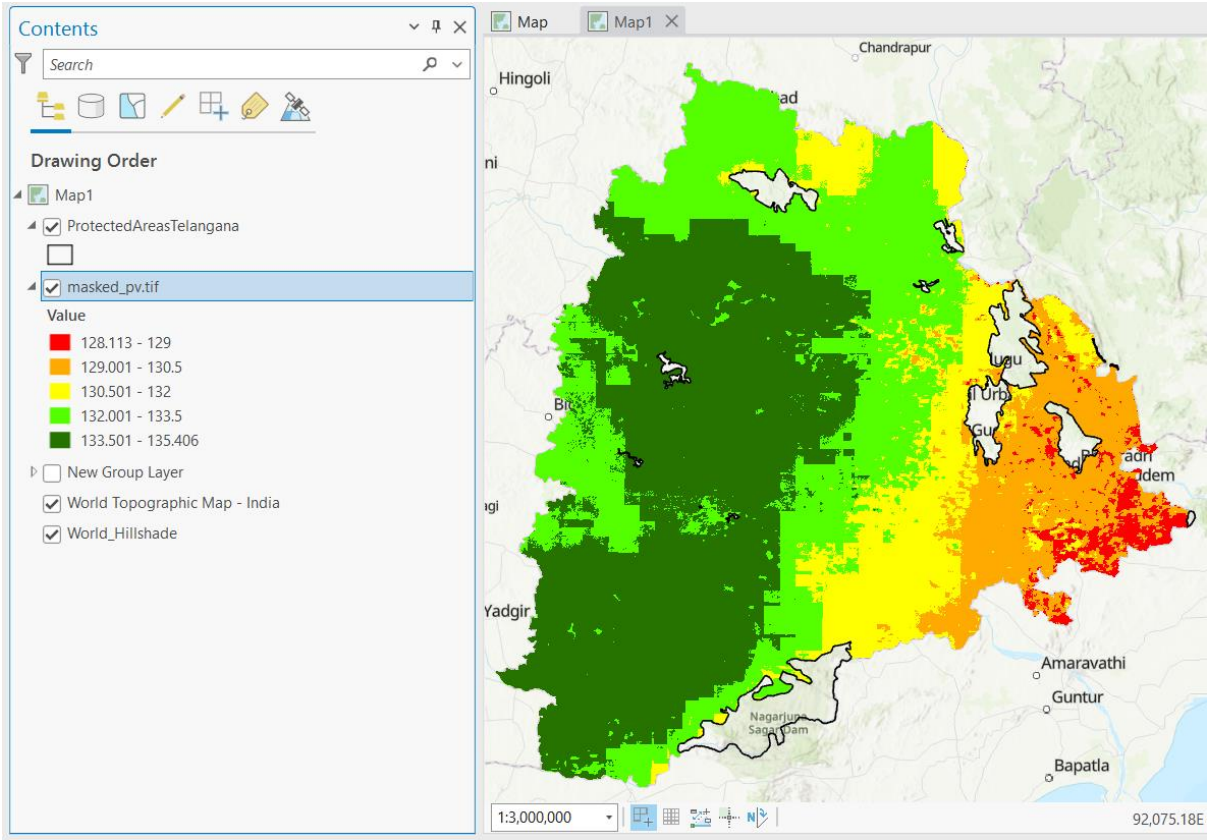
The predicted photovoltaic suitability layer reveals values between 128.1 and 135.4 kWh/m², suggesting that there's a good amount of solar energy available consistently throughout the study area. This range indicates a strong potential for solar resources, where higher numbers point to spots that can produce more energy per unit area, enhancing system efficiency and lowering the levelized cost of electricity. The division into five suitability classes (from Very Low to Very High) emphasizes variations in energy productivity rather than strict limits on feasibility. While the numerical differences might seem small, even a slight variation (1–2 kWh/m²) can have a big impact on long-term energy output and financial returns for large-scale solar projects. Thus, regions in the higher classes are prioritized for solar development, while those in the lower classes might still be technically feasible but are generally less efficient or cost-effective.



GSA PV_out Layer



ML – predicted Suitability classified Layer using RF



Predicted Classified Layer with Masked out Restricted Areas

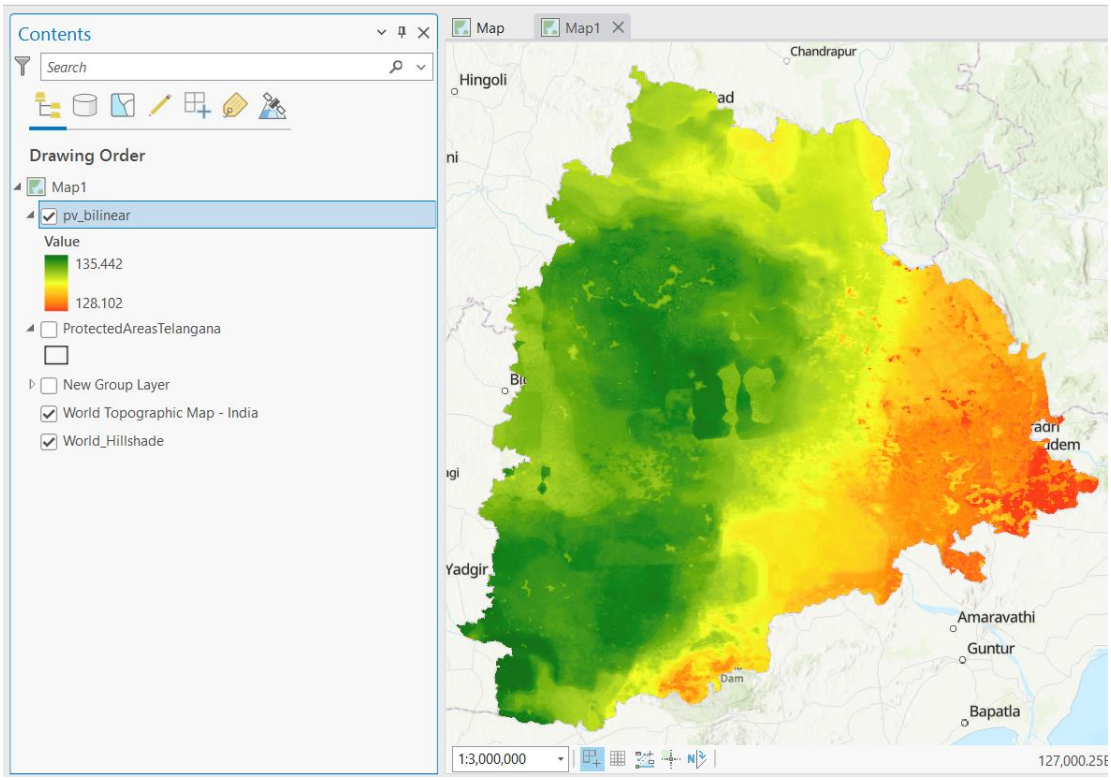
Bilinear Solar Suitability (PV) Map Predicted using Random Forest:

For comparison and to obtain a smoother output surface, the coarser datasets—GHI, temperature, cloud cover, and rainfall—were resampled to 100 m resolution using bilinear interpolation. The remaining layers were kept unchanged, as they represent distance-based proximity maps and were already suitable for resampling at 100 m resolution. This methodology was adopted to generate the bilinear-interpolated photovoltaic suitability zones.

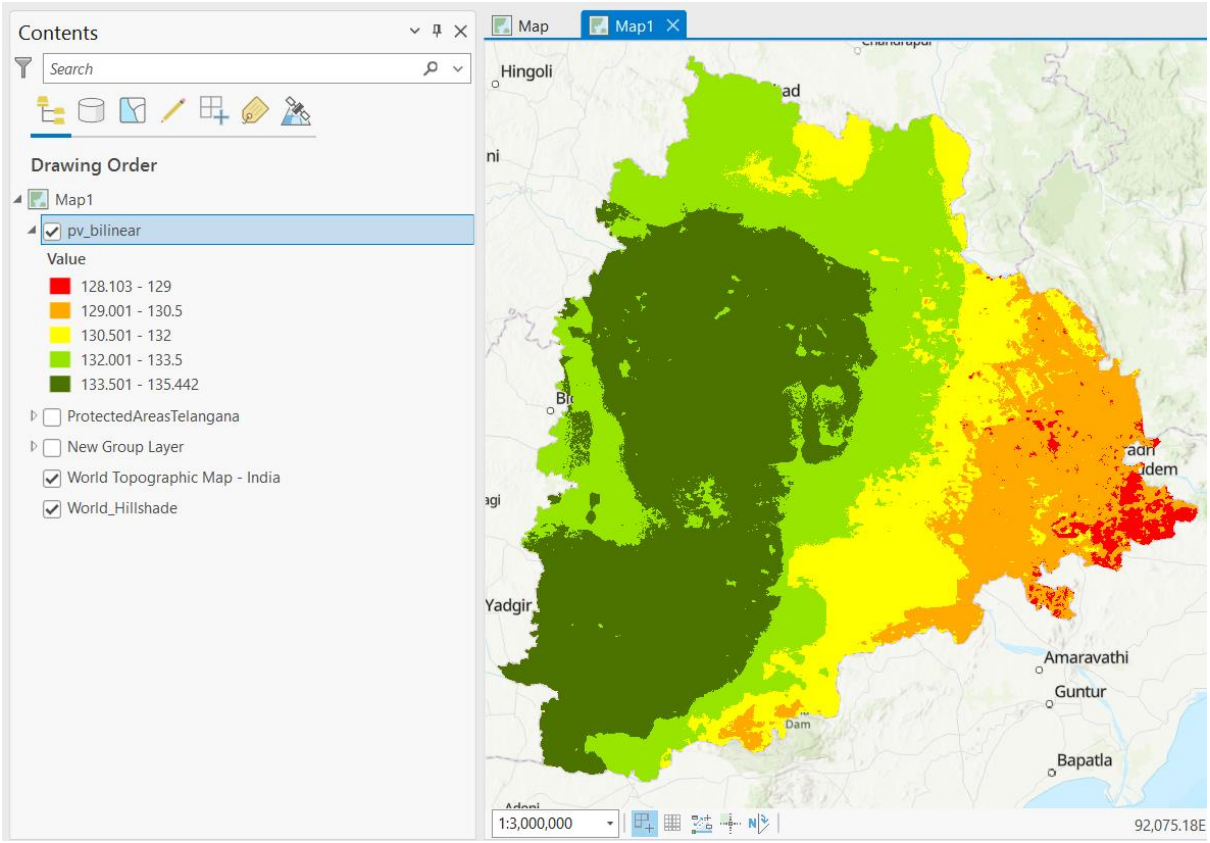
For model training, all ten input layers were used, including the four bilinear-resampled layers along with the PV target variable. The dataset was divided into training and validation sets using a 70:30 split. The model achieved an R^2 value of 0.874 and an RMSE of 0.648, which represent a slight improvement compared to the previous model developed using nearest-neighbour resampled datasets.

The trained model was saved, and feature importance was analyzed. The results showed that GHI had the highest contribution among all input variables. Additionally, an XGBoost (XGB) model was applied to the same dataset, which gave accuracy and RMSE of 0.849 and 0.716 kWh/m² respectively.

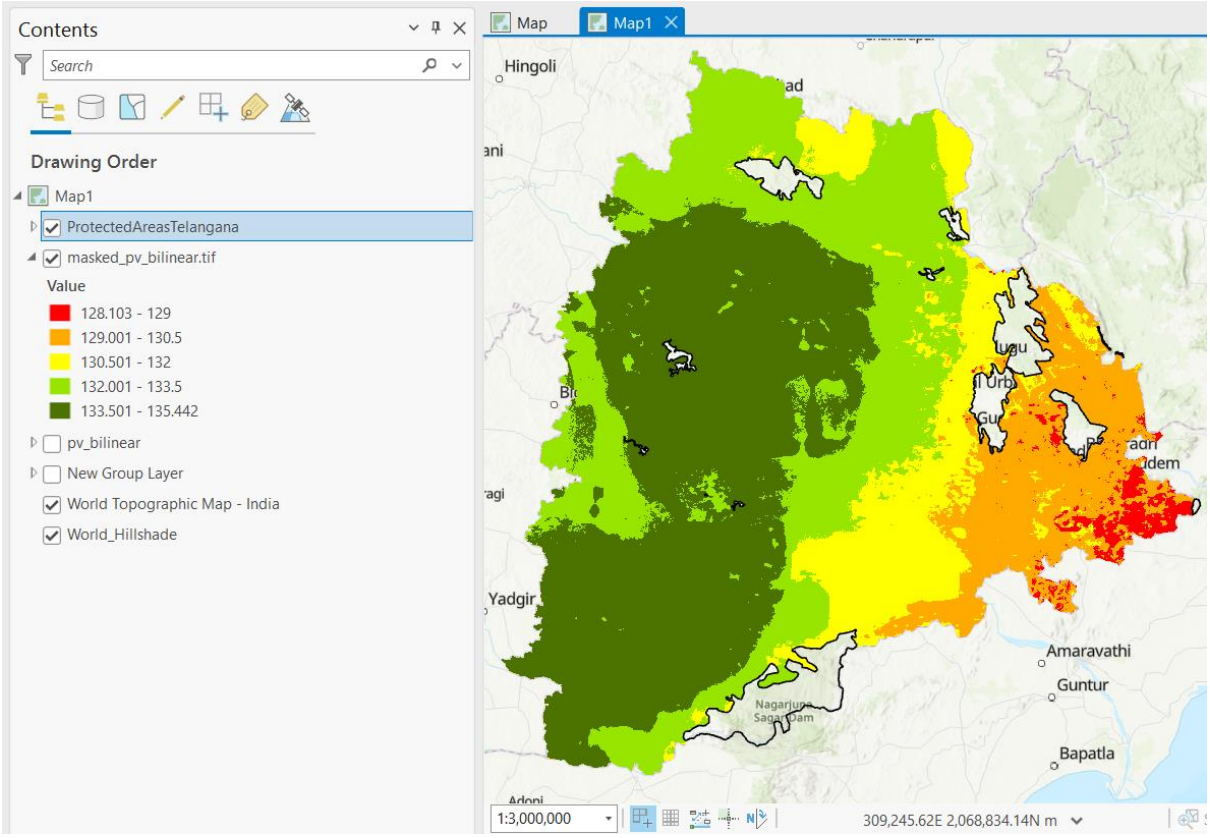
The following Results are obtained using this model.



Predicted Solar Suitability (PV)- Bilinear Layer using Random Forest



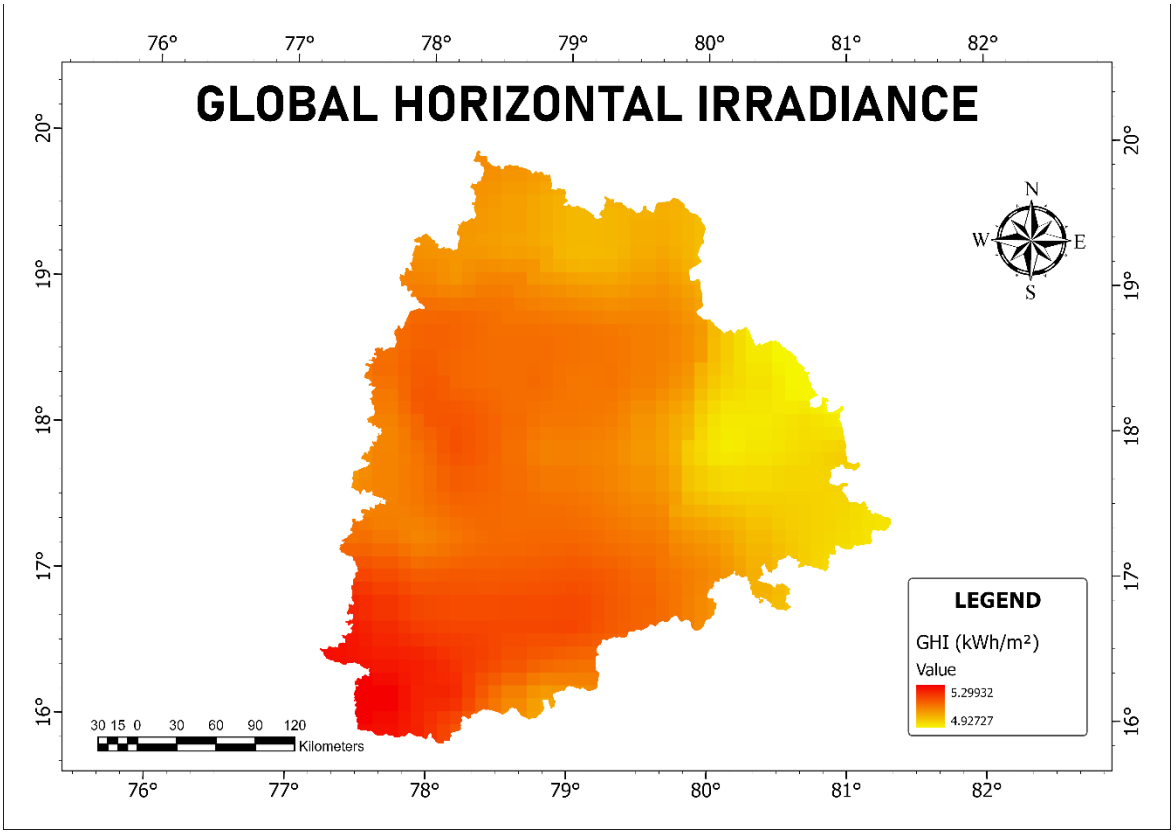
ML – predicted Suitability classified Layer – Bilinear using RF

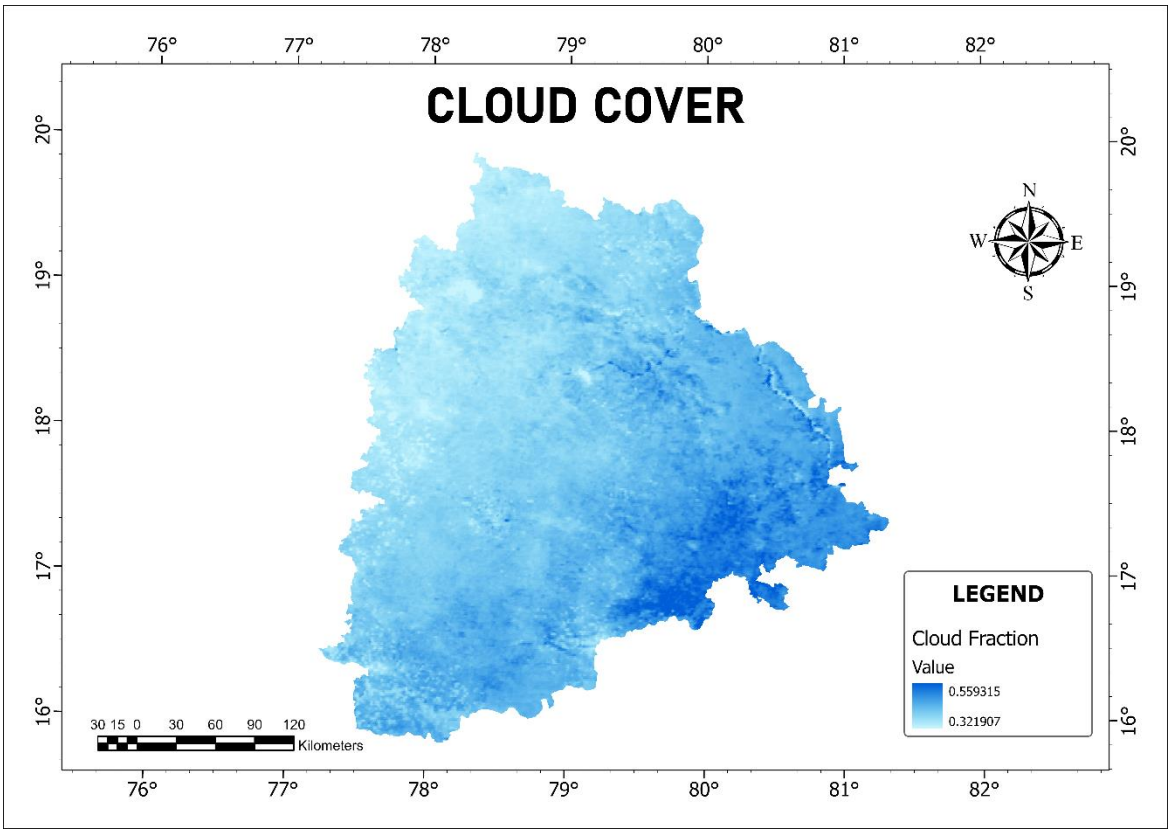


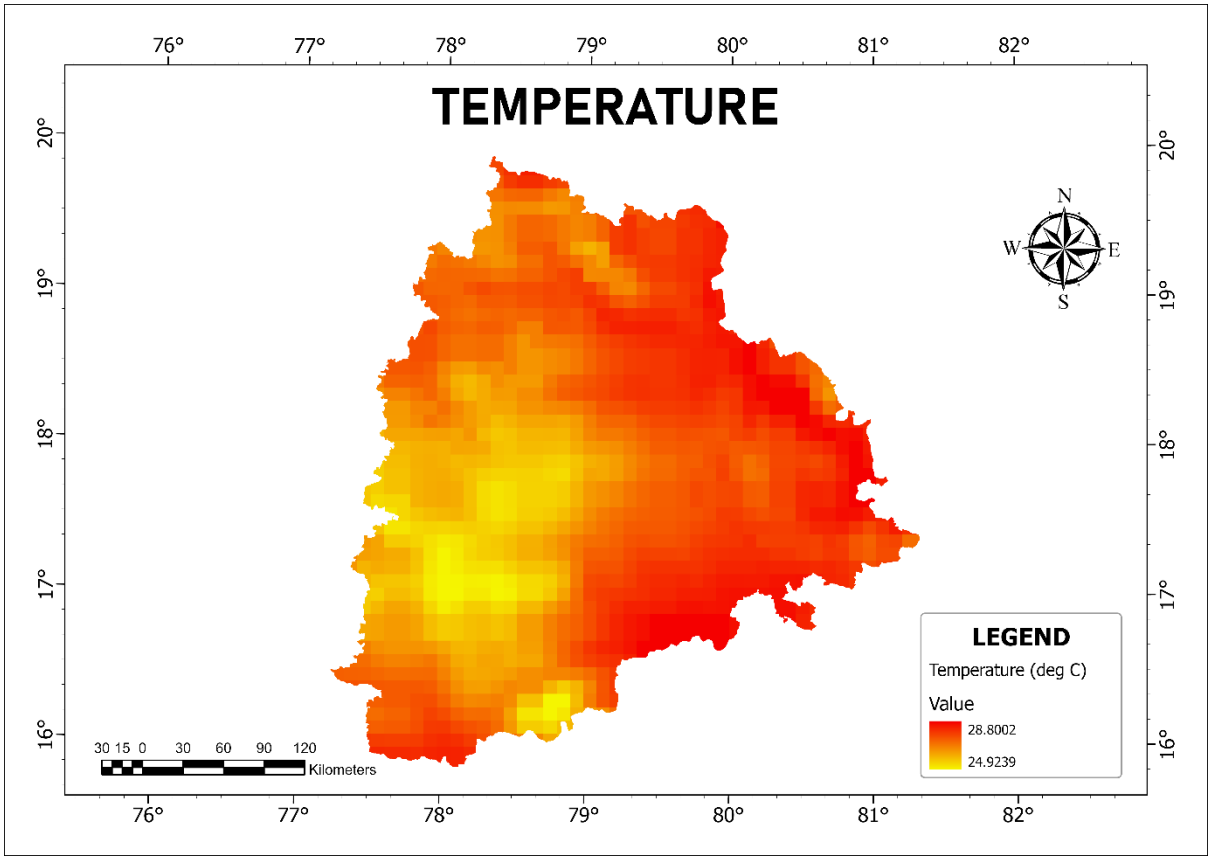
Predicted Classified Layer - Bilinear with Masked out Restricted Areas

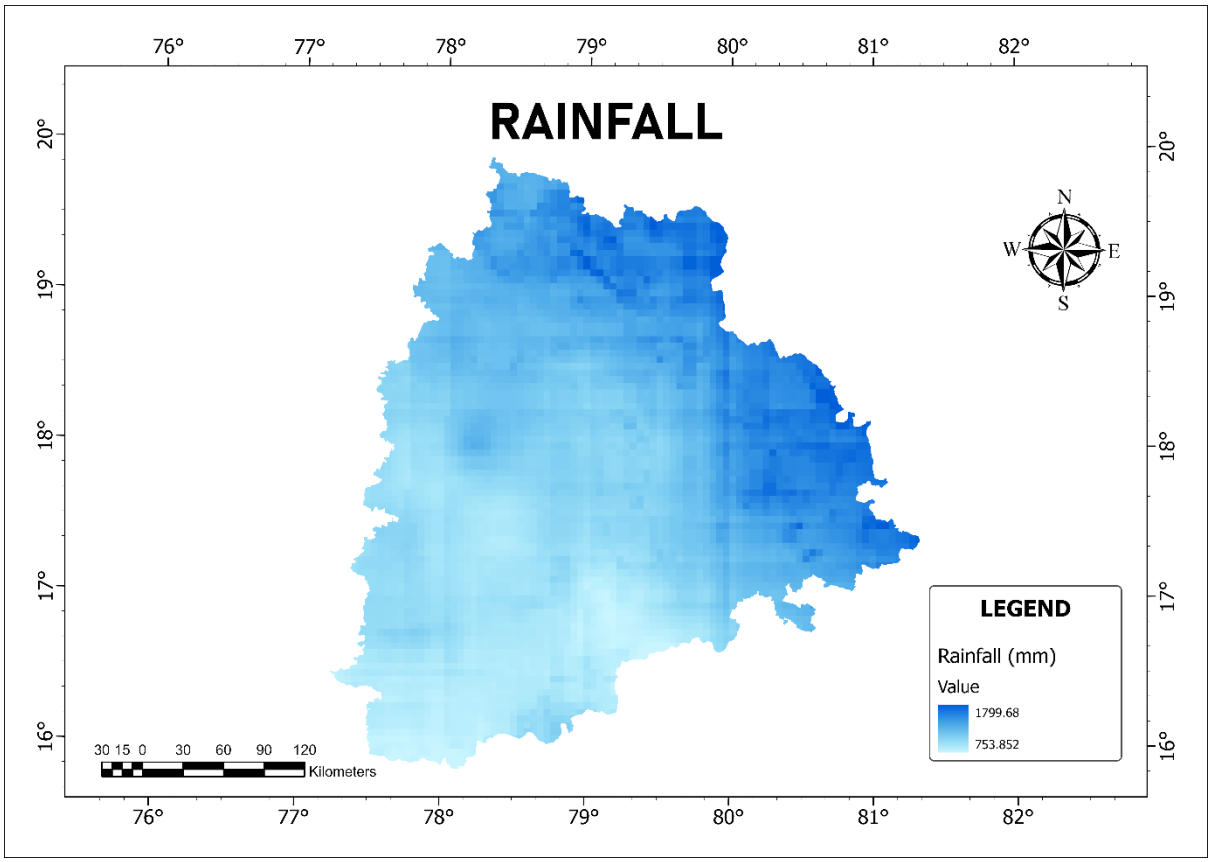
Results:

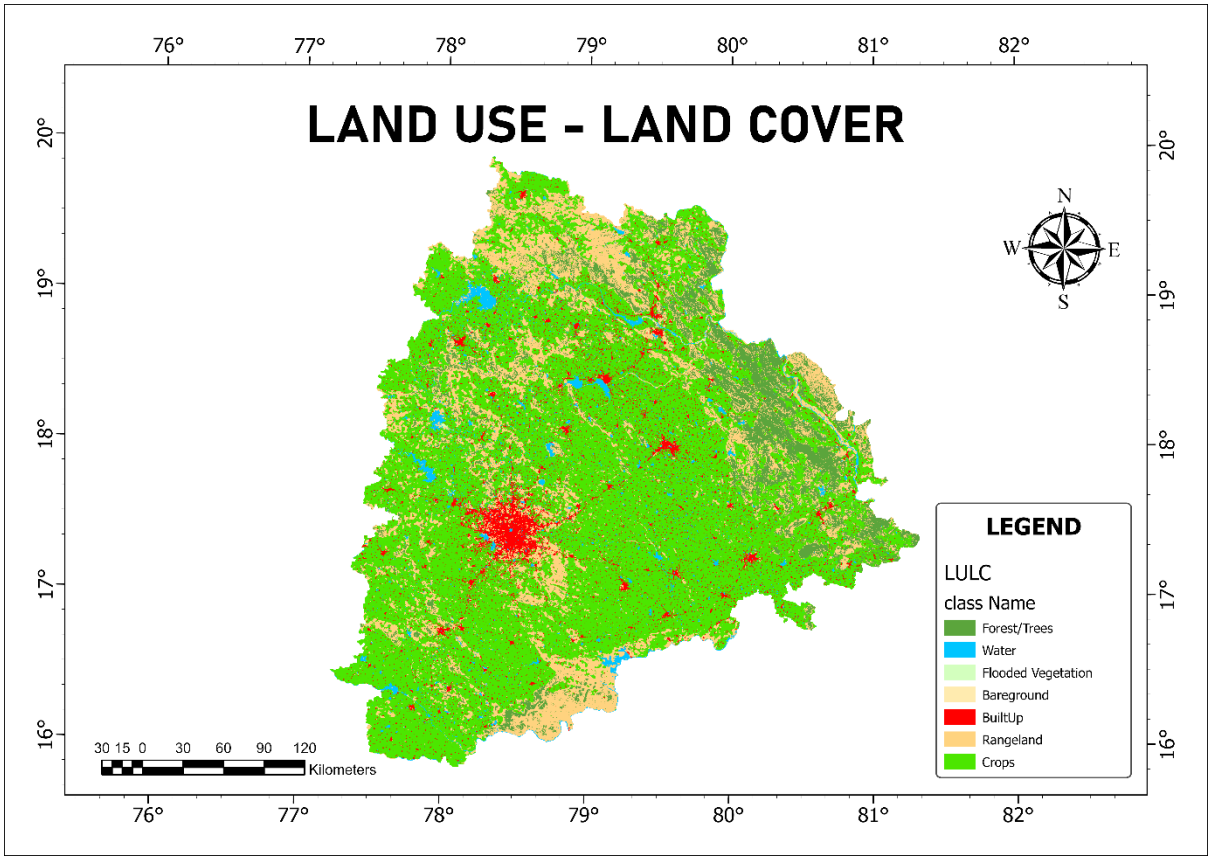
For visualization purposes, layouts were created for the input variables and predicted layers using ArcGIS Pro.

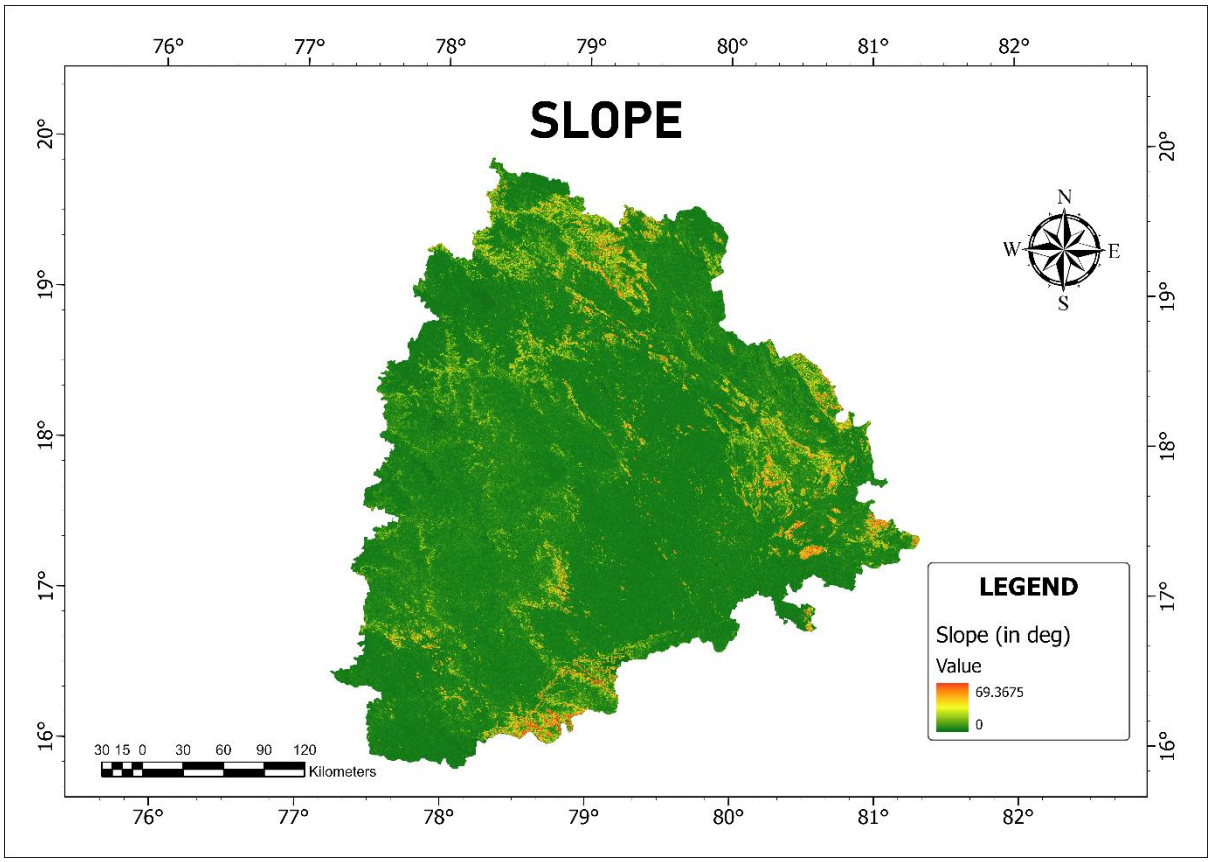


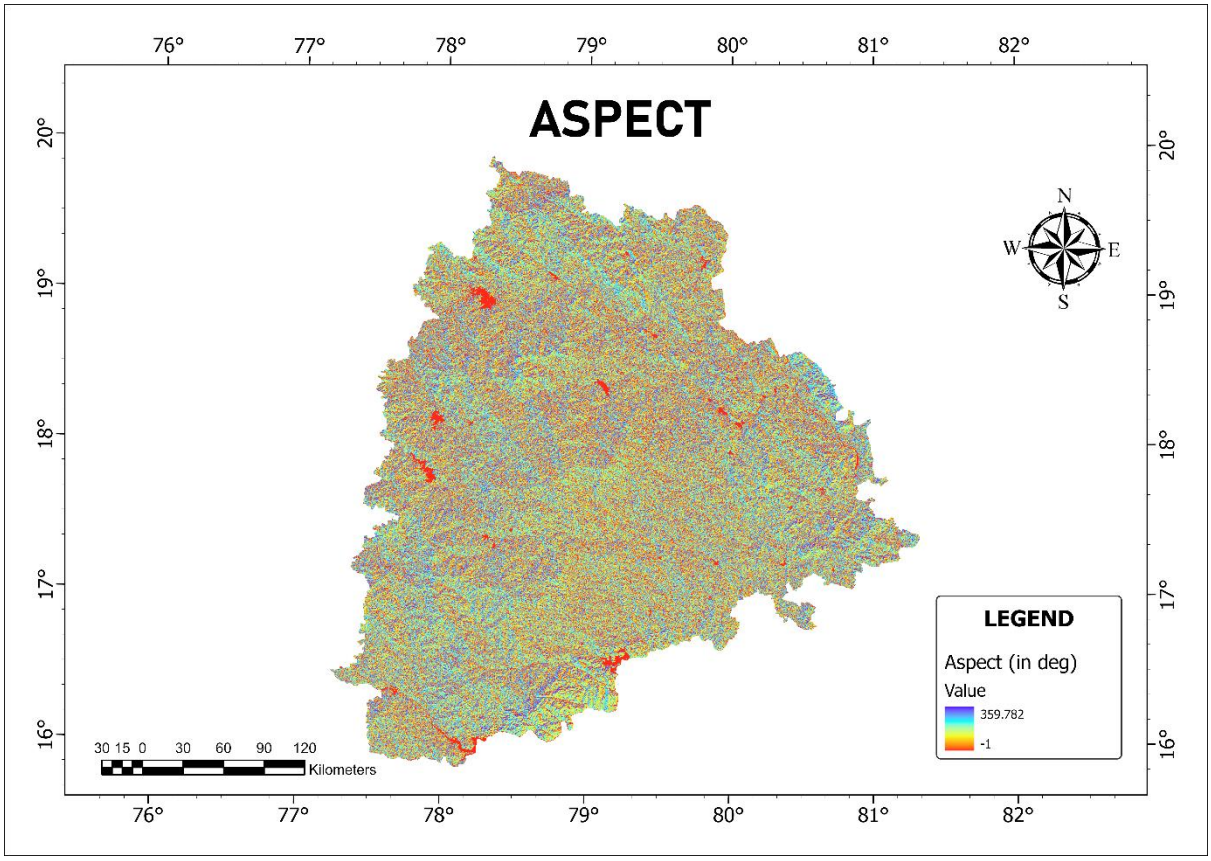


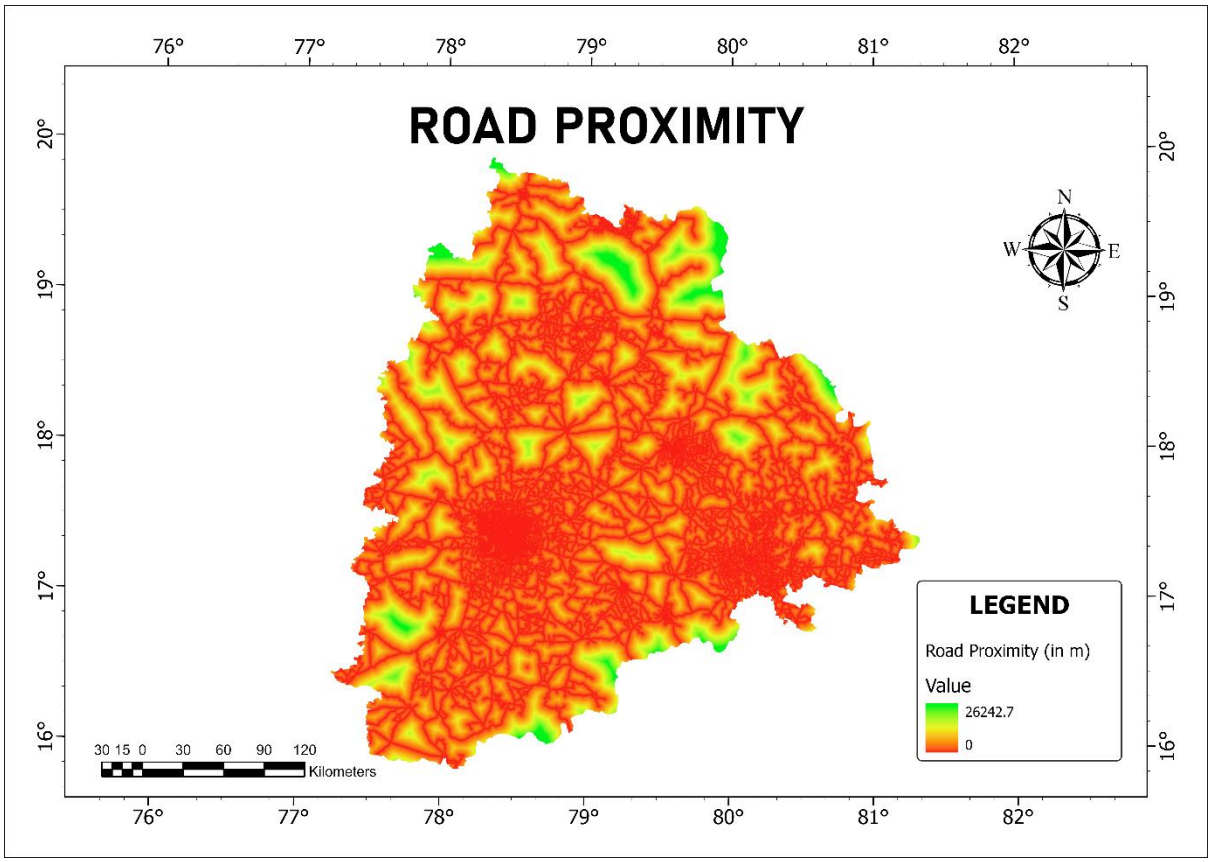

TEMPERATURE
76°
77°
78°
79°
80°
81°
82°
20°
19°
18°
17°
16°
N
W
E
S
LEGEND
Temperature (deg C)
Value
28.8002
24.9239
30 15 0 30 60 90 120
Kilometers

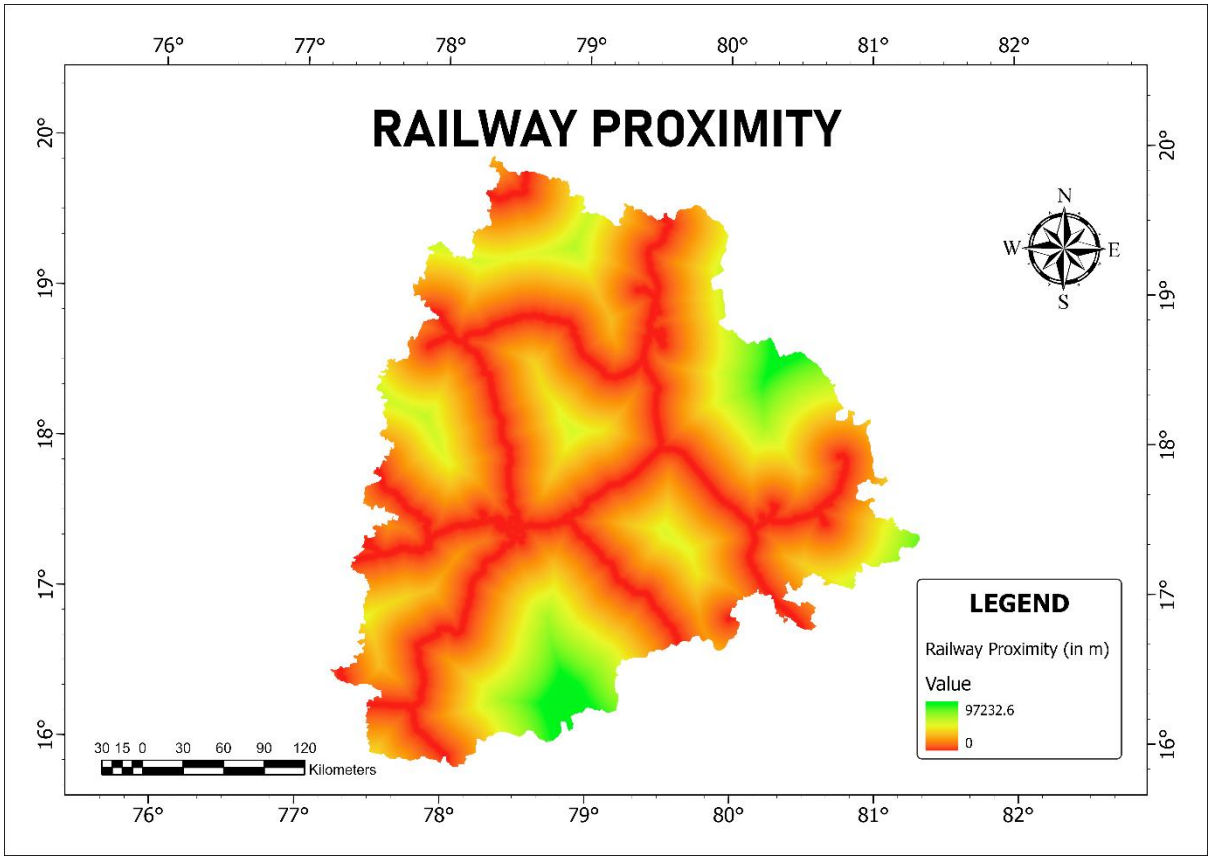

RAINFALL
76°
77°
78°
79°
80°
81°
82°
20°
19°
18°
17°
16°
N
W
E
S
LEGEND
Rainfall (mm)
Value
1799.68
753.852
30 15 0 30 60 90 120
Kilometers

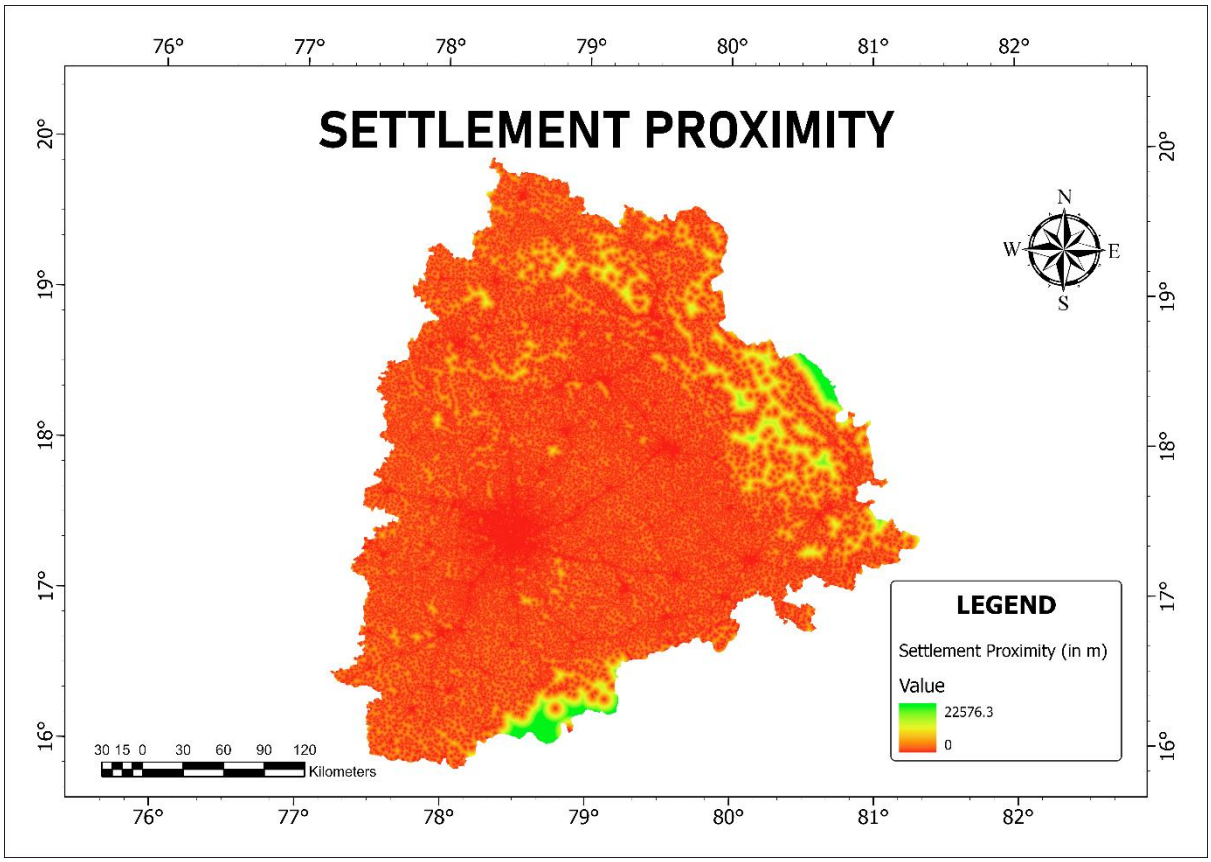

LAND USE - LAND COVER
76°
77°
78°
79°
80°
81°
82°
20°
19°
18°
17°
16°
N
W
E
S
LEGEND
LULC
class Name
Forest/Trees
Water
Flooded Vegetation
Bareground
BuiltUp
Rangeland
Crops
30 15 0 30 60 90 120
Kilometers


SLOPE
76°
77°
78°
79°
80°
81°
82°
20°
19°
18°
17°
16°
N
W
E
S
LEGEND
Slope (in deg)
Value
69.3675
0
30 15 0 30 60 90 120
Kilometers

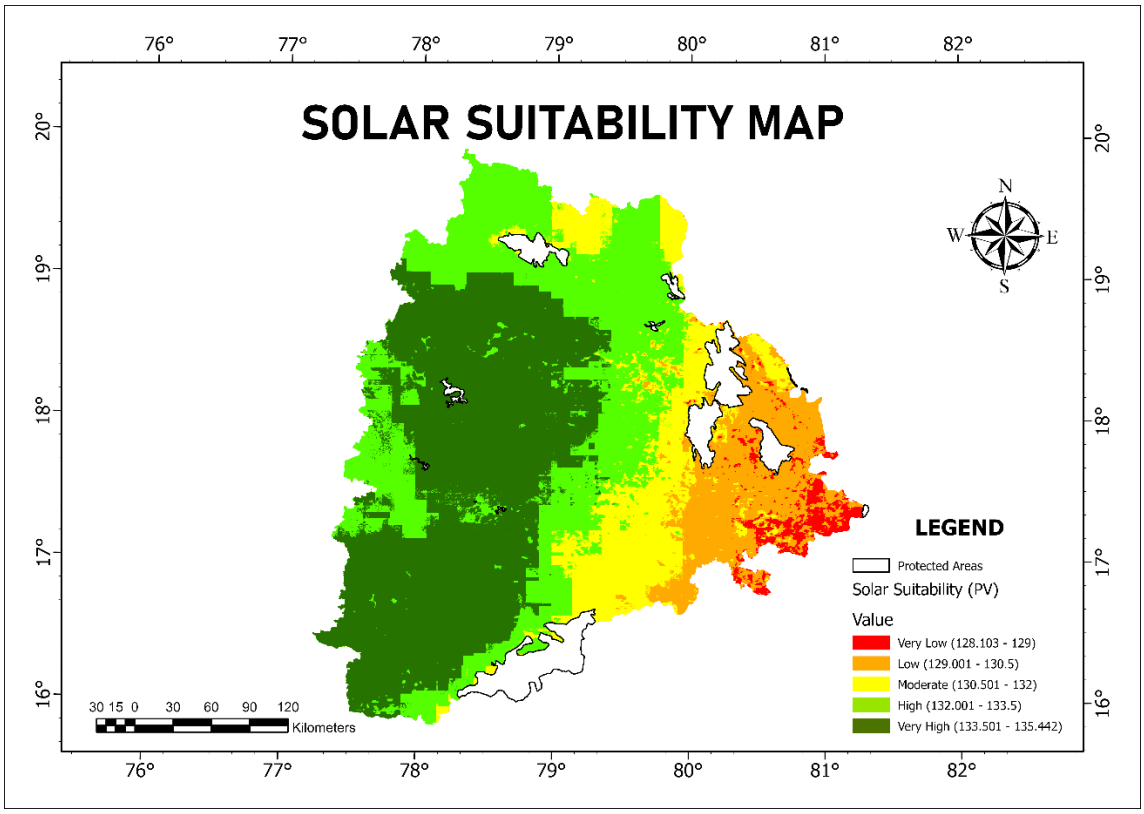

ASPECT
76°
77°
78°
79°
80°
81°
82°
20°
19°
18°
17°
16°
N
W
E
S
LEGEND
Aspect (in deg)
Value
359.782
-1
30 15 0 30 60 90 120
Kilometers


ROAD PROXIMITY
76°
77°
78°
79°
80°
81°
82°
20°
19°
18°
17°
16°
N
W
E
S
LEGEND
Road Proximity (in m)
Value
26242.7
0
30 15 0 30 60 90 120
Kilometers

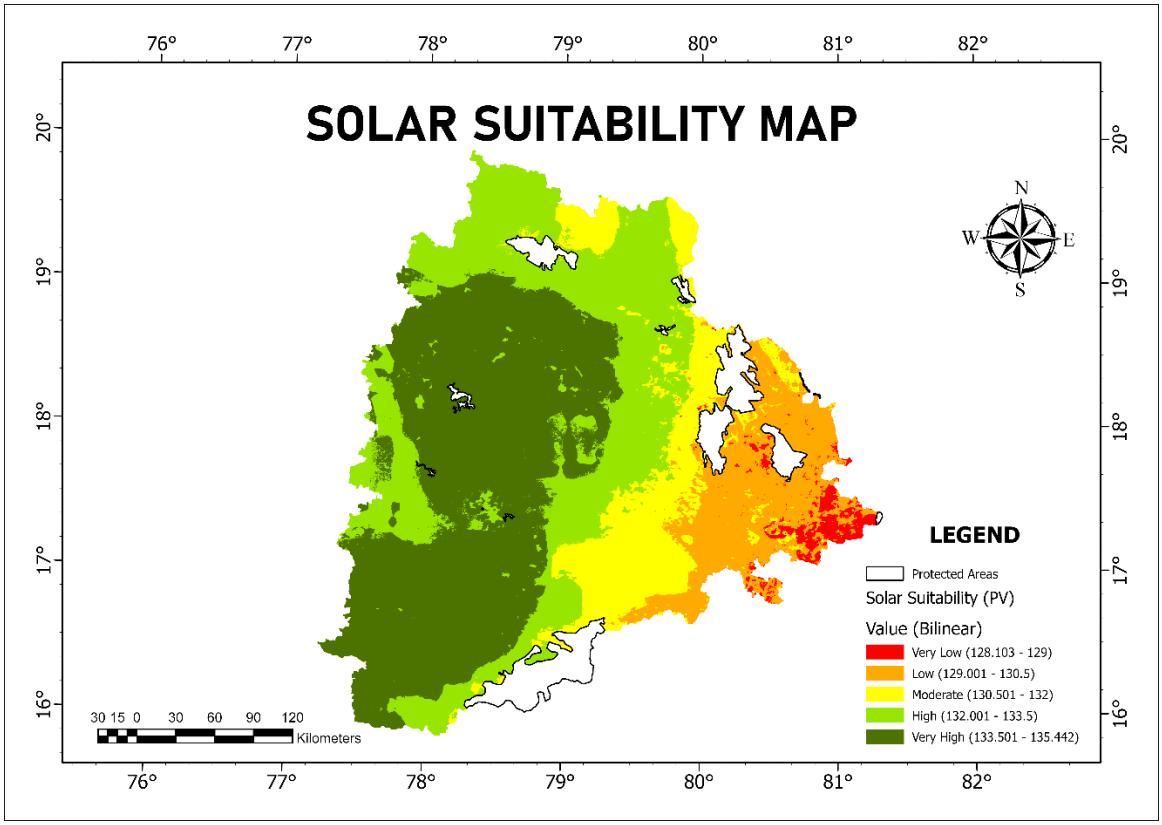

RAILWAY PROXIMITY
76°
77°
78°
79°
80°
81°
82°
20°
19°
18°
17°
16°
N
W
E
S
LEGEND
Railway Proximity (in m)
Value
97232.6
0
30 15 0 30 60 90 120
Kilometers


SETTLEMENT PROXIMITY
76°
77°
78°
79°
80°
81°
82°
20°
19°
18°
17°
16°
N
W
E
S
LEGEND
Settlement Proximity (in m)
Value
22576.3
0
30 15 0 30 60 90 120
Kilometers

Predicted Solar Suitability Layers:



Solar Suitability Map using Nearest Neighbour Resample



Solar Suitability Map using Bilinear Resample

Conclusion:

The model generated a predicted suitability layer using the training data derived from the input datasets and the target variable. The predicted results show strong similarity with the target variable. The western regions of Telangana State were identified as highly suitable, while suitability gradually decreases toward the eastern regions.