

ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL

**A NEW AGRO-METEOROLOGICAL DROUGHT INDEX BASED ON
REMOTE SENSING**



Ph.D. THESIS

Eyyup Ensar BAŞAKIN

Department of Civil Engineering

Hydraulics and Water Resources Engineering Programme

NOVEMBER 2024

ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL

**A NEW AGRO-METEOROLOGICAL DROUGHT INDEX BASED ON
REMOTE SENSING**

Ph.D. THESIS

**Eyyup Ensar BAŞAKIN
(501182505)**

Department of Civil Engineering

Hydraulics and Water Resources Engineering Programme

Thesis Advisor: Prof. Dr. Mehmet Cüneyd DEMİREL

NOVEMBER 2024

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

**UZAKTAN ALGILAMA TEMELLİ YENİ BİR AGRO-METEOROLOJİK
KURAKLIK İNDEKSİ**

DOKTORA TEZİ

**Eyyup Ensar BAŞAKIN
(501182505)**

İnşaat Mühendisliği Anabilim Dalı

Hidrolik ve Su Kaynakları Mühendisliği Programı

Tez Danışmanı: Prof. Dr. Mehmet Cüneyd DEMİREL

KASIM 2024

Eyyup Ensar BAŞAKIN, a Ph.D. student of İTÜ Graduate School student 501182505, successfully defended the thesis/dissertation entitled “A NEW AGRO-METEOROLOGICAL DROUGHT INDEX BASED ON REMOTE SENSING”, which he prepared after fulfilling the requirements specified in the associated legislations, before the jury whose signatures are below.

Thesis Advisor : **Prof. Dr. Mehmet Cüneyd DEMİREL**
Istanbul Technical University

Jury Members : **Assoc. Prof. Dr. Ömer EKMEKCİOĞLU**
Istanbul Technical University

Assoc. Prof. Dr. Kerim KOÇ
Yıldız Technical University

Prof. Dr. Levent ŞAYLAN
Istanbul Technical University

Assoc. Prof. Dr. Eyüp ŞİŞMAN
Yıldız Technical University

Date of Submission : 25 October 2024
Date of Defense : 29 November 2024





To Türkiye,



FOREWORD

I would like to extend my heartfelt gratitude to my advisor, Prof. Dr. Mehmet Cüneyd DEMİREL, for always sharing his knowledge and expertise with me. I also sincerely thank Assoc. Prof. Dr. Ömer EKMEKCİOĞLU, with whom I started my academic journey. His continuous support extended not only to my thesis work but also to all of my academic endeavors and personal life. I owe my deepest thanks to Assoc. Prof. Dr. Kerim KOÇ for his significant recommendations, which greatly contributed to the progress and development of my thesis. I would also like to thank my initial advisor, Prof. Dr. Mehmet ÖZGER, and the jury members.

I would like to extend my heartfelt thanks to Prof. Dr. Levent ŞAYLAN, one of the foremost experts in agricultural drought in our country, who played a significant role in my choice of thesis topic and provided invaluable support throughout. I also thank Prof. Dr. Dursun Zafer ŞEKER for his constant guidance in academic matters.

My gratitude extends to Prof. Dr. Paul C. STOY and his team at UW-Madison, where I conducted a major part of my thesis work with the support of the 2214-A TÜBİTAK scholarship. Their provision of a productive work environment, knowledge sharing, and exceptional assistance over the year I spent there were invaluable. I am also deeply grateful to the Turkish students, academics, and families in Wisconsin/Madison who helped us feel at home and offered companionship during this time.

I extend my sincere thanks to TÜBİTAK for the support provided through the 2214-A - International Research Fellowship Programme for PhD Students (grant number 1059B142201613). I would also like to thank Istanbul Technical University Research Fund (ITU BAP) for their support with the grant number MDK-2021-43075.

My endless gratitude goes to my mother, Sevil BAŞAKIN, and my siblings, Zeynep BAŞAKIN, Betül Gülfem BAŞAKIN, and İsmail Erhan BAŞAKIN, for their unwavering support and the significant roles they played in helping me reach this point. I would also like to thank all of my relatives, friends, and loved ones for their financial and moral support.

Finally, I owe boundless love and gratitude to my beloved wife, Sevim Seda BAŞAKIN, for her extraordinary patience throughout the challenging doctoral and thesis-writing process. I apologize for the time I took away from her and am eternally grateful for her steadfast support and constant presence by my side during this journey.

November 2024

Eyyup Ensar BAŞAKIN
(Civil Engineer)



TABLE OF CONTENTS

	<u>Page</u>
FOREWORD	ix
TABLE OF CONTENTS.....	xi
ABBREVIATIONS	xv
SYMBOLS	xxi
LIST OF TABLES	xxiii
LIST OF FIGURES	xxv
SUMMARY	xxix
ÖZET.....	xxxiii
1. INTRODUCTION.....	1
2. PROVIDING A COMPREHENSIVE UNDERSTANDING OF MISSING DATA IMPUTATION PROCESSES IN EVAPOTRANSPIRATION-RELATED RESEARCH: A SYSTEMATIC LITERATURE REVIEW	5
2.1 Introduction	5
2.2 Mechanisms And Patterns Of Missing Data	7
2.2.1 Missing data mechanisms	7
2.2.2 Missing data patterns	7
2.3 Methods For Measuring And Calculating ET	8
2.3.1 The eddy covariance method	8
2.3.2 Bowen ratio energy balance method	9
2.3.3 Weighing lysimeter method	9
2.3.4 FAO-56 Penman-Monteith method	10
2.4 Literature Review Rationale.....	10
2.5 Results	11
2.5.1 Missing data imputation techniques.....	12
2.5.2 Missing data imputation attempts concentrated on Bowen ratio data	15
2.5.3 Missing data imputation attempts concentrated on EC data	15
2.5.4 Missing data imputation attempts concentrated on lysimeter data	17
2.5.5 Missing data imputation attempts concentrated on humidity data.....	18
2.5.6 Missing data imputation attempts concentrated on solar radiation data ..	19
2.5.7 Missing data imputation attempts concentrated on wind speed data	20
2.5.8 Missing data imputation attempts concentrated on temperature data	21
2.6 Discussion	23
2.6.1 Assessment of the pre-imputation stage.....	24
2.6.2 Assessment of the during-imputation stage	25
2.6.3 Assessment of the post-imputation stage	26
2.7 Conclusions	29
3. DEVELOPING A NOVEL APPROACH FOR MISSING DATA IMPUTATION OF SOLAR RADIATION: A HYBRID DIFFERENTIAL EVOLUTION ALGORITHM BASED EXTREME GRADIENT BOOSTING MODEL.....	31

3.1 Introduction	31
3.2 Study Area And Data.....	37
3.3 Research Framework	40
3.4 Methods	43
3.4.1 Interpolation methods.....	44
3.4.2 eXtreme gradient boosting (XGBoost)	44
3.4.3 Multivariate adaptive regression splines	45
3.4.4 Random forest	45
3.4.5 Differential evolution	46
3.4.6 Performance evaluations	46
3.5 Results	47
3.5.1 Interpolation-based results (scenario 1)	47
3.5.2 Machine learning-based results	49
3.5.3 Scenario 2 results	50
3.5.4 Scenario 3 results	55
3.5.5 Overall results	58
3.6 Discussions, Limitations, And Insights For Future Research	61
3.7 Concluding Remarks	65
4. ESTIMATION OF DAILY REFERENCE EVAPOTRANSPIRATION BY HYBRID SINGULAR SPECTRUM ANALYSIS-BASED STOCHASTIC GRADIENT BOOSTING	67
4.1 Introduction	67
4.2 Study Area And Data.....	69
4.3 FAO-56 Penman-Monteith Reference Evapotranspiration Model.....	70
4.4 Method Details	71
4.4.1 Singular spectrum analysis.....	71
4.4.2 Stochastic gradient boosting	73
4.5 Method Implementation	74
4.6 Methods Validation	76
4.7 Conclusion.....	78
5. GROSS PRIMARY PRODUCTIVITY RESEARCH: ONGOING TRENDS AND FUTURE TRAFECTORIES	81
5.1 Introduction	81
5.2 Background.....	82
5.2.1 Estimating GPP from ecosystem scale measurements.....	82
5.2.2 Estimating GPP from remote sensing	83
5.2.3 Model-data fusion and machine learning.....	84
5.3 Methodology For Literature Review	85
5.4 Database And Search Strategies	87
5.5 Results and Discussion.....	88
5.5.1 Basic information of field research	88
5.5.2 Top countries.....	90
5.5.3 Top journals.....	91
5.5.4 Top institutions.....	92
5.6 Bibliometric Analysis.....	92
5.6.1 World cloud analysis.....	92
5.6.2 Co-occurrence network of keywords	94
5.6.3 Institution network analysis	96
5.6.4 Journal network analysis	96
5.6.5 Thematic map.....	98

5.6.6 Trend topics.....	100
5.7 Conclusion, Limitations And Future Research Opportunities	102
6. SPATITEMPORAL VARIABILITY OF GROSS PRIMARY PRODUCTIVITY IN TÜRKİYE: A MULTI-SOURCE AND MULTI- METHOD ASSESMENT	105
6.1 Introduction	105
6.2 Materials.....	107
6.2.1 Study area.....	107
6.2.2 GPP datasets.....	109
6.2.2.1 MODIS.....	110
6.2.2.2 MuSyQ	110
6.2.2.3 PMLV2.....	111
6.2.2.4 TL-LUE.....	111
6.2.2.5 GOSIF	111
6.2.2.6 FLUXCOM	112
6.2.3 Validation of GPP datasets	112
6.3 Methods	114
6.3.1 Kruskal-Wallis and Mann-Whitney U	114
6.3.2 Modified Mann-Kendal test	114
6.3.3 Innovative trend analysis	114
6.3.4 Empirical mode decomposition	116
6.4 Study Design	119
6.5 Results	120
6.5.1 Statistical evaluation of the GPP datasets	120
6.5.2 Kruskal-Wallis and Mann-Whitney U test results	122
6.5.3 Modified Mann-Kendall results	124
6.5.4 Innovative trend analysis results	127
6.5.5 Empirical mode decomposition results	128
6.6 Discussion	131
6.6.1 Overview.....	131
6.6.2 Spatial analysis of GPP datasets	133
6.6.3 Temporal analysis of GPP datasets.....	134
6.6.4 Differences and similarities between GPP datasets	135
6.6.5 The importance of the findings	136
6.6.6 Limitations of study	137
6.7 Conclusions	137
7. COMBINED DROUGHT INDEX USING HIGH-RESOLUTION HYDROLOGICAL MODELS AND EXPLAINABLE ARTIFICIAL INTELLIGENCE TECHNIQUES IN TÜRKİYE	139
7.1 Introduction	139
7.2 Materials.....	144
7.2.1 Study area.....	144
7.2.2 Data description	145
7.3 Methods	148
7.3.1 Calculations of drought indices.....	148
7.3.1.1 The precipitation condition index	148
7.3.1.2 The temperature condition index	148
7.3.1.3 The vegetation condition index.....	148
7.3.1.4 The soil moisture condition index.....	149
7.3.1.5 The normalized difference water index.....	149

7.3.1.6 The gross primary productivity	149
7.3.2 Mesoscale hydrological model.....	150
7.3.3 Crop yield detrending.....	150
7.3.4 Benchmarking methods.....	151
7.3.4.1 Standardized precipitation evapotranspiration index	151
7.3.4.2 Self-calibrated palmer drought severity index	152
7.3.5 eXtreme gradient boosting	152
7.3.6 Shapley additive explanations.....	153
7.3.7 Principal component analysis.....	153
7.3.8 Model evaluation metrics	154
7.3.9 Drought classification	155
7.4 Model Development	155
7.5 Result And Discussion	158
7.5.1 Machine learning results	158
7.5.2 Weight obtained by SHAP	158
7.5.3 Weight obtained by PCA	158
7.5.4 Crop yield correlation results obtained by three different methods	160
7.5.5 Spatiotemporal distribution of drought indices.....	163
7.5.6 Long term evolution of drought indices.....	168
7.5.7 In-depth analysis of feature importance and dependencies.....	169
7.5.8 Future directions, recommendations and practical applications	171
7.6 Conclusion.....	173
8. CONCLUSIONS AND RECOMMENDATIONS.....	175
REFERENCES.....	179
APPENDICES	223
APPENDIX A: Additional explanations, figures, and tables for Chapter 6.	224
APPENDIX B : Additional figures and tables for Chapter 7.	245
CURRICULUM VITAE	253

ABBREVIATIONS

AHP	: Analytic Hierarchy Process
AIC	: Akaike Information Criterion
ANFIS	: Adaptive Neuro Fuzzy Inference System
ANN	: Artificial Neural Networks
APAR	: Absorbed Photosynthetically Active Radiation
ARIMA	: Autoregressive Integrated Moving Average
BC	: Box-Cox
BCS	: Bayesian Compressive Sensing
BPNN	: Back Propagation Neural Network
BREB	: Bowen Ratio Energy Balance
CDI	: Combined Drought Indices
CFNN	: Counter Propagation Fuzzy-Neural Network
CHIRPS	: Climate Hazards Group InfraRed Precipitation with Station
CI	: Clearness Index
CSD	: Classical Seasonal Decomposition
DE	: Differential Evolution
DF	: Deterministic Formula
DHI	: Diffuse Horizontal Irradiance
DINEOF	: Data Interpolating Empirical Orthogonal Functions
DNI	: Direct Normal Irradiance
DRI	: Drought Reconnaissance Index
DWT	: Discrete Wavelet Transform
EC	: Eddy Covariance
EDI	: Effective Drought Index
EDDI	: Evaporative Demand Drought Index
E_i	: Evaporation of Intercepted Rainfall
ELM	: Extreme Learning Machine
EM	: Expectation-Maximization
EMD	: Empirical Mode Decomposition
E_s	: Soil Evaporation

ET	: Evapotranspiration
ETDI	: Evapotranspiration Deficit Index
ETwater	: Evaporation From Ice, Snow, And Water Bodies
EVI	: Enhanced Vegetation Index
EWMA	: Exponential Weighted Moving Average
GA	: Genetic Algorithm
GAM	: Generalized Additive Model
GBRT	: Gradient Boosting Regression Tree
GHI	: Global Horizontal Irradiance
GIS	: Geographic Information System
GLAD	: Global Land Analysis and Discovery
GLDAS	: Global Land Data Assimilation System
GP	: Genetic Programming
GPP	: Gross Primary Productivity
GPR	: Gaussian process regression
GRNN	: Generalized Regression Neural Network
GRU	: Gated Recurrent Units
HDI	: Hot Deck Imputation
HS	: Hargreaves and Samani
IDW	: Inverse Distance Weighting
ID	: Keywords Plus
IMFs	: Intrinsic Mode Functions
IT	: Interpolation Technics
ITA	: Innovative Trend Analysis
KF	: Kalman filter
KGE	: Kling-Gupta efficiency
kNDVI	: Kernel-NDVI
KNN	: K-Nearest Neighbours
KW	: Kruskal–Wallis
LAI	: Leaf Area Index
LE	: Latent Heat Flux
LIME	: Local Interpretable Model-agnostic Explanations
LOWESS	: Locally Weighted Regression Model
LSSVM	: Least-Squares Support Vector Machine
LST	: Land surface temperature

LSTM	: Long Short-Term Memory
LR	: Linear Regression
LUE	: Light Use Efficiency
MA	: Moving Average
MAE	: Mean Absolute Error
MASE	: Mean Absolute Scaled Error
MAR	: Missing at Random
MARS	: Multivariate Adaptive Regression Splines
MB	: Mean Bias
MBE	: Mean Bias Error
MCAR	: Missing Completely at Random
MCP	: Multi Country Publication
MERRA-2	: Modern-Era Retrospective analysis for Research and Applications
MhM	: Mesoscale Hydrological Model
MI	: Mean Imputation
MICE	: Multiple Imputations by Chained Equations
ML	: Machine Learning
MLP	: Multi-Layer Perceptron
MLR	: Multiple Linear Regression
MMK	: Modified Mann–Kendall
MNAR	: Missing Not at Random
MODIS	: Moderate Resolution Imaging Spectroradiometer
MPI	: Multiple Imputation
MuSyQ	: Multisource Data Synergized Quantitative
MVMD	: Multivariate Variational Mode Decomposition
MWU	: Mann–Whitney U
NA	: Not Available
NDVI	: Normalized Difference Vegetation Index
NDWI	: Normalized Difference Water Index
NEE	: Net Ecosystem Exchange
NIRv	: Near-Infrared Reflectance of Vegetation
NP	: Number of Publications
NSE	: Nash-Sutcliffe efficiency
OCO-2	: Orbiting Carbon Observatory-2
P	: Photosynthetically Active Radiation

PACF	: Partial Autocorrelation Function
P-BSHADE	: Biased Sentinel Hospitals-based Area Disease Estimation
PAR	: Photosynthetically Active Radiation
PCA	: Principal Component Analysis
PDSI	: Palmer Drought Severity Index
PM	: Penman–Monteith
PML	: Penman–Monteith–Leuning
PMLV2	: Penman-Monteith-Leuning Evapotranspiration V2
PPCCBC	: Positivity Preserving Condition for Cubic Bézier Curve
PPFD	: Photosynthetic Photon Flux Densities
PY start	: Year of First Publication
R	: Spearman’s Correlation Coefficient
RBF	: Radial Basis Function
RECO	: Ecosystem Respiration
RF	: Random Forest
RH	: Relative Humidity
RMSE	: Root Mean Square Error
RNN	: Recurrent Neural Network
RS	: Random Sample
RSWin	: Short Wave Incoming Solar Radiation
SASR	: Statistically Adjusted Solar Radiation
scPDSI	: Self-Calibrated Palmer Drought Severity Index
SCPs	: Single Country Publications
SDI	: Streamflow Drought Index
SGB	: Stochastic Gradient Boosting
SHAP	: SHapley Additive exPlanations
SI	: Spline Interpolation
SIF	: Solar-Induced Chlorophyll Fluorescence
SM	: Soil Moisture
SMA	: Soil Moisture Anomaly
SMCI	: Soil Moisture Condition Index
SPEI	: Standardized Precipitation-Evapotranspiration Index
SPI	: Standardized Precipitation Index
SS	: Seasonal Splitting
SSA	: Singular Spectrum Analysis

STI	: Stine Interpolation
SRT	: Spatial Regression Testing
SVD	: Singular Value Decomposition
SVM	: Support Vector Machine
SVR	: Support Vector Regression
TBA	: Temperature-Based Approach
TC	: Total Number of Citations
TCI	: Temperature Condition Index
TDRNN	: Time Delay Recurrent Neural Network
TLFN	: Time-Lagged Feedforward Network
TL-LUE	: The Two-Leaf Light Use Efficiency
TUIK	: Turkish Statistical Institute
VCi	: Vegetation Condition Index
VCR	: Variance Contribution Rate
VHI	: Vegetation Health Index
VI_s	: Vegetation Indices
VMD	: Variational Mode Decomposition
VPD	: Vapor Pressure Deficit
WANN	: Wavelet Artificial Neural Network
WMA	: Weighted Moving Average
WS	: Wind Speed
WoS	: Web of Science
XGBoost	: eXtreme Gradient Boosting



SYMBOLS

C	: Carbon
C_a	: Atmospheric Carbon Dioxide Concentration
°C	: Degree Celsius
C_p	: Capillary Rise
CO₂	: Carbon Dioxide
cosθ	: Solar Zenith Angle
d	: Day
D_p	: Deep Infiltration
ET_a	: Actual Evapotranspiration
ET_o	: Reference Evapotranspiration
e_a	: Actual Vapor Pressure
e_s	: Saturation Vapor Pressure Deficit
g	: Gram
G	: Soil Heat Flux
G_c	: Biophysical Canopy Conductance
G_s	: Biophysical Model for Surface Conductance
H	: Sensible Heat Flux
m	: Meter
N	: Sunshine Duration
p	: Probability
P	: Precipitation
R	: Pearson correlation coefficients
R²	: Determination Coefficient
R_{fo}	: Runoff
R_n	: Net Radiation
S₀	: Solar Constant (1367 W/m ²)
T	: Temperature
t	: Time
u	: Wind Speed
α	: Angle of deviation from the direction of the principal stresses

β	: Bowen Ratio/ Initial Photochemical Efficiency
η	: Initial Value of The Slope of Carbon Dioxide Response Curve
γ	: Psychrometric Constant
φ	: Latitude
δ	: Solar Declination
Ω	: Grouping of Leaves
ε_{max}	: Maximum Light Use Efficiency
ρ_a	: Air Density
Δ	: Slope of Vapor Pressure Curve
ΔW	: Changes in Soil Water Content



LIST OF TABLES

	<u>Page</u>
Table 2.1 : Commonly applied missing data imputation methods.....	13
Table 2.2 : Details regarding the missing value imputation studies utilizing the Bowen-Ratio technique	15
Table 2.3 : Details regarding the missing value imputation studies utilizing the Eddy-covariance technique.....	16
Table 2.4 : Details regarding the studies devoted to impute missing humidity data.	18
Table 2.5 : Details regarding the studies devoted to impute missing solar radiation data.....	19
Table 2.6 : Details regarding the studies devoted to impute missing wind speed data	21
Table 2.7 : Details regarding the studies devoted to impute missing temperature data	22
Table 3.1 : The descriptive statistics of the dataset acquired from Aksaray station .	39
Table 3.2 : The descriptive statistics of the solar radiation data acquired from neighboring stations.....	40
Table 3.3 : Model combinations constituted within Scenario 2.	41
Table 3.4 : Model combinations constituted within Scenario 3.....	43
Table 3.5 : Details of the performance indicators considered in the models' evaluation.....	47
Table 3.6 : Test performances of Interpolation-Based models.	48
Table 3.7 : Test performances of the models within Scenario 2.	53
Table 3.8 : Test performances of the models within Scenario 3.	58
Table 3.9 : Benchmarking the proposed comprehensive framework with the existing attempts.....	63
Table 4.1 : Descriptive statistics of meteorological variables.....	70
Table 4.2 : Model performance results.....	77
Table 5.1 : Results of a Scopus database search on GPP research.	89
Table 5.2 : Publication numbers of the top 10 corresponding author's countries.	90
Table 5.3 : The top 10 journals by source impact for the search criteria related to GPP research.....	91
Table 5.4 : The top 15 productive institutions.	92
Table 5.5 : The top 20 most frequently repeated words following the VOSviewer Co-Occurrence Network analysis.	94
Table 6.1 : Temporal and spatial information of GPP dataset.	109
Table 6.2 : Monthly MMK trend analysis results.	126
Table 6.3 : Modified Mann–Kendall trend analysis results for GPP estimates from different remote sensing data products from selected areas in different regions of Türkiye.	128
Table 7.1 : Descriptive information of the data.	147
Table 7.2 : Classification Thresholds for CDI, SPEI, and scPDSI.	155
Table 7.3 : Correlation Results for Different CDI Weight Calculation Methods. ..	160

Table A.1 : Mann-Kendall and Sens' Slope trend analysis results of Antalya region.....	232
Table A.2 : Mann-Kendall and Sens' Slope trend analysis results of Erzincan region.....	232
Table A.3 : Mann-Kendall and Sens' Slope trend analysis results of Izmir region.....	232
Table A.4 : Mann-Kendall and Sens' Slope trend analysis results of Kirklareli region.....	233
Table A.5 : Mann-Kendall and Sens' Slope trend analysis results of Konya region.....	233
Table A.6 : Mann-Kendall and Sens' Slope trend analysis results of Samsun region.....	233
Table A.7 : Mann-Kendall and Sens' Slope trend analysis results of Sanliurfa region.....	234
Table A.8 : Variance contribution rate of IMFs.....	235
Table B.1 : The mean [minimum: maximum] winter wheat yield (t/ha) per district in the study area from 2001 to 2020.....	245
Table B.2 : XGBoost hyperparameter range and optimum values.....	245
Table B.3 : XGBoost model performance evaluation results.....	246
Table B.4 : CDI weights obtained from PCA of the training set.....	247
Table B.5 : Cumulative variance explanation values of the principal components.....	248
Table B.6 : Correlation test results between drought indices and crop yield.....	249

LIST OF FIGURES

	<u>Page</u>
Figure 2.1 : Mechanisms and patterns of missing data. MAR = Missing at Random; MCAR = Missing Completely at Random; MNAR = Missing Not at Random.....	6
Figure 2.2 : The steps of the literature review conducted via Scopus.	11
Figure 2.3 : Keyword networks generated by VOSviewer based on the Scopus database.	12
Figure 3.1 : Study area.	39
Figure 3.2 : Pearson correlation matrix of the meteorological parameters and solar radiation measurements.	41
Figure 3.3 : Pearson correlation matrix of the neighboring stations' solar radiation measurements and Aksaray station's.....	42
Figure 3.4 : Flowchart of the current research.	43
Figure 3.5 : Scatters plots of interpolation-based missing data imputations for the testing set a. Model 1, b. Model 2.	48
Figure 3.6 : Scatters plots of missing data imputations for the testing set with Model 3.	50
Figure 3.7 : Scatters plots of missing data imputations for the testing set with Model 10.	52
Figure 3.8 : Scatters plots of missing data imputations for the testing set with Model 11.	56
Figure 3.9 : Scatters plots of missing data imputations for the testing set with Model 14.	57
Figure 3.10 : Holistic evaluation of adopted models at 5% gap size with Taylor diagram.	60
Figure 3.11 : Holistic evaluation of adopted models at 10% gap size with Taylor diagram.	60
Figure 3.12 : Holistic evaluation of adopted models at 20% gap size with Taylor diagram.	61
Figure 3.13 : Holistic evaluation of adopted models at 30% gap size with Taylor diagram.	61
Figure 4.1 : Study area.	69
Figure 4.2 : FAO-56-PM ETo time series starting from January 1, 2009.	70
Figure 4.3 : Flowchart of the current research.	74
Figure 4.4 : Components of FAO-56-PM time series that decomposition by SSA..	75
Figure 4.5 : Partial autocorrelation function results.....	76
Figure 4.6 : Models scatter plots.	78
Figure 5.1 : Schematic view of Gross Primary Productivity (GPP) and its measurements methods.	84
Figure 5.2 : Annual publications in GPP research listed in Scopus from 2000 to 2023.	89
Figure 5.3 : The geographical distribution of papers based on all authors' affiliations.	91

Figure 5.4 : The word cloud representing the 50 most frequently repeated words (A), the frequencies of the 20 most recurrent words, and their representative areas within the search space (B).....	93
Figure 5.5 : The network visualization analysis of keywords.....	95
Figure 5.6 : The overlay visualization analysis of keywords.....	95
Figure 5.7 : International collaboration network.....	96
Figure 5.8 : The visualization map of journals citation network.	97
Figure 5.9 : The visualization map of journals citation overlay.	98
Figure 5.10 : Thematic map of author’s keywords.	100
Figure 5.11 : Trend topics.	101
Figure 6.1 : Maps of elevation (a), annual average precipitation (1980–2020) (b), annual average temperature (1980–2020) (c), and land use/land cover (2021) (d) of Türkiye.	108
Figure 6.2 : Gross Primary Productivity across Türkiye as estimated by the TL-LUE dataset (Bi et al, 2022) for June 2020.	110
Figure 6.3 : ITA example.	116
Figure 6.4 : Flowchart of the study.	120
Figure 6.5 : Correlation matrices of GPP datasets from the (a) Antalya, (b) Erzincan, (c) Izmir, (d) Kirklareli, (e) Konya, (f) Samsun, and (g) Sanliurfa regions.....	121
Figure 6.6 : GPP dataset similarity test results for the (a) Antalya, (b) Erzincan, (c) Izmir, (d) Kirklareli, (e) Konya, (f) Samsun, and (g) Sanliurfa regions of Türkiye (**** $p < 0.0001$, *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$).	123
Figure 6.7 : ITA results for (a) Antalya, (b) Erzincan, (c) Izmir, (d) Kirklareli, (e) Konya, (f) Samsun, and (g) Sanliurfa.	129
Figure 6.8 : Example of EMD subseries for MuSyQ GPP estimates from Konya, Türkiye.....	130
Figure 6.9 : Trends in GPP extracted from the last mode of an Empirical Mode Decomposition (EMD) on GPP estimates from different remote sensing data products from different regions of Türkiye, (a) Antalya, (b) Erzincan, (c) Izmir, (d) Kirklareli, (e) Konya, (f) Samsun and (g) Sanliurfa.....	131
Figure 7.1 : (a) The study area in Türkiye, (b) districts in the study area, (c) a digital elevation map, and (d) cropland areas.	145
Figure 7.2 : Study flowchart.....	157
Figure 7.3 : Loading plots of PCA analysis results represent the four selected provinces within the study area, with the selected districts for Aksaray, Karaman, Konya, and Niğde being Gülagaç, Ayrancı, Derbent, and Çiftlik, respectively.....	159
Figure 7.4 : Box plots of correlation coefficients between drought indices and crop yield (a) train set, (b) test set.....	161
Figure 7.5 : Spatial distribution of correlation results between drought indices and crop yield	162
Figure 7.6 : Annual time series of drought indices and crop yield of the test districts.	163
Figure 7.7 : CDI, SPEI and scPDSI maps for all months of the winter wheat growing season during the 2016 dry year.	164
Figure 7.8 : The CDI drought map for winter wheat growing seasons in the study region.	166

Figure 7.9 : SPEI's drought map for the winter wheat growing season across the study.....	167
Figure 7.10 : The same as Figure 9 but for scPDSI.	167
Figure 7.11 : Long-term time series of CDI, SPEI and scPDSI for the districts in the test set.	169
Figure 7.12 : SHAP summary plot calculated from predictions obtained using the XGBoost method.	170
Figure 7.13 : Dependences plots obtained after SHAP analysis of single indices used to calculate CDI.	171
Figure A.1 : MODIS GPP algorithm flow chart.....	225
Figure A.2 : The MuSyQ GPP algorithm flowchart.....	227
Figure A.3 : Histograms and time series of MODIS datasets (a) Antalya, (b) Erzincan, (c) Izmir, (d) Kirklareli, (e) Konya, (f) Samsun, (g) Sanliurfa.....	236
Figure A.4 : Histograms and time series of MuSyQ datasets (a) Antalya, (b) Erzincan, (c) Izmir, (d) Kirklareli, (e) Konya, (f) Samsun, (g) Sanliurfa.....	237
Figure A.5 : Histograms and time series of PMLV2 datasets (a) Antalya, (b) Erzincan, (c) Izmir, (d) Kirklareli, (e) Konya, (f) Samsun, (g) Sanliurfa.....	238
Figure A.6 : Histograms and time series of TL-LUE datasets (a) Antalya, (b) Erzincan, (c) Izmir, (d) Kirklareli, (e) Konya, (f) Samsun, (g) Sanliurfa.....	239
Figure A.7 : Histograms and time series of GOSIF datasets (a) Antalya, (b) Erzincan, (c) Izmir, (d) Kirklareli, (e) Konya, (f) Samsun, (g) Sanliurfa.....	240
Figure A.8 : Histograms and time series of FLUXCOM datasets (a) Antalya, (b) Erzincan, (c) Izmir, (d) Kirklareli, (e) Konya, (f) Samsun, (g) Sanliurfa.....	241
Figure A.9 : ITA trend analysis results (horizontal and vertical axes show GPP values in g C m ² d ⁻¹), (a) Antalya, (b) Erzincan, (c) Izmir, (d) Kirklareli, (e) Konya, (f) Samsun, (g) Sanliurfa.....	242
Figure A.10 : EMD trend analysis results (the horizontal axis shows time; the vertical axis shows GPP values), (a) Antalya, (b) Erzincan, (c) Izmir, (d) Kirklareli, (e) Konya, (f) Samsun, (g) Sanliurfa.....	243
Figure B.1 : The mHM workflow.....	250
Figure B.2 : SHAP feature importance graph.....	251



A NEW AGRO-METEOROLOGICAL DROUGHT INDEX BASED ON REMOTE SENSING

SUMMARY

Drought is one of the most significant natural disasters negatively impacting ecosystems and living organisms. Unlike other natural disasters, drought cannot be directly measured due to its gradual formation and is influenced by numerous independent variables. In the literature, drought is generally categorized into four main types: meteorological drought, which occurs when precipitation is below long-term averages; hydrological drought, caused by reduced surface water flows; agricultural drought, which arises when there is insufficient moisture in the soil for crop development; and socio-economic drought, where indirect impacts affect human activities and lead to disruptions in socio-economic life. Drought cannot be directly measured; hence, drought indices are employed to mathematically define and model drought events. These indices provide valuable information about the severity, frequency, and duration of drought. The main inputs for these indices come from meteorological, hydrological, and socio-economic variables. For meteorological drought indices, precipitation is the primary input, while streamflow data are used for hydrological drought indices. For agricultural drought, soil moisture or vegetation health inputs are typically used. Drought indices can be constructed using single or multiple inputs. For example, the Standardized Precipitation Index (SPI), one of the most frequently used meteorological drought indices, is based solely on precipitation data. In contrast, the Standardized Precipitation Evapotranspiration Index (SPEI) also incorporates evaporation data, allowing for a more physically accurate representation of drought conditions in a region. Data used in calculating drought indices generally come from ground observation stations. Various measurement devices are employed for these observations, but they often result in missing or inaccurate data due to device malfunctions, human errors, or environmental factors. For instance, improperly calibrated or broken devices can cause significant data gaps. Human errors in manually collected data and environmental factors such as dust, contamination from living organisms, or nearby objects (e.g., trees) can lead to incomplete or faulty data. Missing data in these measurements can have detrimental effects on all stages of natural disaster modeling and analysis, including drought studies. Numerous data processing techniques are used to mitigate these negative impacts, such as filtering or completing missing or faulty data, thereby improving the dataset.

This thesis focuses on the development of a novel drought index, utilizing both ground observations and remote sensing data. To ensure that this new index makes a significant contribution to the literature, alternative inputs to traditional ones were explored. Initially, an attempt was made to model and use evaporation data alongside precipitation, but due to critical gaps in the remote sensing data (specifically for Türkiye), the second phase of the study shifted focus towards Gross Primary Productivity (GPP) values. GPP, representing both carbon and water cycles, was examined to assess how it could be obtained and its current status in Turkey before moving on to the index development phase.

The thesis comprises six chapters. The first chapter presents a review article on methods for addressing missing data in the field of evaporation. The second chapter features a study proposing an innovative model for completing missing values in solar radiation measurements, which directly impact evaporation (and, by extension, drought). The third chapter covers a study utilizing modern modeling techniques to forecast future evaporation values. In the first study of the thesis's second phase, covered in the fourth chapter, a bibliometric analysis of the literature on GPP estimation was conducted. The fifth chapter presents a study analyzing the current state of GPP values in Turkey using statistical methods across seven geographic regions. The analysis incorporates classical trend analysis methods alongside innovative graphical and signal processing techniques. Finally, the last chapter introduces a new agricultural drought index based on remote sensing data.

In the first chapter, an extensive literature review revealed that traditional methods for completing missing data in evaporation studies predominantly use classical methods. Some recent studies have employed modern machine learning-based methods, but most studies lack careful analysis before, during, and after applying these methods. For instance, assumptions of traditional methods were often not explored, and hyperparameter optimization for modern methods was generally neglected. Moreover, the accuracy of the data completion post-process was rarely investigated. The review concluded that studies in the literature should be encouraged to use deep learning methods more extensively.

In the second chapter, a more accurate and innovative missing data completion method was proposed based on the detailed review in the first chapter. The XGBoost algorithm, a modern machine learning method, was employed for the data completion process, and a novel hybrid model using differential evolution algorithm for hyperparameter optimization was developed. The proposed model's success was compared with classical methods, and it was found to significantly outperform them. This innovative method introduced a fast and highly accurate missing data completion algorithm to the literature.

The third chapter presents future projections of evaporation values calculated using ground observation data and the FAO-56 Penman-Monteith equation. Since evaporation greatly influences drought, understanding its future status is of critical importance. Therefore, a time series forecasting study was conducted. Time series data for evaporation were obtained, and machine learning methods combined with signal processing techniques were used for future value predictions. After performing SSA analysis and separating the time series into sub-bands of different frequencies, SGB was used for the prediction. The study demonstrated that the hybrid method, which incorporated signal processing, achieved much higher accuracy than simpler models.

The fourth chapter conducted a bibliometric analysis to identify how GPP is estimated, the key research terms in the field, and the areas with the most potential for future research. A total of 3442 studies from the Scopus database were analyzed. Several quantitative analyses were conducted, including the number of studies over time, the countries and institutions that produced the most research, and the trend topics. The analysis identified machine learning and remote sensing as areas with high potential for development, while the keyword "drought" emerged as a niche theme.

In the fifth chapter, a trend analysis of GPP values specific to Turkey was conducted. Using data from remote sensing systems, the spatial and temporal analyses of various GPP datasets were carried out. To represent Turkey's overall situation, one pilot study

area was selected from each of the seven geographic regions, focusing on forested or agricultural areas. For the temporal analysis of datasets with different time intervals, innovative methods based on signal processing and graphical models were utilized alongside classical trend analysis techniques.

In the final section of this study, a new remote sensing-based agricultural drought index was developed, incorporating NDWI and GPP alongside traditional drought indices. Winter wheat yield data from 40 rainfed districts were used to ensure the index accurately reflected crop conditions. Machine learning methods were applied to derive the index, which showed a strong correlation with yields. Compared to classical methods and widely used indices like SPEI and scPDSI, the new index demonstrated a higher correlation with yield values across most of the study area.

This thesis contributes a new agricultural drought index to the literature. All input data for this index were sourced from remote sensing systems and are freely available. In future studies, higher resolution drought maps can be developed for Turkey using these data. Additionally, the development of different drought indices for various crop types is encouraged, enabling faster and more effective drought analyses.



UZAKTAN ALGILAMA TEMELLİ YENİ BİR AGRO-METEOROLOJİK KURAKLIK İNDEKSİ

ÖZET

Kuraklık canlı yaşamını ve ekosistemlerini olumsuz yönde etkileyen en önemli doğal afetler arasındadır. Oluşumu bakımından diğer doğal afetlerden görece farklılaşan kuraklığın, doğrudan ölçümü yapılamaz ve birçok farklı bağımsız değişkenin etkisi altındadır. Farklı birçok türe sahip olan kuraklık, literatürde genel olarak 4 ana başlık altında toplanmaktadır. Yağışların uzun dönemli ortalamaların altında seyretmesi ile oluşan meteorolojik kuraklık, yeryüzü sularının akışa geçen miktarlarında meydana gelen azalma ile oluşan hidrolojik kuraklık, tarım ürünlerinin gelişimi için toprakta yeterli nemin bulunmaması ile oluşan tarımsal kuraklık ve insanların dolaylı yoldan olumsuz etkilere maruz kaldığı, sosyo-ekonomik yaşamda aksamalara neden olan sosyo-ekonomik kuraklık. Doğrudan ölçümü olmayan kuraklık hadiselerinin matematiksel olarak tanımlanabilmesi ve modellenebilmesi için kuraklık indisleri kullanılmaktadır. Bu indisler kuraklığın şiddeti, sıklığı ve süresi hakkında bilgiler edinmeye yardımcı olmaktadır. İndislerin oluşturulmasında ana girdiler meteorolojik, hidrolojik ve sosya-ekonomik değişkenlerden oluşmaktadır. Meteorolojik kuraklık indislerinin ana girdisini yağışlar oluştururken, hidrolojik kuraklık indislerinin girdisi akış değerleridir. Tarımsal kuraklık için genel olarak toprak nemi veya vejetasyon sağlığı girdileri kullanılır. Bu girdiler ile oluşturulan indisler, tekil veya çoğul girdilere sahip olabilmektedir. Örneğin literatürde çok sık kullanılan ve meteorolojik kuraklık indisi olan Standartlaştırılmış Yağış İndeksi (SPI) sadece yağış değerleri kullanılarak oluşturulur. Yine meteorolojik indislerden biri olan SPEI ise yağış değerlerinin yanında buharlaşma değerlerini de kullanmaktadır. Bu sayede bir bölgedeki kuraklık hadiselerinin, fiziksel olarak gerçeğe daha yakın şekilde ele alınması sağlanabilmektedir.

Kuraklık indislerinin elde edilmesinde kullanılan veriler genel olarak yer gözlem istasyonları aracılığı ile elde edilen verileri kapsamaktadır. Bu gözlemler için birçok farklı ölçüm cihazı kullanılmaktadır ve bu cihazların kullanımı sırasında önemli derecede eksik veya hatalı veri ile karşılaşılabilir. Eksik veya hatalı verilerin kaynakları ölçüm cihazı kaynaklı, insan kaynaklı ve çevresel etkilerden kaynaklı olabilmektedir. Ölçüm cihazlarının doğru kalibre edilmemesi veya bozulan cihazlar en önemli eksik veri kaynağıdır. Ölçüm verilerinin manuel şekilde toplandığı cihazlarda insan hatası da gözlemlenebilmektedir. Son olarak ölçüm cihazının sağlıklı şekilde veri toplamasını etkileyen çevresel faktörler ki bunlar toz, canlı teması sonucu kirlenme ve ağaç gibi cisimlerin fiziksel olarak çok yakında bulunması, ölçüm sonuçlarında eksiklik veya hataya sebep olabilmektedir. Ölçümlerde meydana gelen eksiklikleri tüm doğal afetlerin modellenmesi ve analiz edilmesi aşamalarında olumsuz etkilere neden olabildiği gibi kuraklık çalışmalarında da çok büyük etkilere sebep olabilmektedir. Bu olumsuz etkilerin bertaraf edilmesi için birçok veri işleme yöntemi kullanılmaktadır. Bu yöntemler sayesinde eksik veya hatalı veriler ayıklanabilmekte ya da tamamlama yolu ile veri yapısı iyileştirilebilmektedir.

Yer istasyonları ile elde edilen verilerin yüksek maliyetli ve sürekli bakıma ihtiyacı olması sebebi ile araştırmacılar farklı kaynaklardan elde edilen verileri kullanmak için arayışa girdiler. Bu arayış neticesinde uzaktan algılama sistemlerinden elde edilen veriler öne çıkmıştır. Sabit veya hareketli yörüngeye sahip uydu sistemleri sayesinde günümüzde birçok değişkenin ölçümü veya modellenmesi rahatlıkla yapılabilmektedir. Farklı zamansal ve mekansal çözünürlüğe sahip bu veriler genel olarak ücretsiz bir şekilde tüm araştırmacıların kullanımına açıktır. Bu veriler kullanılarak yer bilimleri alanında birçok analiz ve modelleme kolaylıkla yapılabilmektedir.

Bu tez kapsamında yeni bir kuraklık indisi geliştirmek için çalışmalar yürütülmüştür. Yeni indisi geliştirmek için hem yer gözlemleri hem de uzaktan algılama verileri etkin şekilde kullanılması amaçlanmıştır. Kuraklık indisinin literatüre önemli katkılar sunmasını sağlamak için geleneksel girdilere ek alternatif girdilerin kullanımı araştırılmıştır. İlk olarak, yağış verilerine ek olarak buharlaşma verilerinin modellenmesi ve kullanımı denenmiş ancak bu veriler için kullanılması düşünülen uzaktan algılama sistemlerindeki verilerdeki mevcut önemli eksiklikler (Türkiye özelindeki), çalışmanın ikinci evresinde GPP değerlerine odaklanılmasını doğurmuştur. Karbon ve su döngüsünü beraber şekilde ele alacak olması sebebi ile yeni indis girdileri arasında bulunan GPP'nin nasıl elde edildiği ve Türkiye'deki mevcut durum araştırıldıktan sonra indis geliştirme aşamasına geçilmiştir. Altı bölümden oluşan tezin ilk bölümünde, buharlaşma alanında eksik verilerin nasıl tamamlandığına ait çalışmaları içeren bir derleme makalesi paylaşılmıştır. İkinci bölümde ise buharlaşmayı doğrudan etkileyen (dolayısı ile kuraklık için de çok önemli olan) en önemli meteorolojik değişken olan solar radyasyon ölçümlerinin eksik değerlerinin tamamlanması için yenilikçi bir modelin oluşturulduğu araştırma makalesi paylaşılmıştır. Üçüncü bölümde ise buharlaşma değerlerinin geleceğe dönük tahminlerinin yapıldığı ve modern modelleme tekniklerinin kullanıldığı bir araştırma makalesi ele alınmıştır. Tez çalışmasının ikinci evresinin ilk çalışması ve tezin dördüncü bölümü olan çalışmada ise GPP değerlerinin tahmin edilmesine yönelik literatürdeki çalışmalar ışığında bir bibliometrik analiz yapılmıştır. Beşinci bölümde, tezin çalışma alanı olan Türkiye özelinde mevcut GPP değerlerinin durumu, yedi coğrafi bölgeyi kapsayacak şekilde, istatistiksel yöntemler aracılığı ile analizini içeren bir araştırma makalesi paylaşılmıştır. Bu analiz çalışması kapsamında literatürde sıklıkla kullanılan klasik trend analizi yöntemleri yanında, grafik tabanlı ve sinyal işleme tabanlı yenilikçi yöntemler kullanılmıştır. Tezin son bölümünde ise uzaktan algılama temelli yeni bir tarımsal kuraklık indisinin tanıtıldığı bir araştırma makalesi sunulmuştur.

Tez kapsamındaki birinci bölümde yapılan çalışmaların sonuçları incelendiğinde, literatürdeki buharlaşma alanında veya buharlaşma hesabında kullanılan verilerdeki eksik değerlerin tamamlandığı çalışmalarda büyük oranda klasik tamamlama yöntemlerinin kullanıldığı görülmüştür. Güncel bazı çalışmalarda ise makine öğrenmesi tabanlı modern yöntemlerin kullanıldığı görülmüştür. Ayrıca eksik verilerin tamamlanması için kullanılan klasik veya modern yöntemlerin kullanımı öncesi, sonrası ve sonrasındaki analizlerin çoğu çalışmada özenli şekilde yapılmadığı tespit edilmiştir. Klasik tamamlama yöntemlerinin varsayımlarının genel olarak araştırılmadığı, modern yöntemlerin ise hiperparametre optimizasyonlarının genel olarak yapılmadığı gözlemlenmiştir. Tüm bunlara ek olarak tamamlama sonrasında ne ölçüde doğru tamamlama yapıldığı araştırılmamıştır. Literatürdeki çalışmaların derin öğrenme yöntemlerini kullanmaları önemli ölçüde teşvik edilmesi gerektiği sonucuna da ayrıca varılmıştır.

İkinci bölümde ise, birinci bölümdeki detaylı inceleme sonucundaki analiz adımlarının daha doğru şekilde yürütüldüğü yenilikçi bir eksik veri tamamlama yöntemi önerilmiştir. Bu bölümdeki çalışmada modern makine öğrenmesi yöntemlerinden olan XGBoost yöntemi tamamlama için kullanılmıştır. XGBoost yöntemi hiperparametrelerini optimize edilmesi için ise diferansiyel gelişim algoritması kullanılarak yenilikçi hibrid bir model oluşturulmuştur. Önerilen modelin tamamlama başarısı, klasik tamamlama yöntemleri ile karşılaştırılmış ve önemli ölçüde başarılı olduğu ispatlanmıştır. Önerilen bu yöntem sayesinde hızlı ve yüksek doğrulukta çalışan bir eksik veri tamamlama algoritması literatüre kazandırılmıştır.

Üçüncü bölümde ise yer gözlem istasyonlarından elde edilen veriler ile hesaplaması yapılan FAO-56 Penman-Monteith buharlaşma değerlerinin geleceğe yönelik projeksiyonları yapılmıştır. Buharlaşma değerlerinin kuraklığı önemli ölçüde etkilediği düşünüldüğünde, bu değerlerin gelecekteki durumu hakkında bilgi edinmek hayati öneme sahiptir. Bu sebeple bir zaman serisi tahmin çalışması gerçekleştirilmiştir. Buharlaşma değerlerinin zaman serisi elde edilmiş ve gelecekteki değerlerinin tahmini için makine öğrenmesi yöntemleri ve sinyal işleme yöntemleri birlikte kullanılmıştır. SSA analizi sonrasında farklı frakansta alt bantlara ayrılan zaman serisi değerleri, SGB yöntemi ile tahmin edilmiştir. Kısa ve uzun vadeli tahminlerin yapıldığı çalışmada, yalın modellere nazaran sinyal işleme yöntemleri ile hibridize edilen yöntemin tahmin başarısının oldukça yüksek olduğu sonucuna varılmıştır.

Dördüncü bölümde gerçekleştirilen çalışmada GPP değerlerinin nasıl elde edildiğinin ve gelecekte bu alandaki anahtar kelimelerin neler olabileceği, hangi alanlarda potansiyelin olduğunun tespiti için bibliometrik bir analiz yapılmıştır. Analiz için Scopus veri tabanında bulunan çalışmalar (3442 adet çalışma) incelenmiştir. Çalışmaların yıllar içerisindeki sayısı, bu alanda en çok çalışmayı üreten ülkelerin ve kurumların analizi gibi sayısal birçok analiz gerçekleştirilmiştir. Bu analizlerin yanında trend olan kelimeler ve tematik analizler sonucunda elde edilen yükselmekte veya düşüşte olan araştırma alanları analiz edilmiştir. Analiz sonuçlarına göre makine öğrenmesi ve uzaktan algılama gelişme potansiyeli yüksek alanlar olarak tespit edilmiştir. Ayrıca kuraklık anahtar kelimesinde niş temalar arasında kendine yer bulmuştur.

Beşinci bölümde, Türkiye özelinde GPP değerlerinin mevcut durumu ve trend analizi gerçekleştirilmiştir. Uzaktan algılama sistemeleri tarafından elde edilen veriler ışığında oluşturulan farklı GPP veri setlerinin mekansal ve zamansal analizi yapılmıştır. Türkiye'nin genel durumunu temsil etmesi açısından 7 coğrafi bölgeden birer adet pilot çalışma bölgesi seçilmiştir. Bu alanlar genel olarak ya orman alanları ya da tarım alanlarını içeren bölgeler olmasına dikkat edilmiştir. Farklı zaman aralıklarına sahip veri setlerinin zamansal analizi için klasik trend analizi yöntemleri ve yenilikçi trend analizi yöntemleri karşılaştırmalı olarak kullanılmıştır. Tüm bölgelerdeki GPP veri setlerinin istatistiksel olarak farklılık sergilediği görülmüştür. Analiz sonucunda Türkiye özelinde GPP değerlerinde dünya geneli ile uyumlu olacak şekilde artan bir trendin olduğu sonucuna varılmıştır. Ayrıca sinyal işleme yöntemlerinin kullanımı ile yapılan trend analizleri, zaman içerisindeki trendin nerede başlayıp nerede sonlandığını ya da hangi şekilde bir trendin olduğu konusunda klasik yöntemlere nazaran daha detaylı sonuçlar verdiği gözlemlenmiştir.

Altıncı ve son bölümü kapsayan makalede ise uzaktan algılama temelli yeni bir tarımsal kuraklık indisi geliştirilmiştir. İndisin geliştirilmesi için literatürde sıklıkla

kullanılan tekil kuraklık indislerine ek olarak NDWI ve GPP deęerleri de eklenmiřtir. Kuraklık hadiselerinin tespit edilmesi için aynı bölgede bulunan, 40 adet ilçenin kışık buęday verim deęerleri kullanılmıřtır. Seilen bölgede genel olarak kuru tarımın yapılması geliřtirilen indisin bitki durumu ile yakından iliřkili olmasını saęlamıřtır. Yeni indisin elde edilmesi için makine öęrenmesi yöntemleri ve model açıklama algoritmaları kullanılmıřtır. Literatürde daha önce indis denkleminin aęırlıklarının hesap edilmesi için kullanılan klasik yöntemler ile yapılan karřılařtırmalar neticeisnde elde edilen bileřik kuraklık indisinin kışık buęday verimleri ile iliřkili olduęu tespit edilmiřtir. Ayrıca literatürde sıklıkla kullanılan SPEI ve scPDSI indiseleri ile yapılan karřılařtırmalarda, alıřma alanının büyük bölümünde önerilen indisin verim deęerleri ile daha yüksek korelasyona sahip olduęu görölmüřtür.

Tez alıřması ile literatüre yeni bir tarımsal kuraklık indisi kazandırılmıřtır. Bu indisin tüm girdilerinin uaktan algılama sistemlerinden elde edilmiř ve ücretsiz verileri kapsamaktadır. İlerleyen alıřmalarda bu veriler kullanılarak Türkiye için daha yüksek özünürlükte kuraklık haritaları geliřtirilebilecektir. Ayrıca farklı türden mahsüller için farklı kuraklık indislerinin geliřtirilmesi teřvik edilmekte, böylece daha hızlı ve etkin kuraklık analizlerinin yapılması hayata geirilebilecektir.

1. INTRODUCTION

Drought is one of the most critical environmental issues, deeply affecting both the natural environment and human livelihoods. As the impacts of climate change become increasingly evident, the frequency and severity of drought events are rising. Drought is a complex process, initiated by a decrease in the amount of precipitation below the normal levels in a specific region, and progressively leading to the depletion of water resources, a reduction in agricultural productivity, and a disruption of ecosystem balance. This process is characterized by different stages, including meteorological drought, agricultural drought, hydrological drought, and socio-economic drought, each of which negatively affects human activities and the natural balance. Drought has a particularly profound impact on agricultural production, making it one of the key factors threatening food security. Losses in agricultural yields not only disrupt economic stability but also contribute to rising food prices and deepening poverty in rural areas. In countries with semi-arid climates, such as Türkiye, drought events occur more frequently and with greater intensity, posing significant challenges to agricultural production and water resource management.

Effective drought monitoring and management must be based on accurate and reliable information. To this end, drought indices have become crucial tools for assessing the severity and extent of drought events. These indices are calculated using various methods, including meteorological data, remote sensing technologies, and hydrological models, and they serve as essential guides for decision-makers in drought mitigation efforts. In recent years, the development of combined drought indices, which integrate agricultural, meteorological, and hydrological components, has yielded more effective results in drought monitoring and prediction.

The primary objective of this thesis is to develop a combined drought index for a region in Türkiye using individual drought indices. To achieve this main objective, several sub-objectives have been established, and the thesis is structured accordingly. The thesis is divided into two major sections.

In the first section, the focus is on pre-analysis steps, including data processing and data treatment methods. This section is organized into three chapters:

The first chapter presents a systematic review of methods used for imputing missing data, particularly in the context of evaporation data. This comprehensive review explores various techniques for addressing data gaps in environmental and hydrological studies.

The second chapter introduces a novel approach for imputing missing data, specifically targeting the completion of solar radiation data. This new algorithm enhances the accuracy and completeness of meteorological datasets, which is crucial for improving the reliability of drought indices and hydrological models.

The third chapter focuses on the forward-looking prediction of evaporation values using modern techniques. These predictive models play a critical role in understanding evaporation dynamics and their impact on drought conditions, which is especially important for regions where direct evaporation measurements may be sparse or unavailable.

In the second section, the emphasis shifts to the estimation of Gross Primary Productivity (GPP) values, which were selected to replace evaporation data in the drought index model. This section also contains three chapters:

The first chapter conducts a bibliometric analysis of studies focused on estimating GPP values. This analysis provides a comprehensive overview of the methodologies and trends in the literature, highlighting the growing importance of GPP as a key indicator in ecosystem and drought studies.

The second chapter presents a spatio-temporal analysis of GPP values observed across Türkiye. Using advanced trend analysis techniques, this chapter offers insights into the temporal and geographical variations in GPP, providing a critical understanding of how ecosystem productivity responds to climatic and environmental changes in different regions of Türkiye.

The third chapter details the development of a new combined drought index (CDI). This chapter outlines the integration of agricultural, meteorological, and hydrological data—specifically GPP—using machine learning techniques. The resulting CDI is

tailored to regional and crop-specific drought monitoring, leveraging remote sensing data and machine learning to create a more accurate and responsive tool for assessing agricultural drought. The contributions of this thesis to the existing body of literature can be summarized as follows:

- A systematic review of methods for addressing missing data in evaporation studies, providing a foundational reference for future research in hydrology and environmental sciences.
- The development of a novel algorithm for imputing missing solar radiation data, significantly improving the quality and completeness of meteorological datasets.
- A forward-looking predictive model for evaporation values, utilizing modern forecasting techniques to enhance understanding of evaporation dynamics under varying climatic conditions.
- A bibliometric analysis of Gross Primary Productivity (GPP) research, offering a detailed examination of global trends and methodologies in this rapidly evolving field.
- A spatio-temporal analysis of GPP values in Türkiye, providing critical insights into how climate change and environmental variability affect ecosystem productivity over time and across regions.
- The development of a crop and region specific agricultural drought index, designed using machine learning methods and incorporating entirely remote sensing-based inputs, offering a powerful new tool for drought monitoring and agricultural management in semi-arid regions.

This thesis makes significant contributions by integrating advanced data imputation techniques, spatio-temporal analysis, and machine learning-based drought indices, offering new insights and practical tools for drought monitoring and prediction in Türkiye and beyond.



2. PROVIDING A COMPREHENSIVE UNDERSTANDING OF MISSING DATA IMPUTATION PROCESSES IN EVAPOTRANSPIRATION-RELATED RESEARCH: A SYSTEMATIC LITERATURE REVIEW¹

2.1 Introduction

The quality of measurements is of utmost importance for the majority of research areas, such as hydrology, meteorology, and earth science (Yozgatligil et al, 2013). Despite the notable advancements in technology over the past few decades, the acquisition of highly accurate measurements remains a significant challenge due to the potential for measurement errors. Such errors can occur due to various factors such as inadequate calibration, imprecise measurement readings, and environmental influences (e.g. proximity to man-made or natural structures, air pollution, etc.), which can negatively impact all stages of data analysis (Lin and Tsai, 2020). To tackle these challenges, one can either remove missing data or employ imputation methods to assign substitute values (Langkamp et al, 2010). Here, the amount of missing data plays a critical role; missing data accounting for less than 10% or 15% of the entire dataset can be removed (Strike et al, 2001), whereas one should pay considerable attention to addressing missing values exceeding 15%.

The research community has increasingly focused on statistical models with varying degrees of complexity to minimize the efforts required to handle the missing data problem. The aim is to develop methods that require minimal assumptions while producing accurate estimations that closely align with the actual measured data. The pertinent literature recognized the critical role played by the classical methods, e.g. linear regression (LR) analysis, principal component analysis (PCA), and expectation-maximization (EM) (Hron et al, 2010). Despite their practical utilization, these algorithms require some prerequisites that pushed researchers to adopt different

¹ This chapter is based on the paper “Başakın, E. E., Ekmekcioğlu, Ö., and Özger, M. (2023). Providing a comprehensive understanding of missing data imputation processes in evapotranspiration-related research: a systematic literature review, *Hydrological Sciences Journal*, 68(14), 2089-2104. doi: 10.1080/02626667.2023.2249456.”

methods for accomplishing missing data imputation attempts. In this sense, modern algorithms (mostly concentrated on machine learning applications) are used for a plethora of purposes, such as imputation of river flow data (Sidibe et al, 2018), imputation of water quality data (Tabari and Talaee, 2015; Zhang and Thorburn, 2022), imputation of meteorological data (Sattari et al, 2020; Wang and Shi, 2021), etc. The pertinent literature indicates that modern modelling techniques tend to produce more accurate estimations compared to classical imputation methods, thus offering promising solutions for addressing the challenges of missing data (Gautam and Ravi, 2015; Silva-Ramírez et al, 2015; Zhang et al, 2015). Studies using modern techniques are explained in detail in the following sections.

The current paper focuses on the imputation of missing values observed in various systems to facilitate the computation of actual evapotranspiration (ET_a), along with variables aiding in the calculation of reference evapotranspiration (ET_o). Hence, this research encapsulates a comprehensive investigation of data imputation techniques for the parameters utilized in ET measurements and calculations. In this vein, studies carrying out imputation of actual ET values based on eddy covariance (EC), Bowen ratio, and lysimeter measurements as well as other efforts devoted to reconstructing temperature, humidity, wind speed, and solar radiation to assist in calculating the reference ET were diligently investigated. The present research sheds light on the selection of input parameters for imputation methods, the selection of hyperparameters in algorithms used during the imputation process, and the determination of iteration numbers for multiple imputation techniques. The software tools used in the imputation process were further investigated within the scope of the current study.

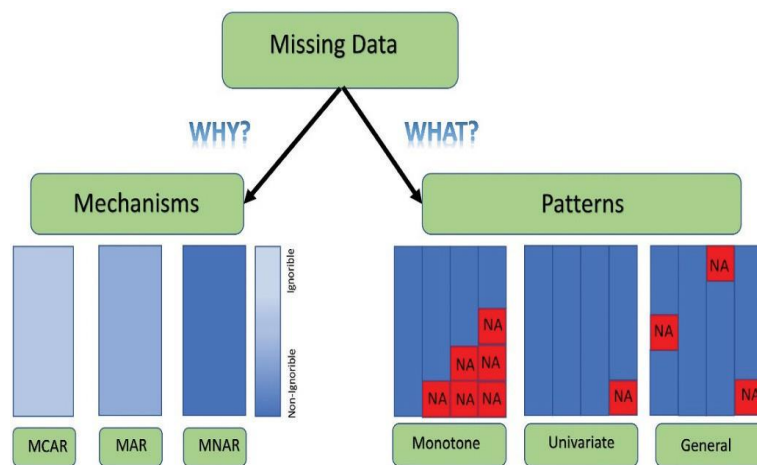


Figure 2.1 : Mechanisms and patterns of missing data. MAR = Missing at Random; MCAR = Missing Completely at Random; MNAR = Missing Not at Random.

2.2 Mechanisms And Patterns Of Missing Data

The mechanisms and patterns also play a significant role in the resultant research outcomes, which may have a greater impact than the extent of missing data itself. Figure 2.1 visualizes the missing data mechanisms and patterns.

2.2.1 Missing data mechanisms

In imputing missing data, it is crucial to first identify the structural features of the missing data. In particular, it is necessary to designate whether the missing data in the observations are randomly distributed or follow a certain pattern. In this context, the classification system of missing data that was summarized by Rubin (1976), which treats missing data as a probabilistic process, has been widely used in recent decades. Here, among three different mechanisms, missing completely at random (MCAR) is regarded as the least challenging mechanism for statistical analyses as it grounded on the independent relationship between the missing data and other variables in the dataset. On the other hand, missing at random (MAR) assumes that there is a dependent relationship between the variable having missing values and other variables. In this mechanism, the probability of missing data relies on the observed values of other variables, rather than the other missing values of the target variable. In addition, missing not at random (MNAR) is based on a direct dependent relationship between missing data and the nature of the corresponding variable. Hence, datasets having relatively longer gaps are generally characterized by the MNAR mechanism.

2.2.2 Missing data patterns

According to Figure 2.1, a potential indicator emerges regarding the possibility of data deletion within certain groups in the context of missing data mechanisms. When missing data is deemed negligible, this implies that the likelihood of observing a variable is not contingent upon the value of that variable, while the converse denotes non-negligible missing data (Beaulieu-Jones et al, 2018). On the one hand, in instances where the missing data mechanism is MCAR or MAR, data deletion may result in minimal information loss. On the other hand, in the case of an MNAR mechanism, it is recommended to employ data imputation as the missingness is attributable to the impact of other independent variables.

Furthermore, the utilization of missing data imputation techniques is based on the patterns of missing data as the corresponding patterns indicate the location of missing values in the focalized time series. The existing literature covers many missing data patterns, mainly divided into two groups, i.e. monotone or non-monotone (general) and univariate or multivariate. Let the data be represented by the $n \times p$ matrix. In the presence of missing data, Y is partially observed. Notation Y_j is the j -th column in Y , and Y_{-j} indicates the complement of Y_j – that is, all columns in Y except Y_j . The missing data pattern of Y is the $n \times p$ binary response matrix R . A missing data pattern can be regarded as univariate if there is only one variable with missing data, whereas, once a set of series are all observed or missing on the same set of cases, the corresponding pattern is called multivariate. In addition, a missing data pattern is regarded as monotone if the variables Y_j can be ordered such that if Y_j is missing then all variables Y_k with $k > j$ are also missing. However, if the pattern is not monotone, it is called non-monotone or general.

2.3 Methods For Measuring And Calculating ET

2.3.1 The eddy covariance method

The EC method is widely used to measure fluxes in ecosystem studies across the globe. Flux refers to the amount of matter passing through a closed surface per unit area in a unit of time. The EC method describes the airflow in the atmosphere as consisting of vertical movements that transport materials through turbulence. The difference between the number of gas molecules transported upwards and those transported downwards at any given moment results in net convection. In the EC method, the concentration of the gas being measured and the covariance of the vertical wind speed are the fundamental principles used for estimating gas exchange between the atmosphere and the surface (Burba et al, 2013). Equation (2.1) shows how the evaporation latent heat flux (LE) is calculated:

$$LE = \rho_a \overline{w'q'} \quad (2.1)$$

Here, LE represents the latent heat of evaporation ($W m^{-2}$), ρ_a is the air density ($kg m^{-3}$), w' represents the deviation of the vertical wind speed from the mean ($m s^{-1}$), q'

represents the deviation of the water vapor concentration from the mean (kg m^{-3}), and $\overline{(w'q')}$ represents the covariance of these deviations (Jiang et al, 2014).

2.3.2 Bowen ratio energy balance method

The Bowen ratio energy balance (BREB) method is used to determine the actual ET based on the ratio of distribution of the convective flux between the evaporative latent heat flux and the sensible heat flux. The BREB method involves measuring the temperature and humidity gradient, as well as net radiation and soil heat flux. The BREB method is based on the energy balance equation (Bowen, 1926):

$$R_n - G = LE + H \quad (2.2)$$

Here; R_n represents net radiation (W m^{-2}); G represents the soil heat flux (W m^{-2}); LE represents evaporative latent heat flux (W m^{-2}), and H represents sensible heat flux (W m^{-2}). The method is based on the Bowen Ratio calculation introduced by Bowen, which shows the distribution ratio of the receiving energy between LE and H (Bowen, 1926).

$$\beta = \gamma \left[\frac{T_1 - T_2}{e_{a1} - e_{a2}} \right] = \frac{H}{LE} \quad (2.3)$$

where β represents the Bowen Ratio, γ indicates the psychrometric coefficient ($\text{kPa } ^\circ\text{C}^{-1}$), T_1 represents the air temperature ($^\circ\text{C}$) measured at the lower level, T_2 represents the air temperature ($^\circ\text{C}$) measured at the upper level, e_{a1} represents the actual vapor pressure (kPa) of the air at the lower level, e_{a2} represents the actual vapor pressure (kPa) of the air at the upper level. With the determination of the Bowen ratio, LE (W m^{-2}) can be calculated by using the expression given in equation 2.4.

$$LE = \frac{R_n - G}{1 + \beta} \quad (2.4)$$

2.3.3 Weighing lysimeter method

The water budget equation was used for ET measurements in lysimeters. For a typical lysimeter tank insulated from the environment, the water budget equation can be written as in equation 2.5.

$$(P + I + C_p + R_{fc}) - (ET + D_p + R_{fo}) = \mp \Delta W \quad (2.5)$$

where P is precipitation, and I is irrigation water. Also, C_p and R_{fc} are denoted as capillary rise and runoff, respectively. In equation 10, ET, D_p , and R_{fo} indicate evapotranspiration, deep infiltration or drainage water, and runoff, respectively. ΔW shows the changes in soil water content. It is worth mentioning that all units are in millimeters. In general, C_p , R_{fc} , and R_{fo} values are zero in lysimeters. Thus, the equation for ET ends up with equation 2.6 as follows:

$$ET = P + I - D_p \mp \Delta W \quad (2.6)$$

2.3.4 FAO-56 Penman-Monteith method

The FAO Penman-Monteith method is an equation derived from the original Penman-Monteith equation along with the aerodynamic and surface resistance equations (Allen et al, 1998).

$$ET_0 = \frac{0.408(R_n - G)\gamma \frac{900}{T + 273} u_2 (e_s - e_a)}{\Delta + \gamma(1 + 0.34u_2)} \quad (2.7)$$

where ET_0 is the reference evapotranspiration [mm day⁻¹], R_n is the net radiation at the crop surface [MJ m⁻² day⁻¹], G is the soil heat flux density [MJ m⁻² day⁻¹], T is the mean daily air temperature at 2 m height [°C], u_2 is the wind speed at 2 m height [m s⁻¹], e_s is the saturation vapor pressure [kPa], e_a is the observed vapor pressure [kPa], $e_s - e_a$ refers to the saturation vapor pressure deficit [kPa], Δ represents the slope of vapor pressure curve [kPa °C⁻¹], and γ denotes the psychrometric constant [kPa °C⁻¹].

2.4 Literature Review Rationale

In this study, missing data imputation studies of the parameters used to obtain actual and reference ET values were examined. This research utilized the Scopus search engine to conduct the comprehensive literature survey. To ensure relevance, the search was limited to articles published in English between 2002 and 2022, reflecting the most recent research trends in the field. A total of 282 articles were subjected to the abstract-level investigation and the number of studies was reduced to 110. Finally, the articles deemed directly relevant to the missing data imputation of the variables taken

into account were further analyzed through the full-text investigation which resulted in a total of 62 articles. The entire process of the literature review is illustrated in Figure 2.2.

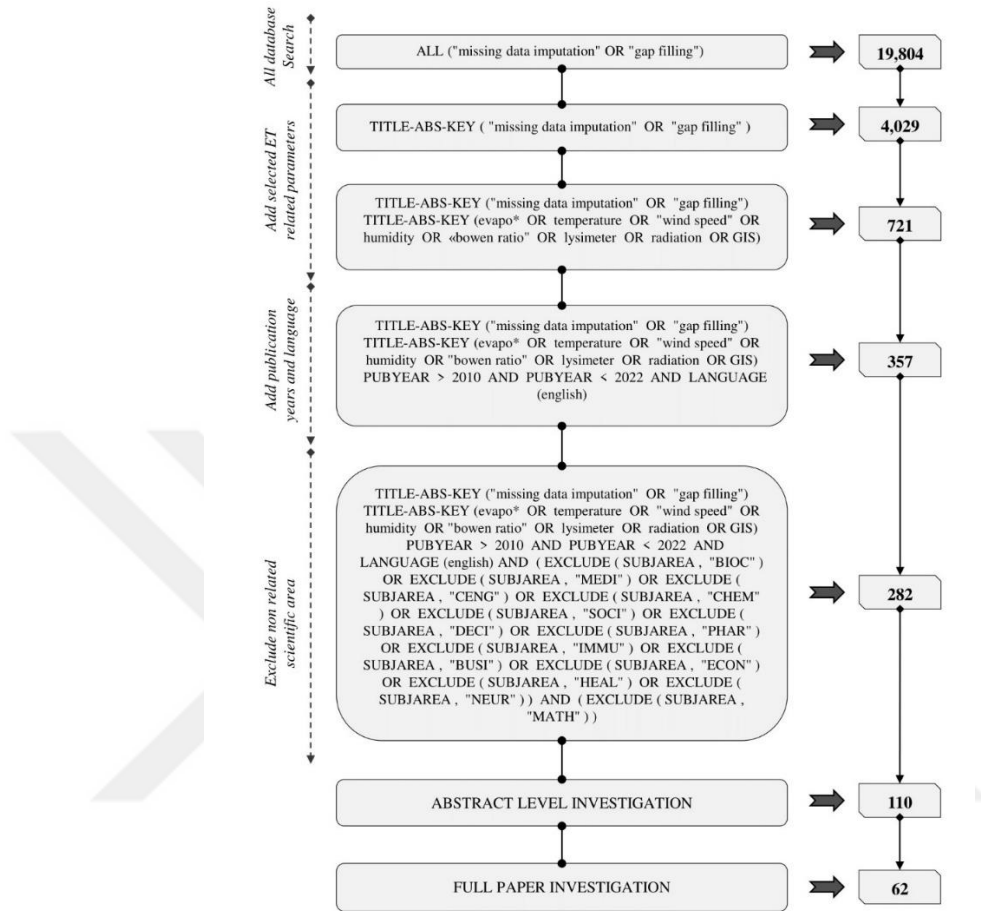


Figure 2.2 : The steps of the literature review conducted via Scopus.

2.5 Results

Three methods, i.e. EC, Bowen ratio, and lysimeter systems, that are used to measure actual ET values were examined in the present research. For reference ET calculation, the parameters (i.e. temperature, wind speed, pressure, insolation, radiation, and humidity) of the FAO-56 Penman-Monteith method, which is one of the most commonly adopted empirical approaches in the literature, were considered. It is also worth noting that the review of the pertinent literature revealed that there is no study on the imputation of missing data during the measurement of pressure and insolation times.

Along with the content-based investigation, the present work also aimed to highlight the scientometrics of the relevant studies. Therefore, a total of 62 articles extracted

through the Scopus engine were subjected to the VOSviewer software and a keyword-based analysis was performed to explore the knowledge distribution over years. In Figure 2.3, elements' size is directly related to the occurrence frequency of the keywords, while the colour represents distinct time periods associated with the keywords, and the links between the nodes symbolize the co-occurrence relationship between the words. The corresponding analysis revealed that the early studies (represented by deep blue) mainly concentrated on the ET calculations were linked with climatic data, data obtained from EC and Bowen-ratio systems. However, most studies primarily focused on filling gaps to perform further processes rather than directly embracing the missing data imputation objective. In addition, recent attempts tend to discover the contribution of air temperature, land surface temperature (LST), and wind speed, while the research community also focalized on the integration of classical techniques in gap filling. Still, few studies have acknowledged the role of advanced modelling techniques such as machine learning approaches.

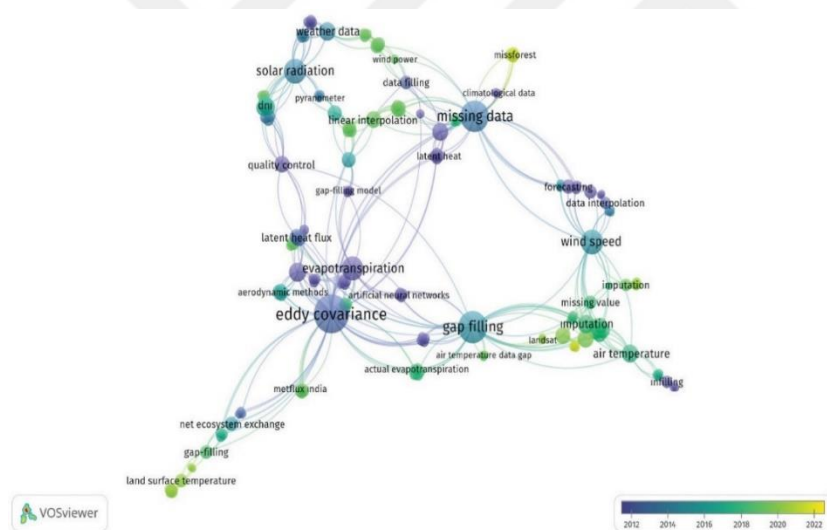


Figure 2.3 : Keyword networks generated by VOSviewer based on the Scopus database.

2.5.1 Missing data imputation techniques

The existing literature recognized the critical role played by divergent approaches (e.g. statistical, data-based, machine learning techniques, etc.) in missing data imputation to acquire precise estimations. Hence, in the present research we diligently reviewed the pertinent literature and extracted the commonly utilized methodologies to provide a better overview regarding the definitions, pros, and cons of the corresponding techniques (Table 2.1).

Table 2.1 : Commonly applied missing data imputation methods.

Method	Basis	Rationale	Assumption	Complexity	H. P.	C. C.	Explanation
Deletion	N/A	N/A	N/A	Low	N/A	Low	It removes observations containing missing values from the data set. Advantages: Easy to implement. Disadvantages: Information loss
Simple Imputation	Statistical	N/A	N/A	Low	N/A	Low	It replaces missing values by the estimations derived from the quantitative attribute of the available values. Advantages: Easy to implement. Disadvantages: Prone to change the original distribution of the variable.
Mean Imputation	Statistical	N/A	N/A	Low	N/A	Low	It replaces the missing value by the arithmetic mean of the remaining dataset. Advantages: Easy to implement. Disadvantages: Potentially distort imputations due to unequal data distribution and the influence of extreme values on the mean.
Regression Imputation	Statistical	N/A	Normal Distribution	Low	N/A	Low	It estimates the relationship between input and output variables, or between a data point and its corresponding variable. Advantages: Easy to implement. Disadvantages: Requiring many assumptions due to its parametric nature.
Hot-Deck Imputation	Data Based	N/A	N/A	Mid	N/A	High	It replaces missing data by values from complete data cases. Advantages: Being robust to misidentification and interaction effects as the imputations are generated using an implicit model. Disadvantages: Underestimating standard errors due to reduced variability that results from the selection of completed data from already observed data. In addition, when the data set is not MCAR, data imputations can result in unpredictable bias.
Expectation–Maximization	Statistical	N/A	Distribution	Mid	N/A	High	It is an iterative approach utilized to compute the maximum likelihood estimation of variables within statistical models. Advantages: Better than mean imputations as it preserves the relationship with other variables. Disadvantages: It does not provide standard error estimations.
Multiple Imputation	Statistical	N/A	N/A	Mid	N/A	High	The regression assignment is re-operated more than once and each dataset is stored for later use, while the obtained errors are calculated. Then, the real dataset is attained by taking the average of the corresponding data. Advantages: It produces statistically valid estimates under the assumption of MAR and yields asymptotically consistent, efficient, and normally distributed estimates. Disadvantages: Due to its inherent variability, it often results in different estimates each time leading to different outcomes among researchers working with the same dataset.
Linear Interpolation	Statistical	Distance-Based	Linearity	Mid	N/A	Low	It performs interpolation between a data point immediately preceding the gap and another data point shortly following the gap. Advantages: When properly utilized, it provides improved estimates for statistical inference. Disadvantages: Providing limited accuracy where continuously long missing data exists.
Kriging	Statistical	Distance-Based	Stationarity and Isotropy	Mid	N/A	Mid	It is a spatial estimation method that utilizes nearby data points to predict the values of unsampled points and uses semi-variogram structural features to make unbiased estimations of positional changes at unsampled locations. Advantages: Measuring the uncertainty of predictions and ensuring uncertainties to the user, enabling more reliable and informed decision-making processes. Disadvantages: Explaining spatial relationships using a linear model, which may result in the inability to capture complex or non-linear relationships present in the dataset.

Table 2.1 (continued) : Commonly applied missing data imputation methods.

Method	Basis	Rationale	Assumption	Complexity	H. P.	C. C.	Explanation
Inverse Distance Weighting	Statistical	Distance-Based	N/A	Mid	N/A	Low	It is grounded on the estimation of the values of unsampled points by utilizing the values of sampled points. The estimated values are a function of both distance and the size of neighboring points, and the importance and impact on the cell being estimated decrease as distance increases. Advantages: When used properly, it enhances the accuracy of statistical inference. Disadvantages: If the data density is low or unevenly distributed, it can lead to inaccurate estimation of missing values in certain regions.
K-Nearest Neighbors	Machine Learning	Distance-Based	N/A	Low	Yes	Low	It works by identifying the closest k (referring to the number of neighbors) number of data points to the data point being evaluated, and assigning it to the same class or estimating its value based on the values of those k data points. Advantages: Easy to implement. Disadvantages: It does not work well with large datasets as calculating the distance between each data instance is costly. Similarly, it is not compatible with high-dimensional data as the calculation process becomes more complex for each dimension. Additionally, the KNN algorithm is sensitive to noisy and missing data.
Decision Trees	Machine Learning	Tree-based	N/A	Mid	Yes	High	It uses a sequential approach to perform the classification or regression process, aiming to minimize the standard error to obtain an optimal decision tree. Advantages: Easy to understand, interpret, visualize and does not require normalization or scaling of data. Disadvantages: Being prone to overfitting.
Artificial Neural Networks	Machine Learning	Black Box	N/A	High	Yes	High	Inspiring from biological neural networks, ANNs are a set of mathematical operations aiming to solve inherent non-linear problems. Advantages: Non-Linear and flexible, having ability to work with incomplete information. Disadvantages: High computational costs, requiring huge amounts of data, limited explanations regarding its internal processes.
Adaptive Neuro-Fuzzy Inference System	Machine Learning	Black Box	N/A	High	Yes	Very High	It defines the relationship between inputs and outputs in fuzzy space, while neural networks try to find the most suitable parameters of the model. Advantages: Less complex compared to other decision-making systems. Disadvantages: Posing significant challenges in estimations having large number of inputs, including the curse of dimensionality, interpretability of rules, and parameter training.
Support Vector Machine	Machine Learning	Distance-Based and Rule-Based	N/A	High	Yes	High	It performs optimal classification with a single differential decision surface and is based on the principle of structural risk minimization. Advantages: Being more effective when there is a clear margin of separation between classes. Disadvantages: Having limited suitability for high dimensional spaces and the need for choosing an optimal kernel which is a challenging task.

2.5.2 Missing data imputation attempts concentrated on Bowen ratio data

The pertinent literature illustrated that studies investigated to impute the missing values in Bowen ratio measurements account for 5% of the total number of studies. The corresponding attempts mostly concentrated on the utilization of classical approaches, such as regression and linear interpolation techniques. For instance, Guo et al. (2007) employed two distinct imputation methods for completing missing LE and H values. Initially, the Priestly-Taylor method was utilized to estimate the missing LE values, followed by the completion of missing H values using the energy balance equation. This sequential imputation approach was effective in simulating both daytime evaporation and night-time condensation. The imputed fluxes closely matched the quality-controlled data for each flow scenario in terms of their temporal distribution. Similarly, in other studies using the Bowen ratio method, linear interpolation was the most widely adopted method, as can be seen from Table 2.2 (Balogun et al, 2009; Burakowski et al, 2018; Sexstone et al, 2016).

Table 2.2 : Details regarding the missing value imputation studies utilizing the Bowen-Ratio technique.

ID	Reference	Imputation methods	Method/Data	Gap size	Performance metric
1.	Burakowski et al. (2018)	LI	Bowen Ratio/SWin	<0.01%	N/A
2.	Sexstone et al. (2016)	LI	Bowen Ratio/Rn	N/A	N/A
3.	Balogun et al. (2009)	LI	Bowen Ratio/Rn	15%	N/A
4.	Guo et al. (2007)	PT, LR	Bowen Ratio/LE	1-7%	R ²

2.5.3 Missing data imputation attempts concentrated on EC data

EC measurements are significant for determining ET as they provide direct and continuous measurements of the turbulent fluxes of water vapor, heat, and momentum between the Earth's surface and the atmosphere. Table 2.3 presents the studies devoted to attempts regarding the imputation of missing values encountered in EC measurements. For instance, Tang et al. (2019) performed the classical LR method to impute missing data in LE and H values that were measured using the EC method over four years in a tropical peat forest in Sarawak, Malaysia. Concerning the software that can be directly used in missing data imputation attempts, Deb Burman et al. (2019) implemented the REddyProc software package compiled with the R program to impute

missing data in H, LE, Rn, short wave incoming solar radiation (RSWin), G, and photosynthetic photon flux densities (PPFD) provided by the EC method in a tropical forest located in Northeast India.

Table 2.3 : Details regarding the missing value imputation studies utilizing the Eddy-covariance technique.

ID	Reference	Imputation methods	Method/Data	Gap size	Performance metric
5.	Goodrich et al. (2021)	PLS, NN, KNN	Eddy Covariance/N ₂ O	15%	R ²
6.	Tang et al. (2019)	LR	Eddy Covariance/H, LE	N/A	N/A
7.	Burman et al. (2019)	REddyProc package	Eddy Covariance/H, LE, Rn, SWin, G, PPFD	N/A	N/A
8.	Abudu et al. (2010)	ANN	Eddy Covariance/LE	5,10,20,30,40%	R ² , RMSE, MAE, MRE
9.	Chen et al. (2012)	MLR, KNN	Eddy Covariance/LE	5,10,20,40%	RMSE
10.	Hui et al. (2004)	MI	Eddy Covariance/NE, LE, H	6-51%	R ²
11.	Alavi et al. (2006)	KF	Eddy Covariance/LE	4-17%	R ² , RMSE, Rbias, R
12.	Kunwor et al. (2017)	Traditional, static window method, Moving window method, Parameter prediction method	Eddy Covariance/NEE	35%	RMSE
13.	Schmidt et al. (2008)	NN	Eddy Covariance/ CO ₂	13%	RMSE
14.	Mahabbati et al. (2021)	ANN, LR, RF, SVM, XGB, FFNN	Eddy Covariance/NE, LE, H	1-32%	R ² , RMSE, MBE, VR, AI
15.	Li et al. (2007)	LI	Eddy Covariance/H, LE	N/A	N/A
16.	Eamus et al, (2013)	Self-organising linear output, ANN	Eddy Covariance/E, C	N/A	RMSE
17.	Foltýnová et al. (2020)	REddyProc package	Eddy Covariance/LE	1-100 half-hour	RMSE, MBE
18.	Boudhina et al. (2018)	REddyProc package, LR, MLR, evaporative fraction	Eddy Covariance/LE	N/A	R ² , RMSE
19.	Zitouna-Chebbi et al. (2018)	REddyProc package	Eddy Covariance/LE	10-70%	B, RB, RRMSE, R ²

Similar attempts further illustrated the effectiveness of the REddyProc package regarding the imputation of missing data in LE values (Foltýnová et al, 2020; Zitouna-Chebbi et al, 2018).

In addition, there are many studies in which missing LE data were imputed with machine learning algorithms, such as artificial neural networks (ANN) (Abudu et al, 2010; Eamus et al, 2013; Mahabbati et al, 2021; Schmidt et al, 2008), k-nearest neighbours (KNN) (Chen et al, 2012) and Kalman filter (KF) (Alavi et al, 2006). Kunwor et al. (2017) used three non-linear regression approaches to fill artificially created gaps of different sizes via light and temperature response curves: (1) “traditional” fixed monthly window, (2) moving window and (3) moving window with parameter estimation using physiological drivers. The results of the incomplete data-imputation simulations showed that the net ecosystem exchange (NEE) estimates made with the moving window and parameter estimation methods were closer to those of observed NEE, while the traditional method had yielded lower overall accuracy.

2.5.4 Missing data imputation attempts concentrated on lysimeter data

Lysimeter measurements are a valuable tool for understanding the water balance of a specific area and can provide important insights into the complex processes that drive ET. They ensure direct measurement of the water balance of a certain region, and can also be used to validate and improve remote sensing- based estimates of ET. Despite their importance, limited studies have been conducted due to the scarcity of the corresponding systems across the globe. Still, Huang et al. (2020) (which is the 62nd study investigated within the scope of the present research) artificially generated random gaps for the three time series attained by the lysimeter systems from Basel (Switzerland), Rheindahlen (Germany), and Rietholzbach (Switzerland) to explore the utility of four techniques: potential ET, ratio, FAO-based water balance and HYDRUS modelling methods. Interestingly, the results suggested that the ratio method was well suited for imputing missing values of lysimeter data for Basel (Switzerland), while the HYDRUS method outperformed its counterparts concerning Rheindahlen (Germany). For Rietholzbach (Switzerland), all methods yielded very similar results but greater root mean square error (RMSE) values were observed with the FAO method.

2.5.5 Missing data imputation attempts concentrated on humidity data

Measuring the humidity for a region plays a crucial role in calculating the vapor pressure deficit (VPD), which is a key factor in the rate of ET and is important in understanding the overall water balance of a region. For instance, Costa et al. (2021) used the multiple imputations by chained equations (MICE) algorithm to impute missing humidity data. Despite the usage of classical approaches, related studies mostly rely on the application of machine learning techniques, such that along with the adoption of Kalman smoothing and autoregressive integrated moving average (ARIMA) in their study, Afrifa-Yamoah et al. (2020) further benchmarked them with multiple linear regression (MLR) methods for missing humidity data estimation. In a similar vein, Hasanpour Kashani and Dinpashoh, (2012) used ANN, adaptive neuro fuzzy inference system (ANFIS), and genetic programming (GP) methods, while Körner et al. (2018) and Lara-Estrada et al. (2018) discovered the performance of gradient boosting and Bayesian network methods, respectively (Table 2.4).

Table 2.4 : Details regarding the studies devoted to impute missing humidity data.

ID	Reference	Imputation methods	Method/Data	Gap size	Performance metric
20.	Costa et al. (2021)	MICE	Humidity	17-54%	RMSE
21.	Kashani and Dinpashoh (2012)	AA, ID, NR, Single best estimator, MLR, UK traditional method, CSM, ANN, ANFIS, GP,	Humidity	N/A	MAE, NSE, SS
22.	Yamoah et al. (2020)	Kalman smoothing, ARIMA, MLR	Humidity	10%	MAE, RMSE, SMAPE
23.	Körner et al. (2018)	GTB	Humidity	50%	RMSE, MAE, R ²
24.	Anjomshoa and Salmanzadeh (2018)	LI, CI	Humidity	1-374 hour	VR, MAE, RMSE
25.	Estrada et al. (2018)	Bayesian Network	Humidity	20%	RMSE, Bias
26.	Lee et al. (2010)	LI	Humidity	0.1%	N/A

2.5.6 Missing data imputation attempts concentrated on solar radiation data

Measurements pertaining to the solar radiation variable help researchers understand the ET mechanisms as they provide insights into the energy available for photosynthesis, transpiration, and soil evaporation, and can be used to estimate the net radiation balance. Among the missing data imputation studies in solar radiation values (Table 2.5), Zainudin et al. (2015) assessed the utility of one of the commonly adopted machine learning algorithms, KNN, by comparing its performance with three classical techniques (i.e. linear interpolation, spline interpolation, and cubic interpolation). Likewise, Kim et al. (2019) employed not only different conventional techniques (including linear interpolation (LI), mode imputation, and MICE) but also the KNN to complete missing solar radiation data. The authors reported that LI produced optimal results when the proportion of missing data was below 15%. However, as the proportion of missing data increased, the KNN method, producing lower error rates, outperformed its counterparts.

Table 2.5 : Details regarding the studies devoted to impute missing solar radiation data.

ID	Reference	Imputation methods	Method/Data	Gap size	Performance metric
27.	Zainudin et al. (2015)	LI, SI, CI, KNN	Solar Radiation	10,20,30,40,50%	RMSE, R ²
28.	Ogunsola and Song (2014)	SSA, SASR, TBA	Solar Radiation	5,10,15,20,25 days	ME, MAE, RMSE
29.	Žliobaite et al. (2014)	LR	Solar Radiation	1-25%	MSE
30.	Kim et al. (2019)	LI, MI, MICE, KNN	Solar Radiation	10, 15, 20 %	RMSE, MRE, RMSD, MRD
31.	Turrado et al. (2014)	MICE, IDW, MLR	Solar Radiation	0.002-1%	RMSE, MAE
32.	Demirhan and Renwick (2018)	36 imputation methods	Solar Radiation	5,10,25,50%	rMAE, rRMSE, MASE

Still, relevant research has generally focused on classical imputation methods, such as the LR method (Dimas et al, 2011; Turrado et al, 2014; Žliobaite et al, 2014). Demirhan and Renwick (2018) created a comprehensive comparison scheme encompassing 36 models with different variants of interpolation, Kalman filters, persistence, weighted moving average (WMA), and random sample techniques. From a different perspective, Ernst and Gooday (2019) utilized an equation consisting of direct normal

irradiance (DNI) and diffuse horizontal irradiance (DHI) values instead of interpolation methods to complete missing solar radiation data.

2.5.7 Missing data imputation attempts concentrated on wind speed data

Wind speeds impact the rate of air movement over the surface, which in turn affects the rate of water vapour transfer from the surface to the atmosphere. As shown in Table 2.6, the use of machine learning methods predominates in studies related to completing wind speed missing data. For instance, the ANFIS approach was adopted by Hocaoglu et al. (2009) and Zhiling Yang et al. (2011) for Türkiye and China, respectively. To illustrate the effectiveness of classical and modern techniques, Liu et al. (2018) combined the multiple imputation (MPI) and Gaussian process regression (GPR) approaches to perform predictions regarding various datasets containing missing values. In this regard, a plethora of research further validated the superiority of the machine learning rationale for similar purposes, such as back propagation neural network (BPNN) (Basheer Shukur and Hisyam Lee, 2015; Mao and Jian, 2016), generalized regression neural network (GRNN) (Du et al, 2019), and recurrent neural network (RNN) (Li et al, 2019), in the pertinent literature. Similar to the NN-based methods, other machine learning techniques have been proven to be practical tools; for example, Chen et al. (2016) integrated the least-squares support vector machine (LSSVM) model with comprehensive learning particle swarm optimization (CLPSO) to reveal that the data tested with the optimized LSSVM model has better imputation sensitivity and the interpolation results of the missing wind speed data are suitable for the variation of the wind speed data at the adjacent level. de Oliveira et al. (2021) developed machine learning models to predict tropospheric ozone concentrations and completed missing values in the input parameters using the MissForest algorithm. Wind speed, which is one of the parameters in the input set, had varying degrees of missing data in different seasons, with the smallest and largest missing rates being 1.69% and 43.75%, respectively. Prior to imputing missing data, the statistical characterization of the dataset was ensured with PCA, and following the imputation process, further controls showed no significant change in the dataset characteristics, leading the researchers to conclude that the imputation process was successful.

Table 2.6 : Details regarding the studies devoted to impute missing wind speed data.

ID	Reference	Imputation methods	Method/Data	Gap size	Performance metric
35.	Yang et al. (2011)	ANFIS	Wind Speed	N/A	RMSE, R ² , MAE
36.	Liu et al. (2018)	GPR, MI	Wind Speed	5,10,15,20,25,30%	RMSE, MAE
37.	Yang and Ma (2016)	NN	Wind Power	N/A	N/A
38.	Du et al. (2019)	NN	Wind Speed	N/A	MSE, MAPE, MAE
39.	Li et al. (2019)	RNN	Wind Speed	N/A	N/A
40.	He et al. (2013)	Mahalanobis distance	Wind Speed	N/A	N/A
41.	Hocaoglu et al. (2009)	ANFIS	Wind Speed	N/A	RMSE, R ²
42.	Xia and Sun (2017)	Low-rank Matrix Approximation, KNN, MPI, MI	Wind Speed	N/A	RMSE
43.	Chen et al. (2016)	LSSVM	Wind Speed	10%	MAPE, RMSE,
44.	Wesonga (2015)	MICE	Wind Speed	28%	N/A
45.	Lin and Wang (2014)	CEM	Wind Speed	1-12 hour	MBE, MAPE, RMSE, r
46.	Oliveira et al. (2021)	MissForest	Wind Speed	1.69-43.75%	Mean_RMSE

2.5.8 Missing data imputation attempts concentrated on temperature data

Table 2.7 demonstrates the attempts focused on the imputation of missing temperature data in the relevant literature. For instance, Xu et al. (2013) introduced a neoteric approach, called Biased Sentinel Hospitals-based Area Disease Estimation (P-BSHADE) point estimation model, and benchmarked it with the kriging, inverse distance weighting (IDW), and spatial regression testing (SRT). The effectiveness of the technique was empirically evaluated in terms of the annual temperature dataset from 1950 to 2000 in China and the scholars concluded that the P-BSHADE is applicable to the annual Chinese temperature datasets. Among other attempts devoted to imputing missing temperature values, several studies have been carried out using machine learning approaches for different regions across the globe, such as multi-layer perceptron (MLP), radial basis function (RBF), RNN, time-lagged feedforward network (TLFN), time delay recurrent neural network (TDRNN), and counter propagation fuzzy-neural network (CFNN) methods for Canada (Coulibaly and Evora, 2007), the LSTM method in South Korea (Park et al, 2019), and NN and KNN methods in Chile.

Table 2.7 : Details regarding the studies devoted to impute missing temperature data.

ID	Reference	Imputation methods	Method/Data	Gap size	Performance metric
47.	Xu et al. (2013)	P-BSHADE, IDW, SRT	Temperature	N/A	RMSE, MAE
48.	Coulibaly and Evora (2007)	MLP, RBF, RNN, TLFN, TDRNN, CFNN	Temperature	2.1-65.2%	MAE, r
49.	Park et al. (2019)	LSTM	Temperature	6,12,24 hour	RMSE, MBE
50.	López et al. (2021)	NN, KNN	Temperature	N/A	R ²
51.	Thevakaran and Sonnadara (2017)	Between-station technique	Temperature	0.55-1.22%	N/A
52.	Altan and Üstündağ (2012)	Wavelet	Temperature	50%	MSE, MAE
53.	Katipoğlu (2022)	SVM, ANFIS, DT	Temperature	20%	RMSE, MAE, R ²
54.	Shtiliyanova et al. (2017)	Kriging	Temperature	16,17.69%	NSE, MAPE, MAE, BIAS
55.	Cheval et al. (2020)	DINEOF	Land surface temperature	1.1-58.4%	R ²
56.	Kadow et al. (2020)	DL	Temperature	12.5%	P, RMSE
57.	Dumitrescu et al. (2020)	MLR, GAM	Land surface temperature	N/A	MAE, RMSE, r
58.	Zhou et al. (2020)	MLR	GIS/land surface temperature	N/A	RMSE
59.	Williams et al. (2018)	EMD, Bayesian compressive sensing	GIS/temperature	1,5,10,20%	RMSE
60.	Lompar et al. (2019)	Debiasing Procedure	GIS/temperature	5 days	RMSE
61.	Koksal et al. (2016)	ANN	Air temperature, relative humidity	30%	R ² , RMSE, MAE, MAPE

In a study using the temperature data of four neighbouring stations with complete temperature data records pertaining to the same climatological area, Thevakaran and Sonnadara (2018) estimated the missing temperature data at the target station and performed accurate predictions due to the inter-stational correlation. Regarding Europe, Shtiliyanova et al. (2017) applied a kriging-based interpolation in the spatial dimension to fill the data gaps in air temperatures and a data driven machine learning-based approach for the temporal dimension. In addition, Geographic Information System (GIS)-related data plays a pivotal role in estimating ET by providing information on the spatial heterogeneity of environmental factors, including topography, land cover, soil type, and meteorological variables. Moreover, GIS data facilitate the assessment of catchment surface area and characteristics that are

fundamental in ET calculations. The GIS data can also be utilized to evaluate energy balance components of the land surface, which are necessary for determining the energy available for ET. Hence, several efforts have been devoted to acquiring complete records regarding GIS- related datasets (Table 2.7). For instance, to accomplish the imputation of raw images containing between 1.1% and 58.4% incomplete LST data, Cheval et al. (2020) utilized the data interpolating empirical orthogonal functions (DINEOF) algorithm via the R software. A void-filled LST dataset with high spatial resolution was used to investigate LST climatology on a monthly, seasonal, and annual scale over an urban area in Bucharest, Romania. In addition, to reduce uncertainties in climate records, Kadow et al. (2020) proposed a new deep learning approach, while Dumitrescu et al. (2020) used two class regression models (i.e. MLR and generalized additive model [GAM]) characterized by differing complexity and flexibility to impute the missing LST data.

The study conducted by Koksal et al. (2017) focused on estimating incomplete hourly climate data in the semi-humid Bafra plain of northern Türkiye, which plays a significant role in terms of agricultural facilities. They utilized the ANN technique and found that it not only accurately mapped the missing ET data but also estimated the data for areas where the ET was not recorded. Along with the adoption of machine learning algorithms, signal processing approaches have also been employed in the imputation of GIS-related data. For instance, Williams et al. (2018) incorporated two signal processing methods into the missing temperature values attained from remote sensing products. They only used time series with missing data and disregarded the information of nearby stations. The results showed that the BCS-on-Signal method out- performed the bayesian compressive sensing (BCS)-on intrinsic mode functions (IMF) method, highlighting the effectiveness of incorporating signal processing methods to estimate missing temperature data from remote sensing products.

2.6 Discussion

Missing data is a common problem in data analysis, and a plethora of algorithms have been introduced to overcome the corresponding challenge. In addition, the reliability and consistency of the imputation attempts should be diligently examined in various aspects. Hence, this research assessed the 62 existing studies based on three major phases, called pre-imputation (in terms of assumptions, mechanism and pattern

control, and standardization), during-imputation (in terms of software used and the number of iterations), and post- imputation (model run times, statistical distribution checks, time series variance, and validation), that must be taken into account to pinpoint the crucial issues in implementing data imputation strategies and ensuring the accuracy and validity of the imputed datasets. The following subsections briefly discuss the corresponding elements.

2.6.1 Assessment of the pre-imputation stage

The pre-imputation stage is a crucial step in missing data imputation, as the foundation for the entire imputation process is set in this stage. Hence, scholars are advised to evaluate the corresponding step concerning different aspects including the pattern and extent of missing data, and assess the causes of missing data, statistical assumptions and data standardization. Failing to adequately address these issues during the pre-imputation stage can lead to biased and inaccurate imputation results, which can have serious consequences for subsequent analysis and the interpretation of the adopted models. This research assessed the pertinent literature based on whether the existing works concerned the statistical assumptions, checked mechanisms and patterns, and performed the standardization of datasets.

Concerning the statistical assumptions, it was found that the vast majority of studies utilizing the classical linear regression-based imputation techniques lack a proper reporting of statistical assumptions. Specifically, Kunwor et al. (2017) rigorously tested the normality of the data and homogeneity of the variances in imputing their missing EC measurements. Furthermore, Afrifa-Yamoah et al. (2020) suggested the use of non-parametric regression models for the imputation of humidity data. This aspect was further emphasized by Rezvan et al. (2015), who reported that such imputations lack reliability and the direct use of data following imputation may lead to spurious results. Therefore, it is imperative that researchers acknowledge the importance of adhering to appropriate statistical assumptions prior to the missing data imputation in order to ensure the validity and accuracy of subsequent analyses.

The present study also examined the existing studies in terms of the mechanisms of the missing data. It is of utmost importance that researchers possess a prior understanding of the mechanism underlying missing data before imputation procedures. Such knowledge is critical for the development of input sets utilized in the

imputations (Ali et al, 2023; Başakın et al, 2023). Here, incorporating parameters believed to be associated with missingness as auxiliary variables into the model can positively influence imputation results. Various methods exist that enable the statistical testing of the corresponding mechanisms, with some R programming language packages offering readymade solutions (Jamshidian et al, 2014). However, only six of the reviewed studies (Afrifa-Yamoah et al, 2020; Alavi et al, 2006; Demirhan and Renwick, 2018; Kim et al, 2019; Liu et al, 2018; Lo Presti et al, 2010) provided information regarding the mechanism underlying missing data. Similarly, the investigation of missing data patterns is of vital importance in missing data imputation studies. However, only three of the reviewed studies (Goodrich et al, 2021; Lin and Wang, 2014; Lo Presti et al, 2010) provided information regarding the pattern of missing data. It is therefore crucial for researchers to consider the mechanism and pattern underlying missing data before imputation and to incorporate such information in their analyses to ensure accurate and valid results.

The last facet inspected under the umbrella of the pre-imputation stage is the standardization of the parameter values incorporated into the imputations. To define the parameters involved in the estimations, it is essential to standardize the input variables (Hasan et al, 2021; Sovilj et al, 2016). Multiple imputation procedures generally focus on solving a regression problem, and reducing the dataset to a dimensionless and evenly distributed structure (such as [0,1]) is regarded as a routine procedure in almost every regression model in the literature. Despite the vitality of standardization, only limited attempts (Chen et al, 2016; Coulibaly and Evora, 2007; Kim and Ahn, 2009; Lee and Chou, 2004; Mao and Jian, 2016; Schmidt et al, 2008) reported information regarding the standardization of parameters among the investigated articles.

2.6.2 Assessment of the during-imputation stage

The existing literature further recognized the critical role played by the essential points during the imputation of missing data. This paper examined the literature based on two aspects, i.e. the software used and the number of iterations in multiple imputations. Concerning the software utilized, the investigated studies reported the use of various programs to perform their imputation attempts. Among them, it was concluded that the most widely used tool is the R programming language (Cheval et al, 2020; Costa et al,

2021; Demirhan and Renwick, 2018; Körner et al, 2018; Meher and Das, 2019; Wesonga, 2015; Williams et al, 2018; Zitouna-Chebbi et al, 2018). In addition, Schmidt et al. (2008) employed both STATISTICA and MATLAB. Likewise, Miró et al. (2017) utilized MATLAB, whereas many other software programs, such as Neuto-Solution (Abudu et al, 2010), SAS (Alavi et al, 2006; Hui et al, 2004), GeneXPro (Hasanpour Kashani and Dinpashoh, 2012), C# (Altan and Ustundag, 2012), and SPSS (Zainudin et al, 2015), were also adopted to perform the imputation of missing values that exist in the records of the variables related to ET.

Furthermore, in multiple imputation algorithms, the imputation process is repeated for different numbers of iterations. The appropriate number of iterations can vary depending on the size of the dataset, the number of independent variables, and the pattern of missingness. For instance, the default value for the number of iterations in the MICE package is 5. When examining studies that utilized multiple imputations, Hui et al. (2004), Turrado et al. (2014), Xie and Sun (2018), Kim et al. (2019), and Costa et al. (2021) selected the iteration value as “ $m = 5$.” However, Miró et al. (2017) chose a value of 20. When considering datasets with varying sizes, it may be beneficial to explore different combinations of iteration values.

2.6.3 Assessment of the post-imputation stage

The evaluation of results obtained regarding imputation processes should essentially be carried out based on several aspects. Various methods can be used to evaluate the imputed data, including comparison with the observed data and comparison of the imputed data with the data from a complete case analysis. Furthermore, visual inspection of the imputed data can be helpful in detecting any systematic patterns or outliers that may indicate problems with the imputation model. Along with these essential aspects, this review further underpinned the role of run time evaluations, statistical distribution checks, assessing time series variance, and validation of the obtained results. Concerning the first aspect, one can conclude that the run time of imputation algorithms may vary depending on the size of the dataset as well as the utilized algorithms and computer capabilities. However, it is well known that the classic methods typically have shorter run times compared to modern techniques such as machine learning algorithms. Here, machine learning methods may require longer run times due to the optimization of hyperparameters. This can pose a challenge for

researchers who need to make quick decisions in time-sensitive situations. Therefore, sharing information regarding the run times of different imputation techniques is crucial to provide a better overview for interested researchers in similar domains. Among the papers investigated within the scope of the current research, only three studies (Körner et al, 2018; Lin and Wang, 2014; Turrado et al, 2014) have comparatively reported the computational costs of the utilized algorithms.

This study also assessed the previous work based on statistical distribution checks. Scholars who perform imputation attempts typically aim to minimize error regardless of the chosen method. Error levels can be measured through various approaches, providing important quantitative results regarding the imputation success. However, these quantitative metrics do not necessarily account for the distributional properties of the imputed values, which ideally should align with the distribution of the original dataset (Zhao et al, 2022). Here, the existing literature highlighted that testing the similarity between distributions can be achieved through statistical tests (Özger et al, 2020). Therefore, if the imputed values do not correspond to the dataset distribution, the imputation procedure should be reviewed and/or reoperated in order to achieve accurate imputation outcomes. We found only eight studies (Demirhan and Renwick, 2018; He et al, 2013; Hui et al, 2004; Körner et al, 2018; Liu et al, 2018; Lo Presti et al, 2010; Wesonga, 2015) that have investigated the similarity between imputed values and dataset distributions, highlighting the need for more research in this area.

Furthermore, when performing imputations for time series data, only the values of the parameter under investigation can be used for the imputation process. Therefore, it is crucial to ensure that the variance of the time series remains the same after imputations (Gao et al, 2018). Among 62 investigated papers, only Anjomshoaa and Salmanzadeh (2019) and Kunwor et al. (2017) checked the variance of the time series. The first used a quantitative metric to measure the variance values of imputed and gapped time series, while the latter provided a detailed analysis of the variance values after the imputation process. In order to reliably impute missing data, another important step that should diligently be taken into account is to conduct a training and validation process using the complete parts of the dataset used (Aghili et al, 2022). This process serves to demonstrate the accuracy of the imputation model and allows for the conclusion that the completed values are close to real values.

Using the entire dataset to obtain the imputation model without subjecting the models to any validation process, i.e. using the entire dataset for training, can lead to overfitting and result in completed values that are underestimated and/or overestimated. To tackle the overfitting phenomenon, cross-validation is one of the most commonly used methods in the pertinent literature (Eum and Yoo, 2022; Yang et al, 2021). Of the studies examined within the scope of current research, 25% of the focalized research considered embracing a validation rationale, while others did not report the differentiation of training and validation stages.

The assessment of whether missing data imputation algorithms achieve optimal performance can be determined by utilizing various statistical evaluation metrics. These metrics are used to assess the degree of agreement between the predicted values and the actual values. Some of these metrics can be used to express the adequacy of a single model/algorithm quantitatively, while others are utilized to compare the results of multiple models (Zouzou and Citakoglu, 2022). In this sense, metrics such as mean square error (MSE), root mean square error (RMSE), mean absolute error (MAE), and mean bias (MB) are used to measure which of two or more models makes predictions with lower errors, while some other metrics (such as coefficient of determination [R^2], Nash-Sutcliffe efficiency [NSE], and Kling-Gupta efficiency [KGE]) measure the performance of a single model. Metrics that measure individual model performance typically range between $[-1,1]$ or $[0,1]$ and are able to measure the goodness of fit independently of the unit used (Bayram and Çıtakoğlu, 2023). Hence, in evaluating the performance of these algorithms, it is crucial to diligently consider the choice of metric and its interpretability, as well as the potential implications of the missing data imputation process for subsequent statistical analyses and modelling outcomes.

It is worth mentioning that nearly 20% of the studies examined do not provide any information regarding the metric used for determining imputation accuracy, while 25% of them paid regard to only one metric. On the other hand, nearly half of the pertinent literature considered two or more metrics in assessing the accuracy of the imputation attempts. Diligent investigation demonstrated that the most commonly used metric is RMSE, with 34 occurrences, followed by MAE and R^2 with 17 and 16 occurrences, respectively. One can conclude that using different performance evaluation indicators helps increase the reliability of the obtained imputation results due to the various data structures representing divergent data-gathering processes.

2.7 Conclusions

Missing data is a common problem in meteorological data for many reasons, including instrument and equipment mal- function, errors in data collection, natural events, and lack of resources. In addition, the existence of missing data poses serious challenges for many modelling approaches in the hydrology domain, especially considering the importance of having complete datasets that allow for accurate and reliable analysis of water-related phenomena. Hence, the present research carried out an investigation of the existing literature with reference to calculating ET, which is one of the significant components of the hydrologic cycle. Among nearly 20,000 studies directly or indirectly related to the ET, a total of 62 studies were extracted to draw a comprehensive representation regarding the efforts devoted in the last two decades. As a result, the following conclusions and implications are drawn:

- The applicability and validity of the classical/parametric imputation methods rely heavily on various assumptions; hence, it is essential to approach these methods with caution and rigorously assess their underlying assumptions to avoid inaccurate conclusions or biased results.
- Assessment of the distribution of data is essential prior to commencing imputation procedures, while optimizing the run time of the imputation model and finetuning the model parameters allow efficient performance to be achieved in the during-imputation stage. Furthermore, within the post-imputation step, it is important to verify the stability of the imputed values; hence, assessing the constancy of variance and similarity of the distributions can help identify potential issues and provide insight into the robustness of the imputation model.
- Although various conventional techniques have been employed to handle missing data due to their ease of implementation, advanced imputation techniques that are particularly leveraged by the introduction of machine learning algorithms have emerged as more reliable alternatives to address missing data.
- Although attempts regarding the incorporation of deep learning techniques are limited thus far, these techniques show great potential for enhancing the performance of imputation models, particularly in datasets with a large number of observations.
- Interpretability of machine learning algorithms is an emerging concept in recent years; however, there was no study exploring the comprehensibility of missing

data imputation procedures. Hence, integration of explainable artificial intelligence (e.g. SHapley Additive exPlanations [SHAP], Local Interpretable Model Agnostic Explanations [LIME], etc.) can enhance the interpretability and transparency of the imputation results.

Furthermore, it is especially worth mentioning that, despite numerous studies examined within the scope of the current research, Scopus was the only database used to extract the relevant studies that exist in the literature. However, subsequent attempts could include different research databases to extend the scope of missing data imputation studies by the inclusion of conference papers, reports, and dissertations. Overall, a thorough investigation of the ET-related literature and the conclusions made in the current study are expected to assist the pertinent research community in understanding the dynamics of missing data imputation strategies and provide valuable insights regarding decision making processes for downstream analyses.

3. DEVELOPING A NOVEL APPROACH FOR MISSING DATA IMPUTATION OF SOLAR RADIATION: A HYBRID DIFFERENTIAL EVOLUTION ALGORITHM BASED EXTREME GRADIENT BOOSTING MODEL²

3.1 Introduction

The use of coal and petroleum products to generate energy results in substantial carbon emissions into the environment. In 2021 alone, the world experienced a total of 36.3 gigatons of carbon emissions from energy-related activities. These emissions are contributing to the greenhouse effect, driving the acceleration of global warming and leading to observable climate change impacts (Gao et al, 2022; Lu et al, 2022). Hence, in recent years, there has been a shift towards clean and renewable energy sources to meet energy needs (Ang et al, 2022). Among clean energy sources, solar energy is one of the sources with the highest energy generation potential, and thereby, receives significant investment (Kabir et al, 2018). The share of energy generated from solar energy sources has dramatically increased from 61.85 TWh in 2011 to 1021 TWh in 2021. In line with this increase, it is important to identify the most efficient areas (such as urban areas, building roofs, rural areas, etc.) to leverage the potential of solar energy. Hence, accurate solar radiation assessments, including determining the correct solar radiation values for the area where the energy facility is planned to be installed, are crucial in the identification and feasibility studies (Marzo et al, 2017). These assessments provide valuable information on the amount of solar energy that can be captured by solar panels and are used to estimate the size and capacity of the solar energy system required to meet the energy needs of a specific site, as well as to determine the expected energy yield and performance of the system over time (Ozcan and Ersoz, 2019). Furthermore, accurate approximations ensure that the solar energy system is optimally sized, maximizing its efficiency and financial viability. The data

²This chapter is based on the paper “Başakın, E. E., Ekmekcioğlu, Ö., and Özger, M. (2023). Developing a novel approach for missing data imputation of solar radiation: A hybrid differential evolution algorithm based eXtreme gradient boosting model, *Energy Conversion and Management*, 280, 116780. doi: 10.1016/j.enconman.2023.116780.”

collected from these assessments can be used to prioritize investment in renewable energy infrastructure in regions with high solar energy potential, and to support the development of renewable energy policies and planning at the national and international levels.

There are different devices fostered for continuously measuring solar radiation, such as pyranometers (Cao et al, 2022) and pyrheliometers (Vignola et al, 2016), while pyranometers are the one that most widely used. These devices directly measure the solar radiation from the sun. In this regard, several challenges are encountered during the measurement of solar radiation as possibly experienced in the measurement of other meteorological variables, i.e., temperature, precipitation, etc., and atmospheric, i.e., humidity, pressure, etc. (Narvaez et al, 2021). Here, lack of regular calibration of devices, objects covering the measuring device due to inadequate maintenance, or contamination of the device over time poses several problems including inaccurate data transmission (Schwandt et al, 2014). In addition, there may not only be erroneous data transmissions but also missing values originating from human factors (Saaban et al, 2016). Considering all these facts, the existence of missing data is of great importance for research society working in the data analysis field and many methods have been introduced to cope with potential challenges that emerged from the missing values (Moradi, 2009). In a similar vein, the deployment of energy systems with high initial investment costs necessitates the utilization of comprehensive measurements to accurately determine the energy potential and to minimize estimation errors. The imputation of missing values in measurements with a high degree of accuracy is therefore pivotal for determining the energy potential (Mohamad et al, 2022). This approach will minimize errors in the analyses and result in more accurate cost projections helping stakeholders make informed decisions regarding the development of the farm. To exemplify, a 100 Joule calculation error in a 20 million m² solar farm, where the average daily solar radiation values are 15 MJ/day/m², can lead to an erroneous calculation of 2 billion Joules per day.

The first step in missing data analysis is to identify the pattern (Gravesteyn et al, 2021) and mechanism (Iodice D'Enza et al, 2022) of missing data; such that there is a great deal of patterns and mechanisms leading to augment the complexity of the analysis. In this vein, it is vital to identify whether the missing data is randomly distributed or existed due to a parameter-related factor. Mechanisms of missing data are mainly

relied upon the three mechanisms, i.e., missing completely at random (MCAR), missing not at random (MNAR), and missing at random (MAR) (Haliduola et al, 2022). Based on the MCAR approach, the missing data occur completely randomly behaving regardless of any independent parameter, while the missing data matching the MNAR and MAR mechanisms represent data that is reduced by the effect of one or more independent variables. Furthermore, there are various patterns indicating the distribution of missing data, such as patterns addressing the fact that data is missing randomly or patterns capitalizing the missing data existing by following a chain. These patterns are called univariate and multivariate, monotone and non-monotone, and general patterns (Ben Aissia et al, 2017).

The second crucial part of the missing data analysis is to decide whether to remove the data from the data set or preserve it. Some of the pertinent literature argued that missing data can be removed from the main dataset once the amount of missing or incorrect data accounted for 5 percent or less of the entire dataset (Strike et al, 2001). On the other hand, missing data imputation strategies can be performed in order to fill the gaps that exist in the dataset through embowring the nature of the original dataset (Camastra et al, 2022). The existing body of knowledge covers many methodologies to impute the missing values. The traditional methods embrace the strategies that are grounded on the descriptive statistics of the original dataset, such as mean, mod, and median (Zhao et al, 2022). Despite these strategies being more practical compared to the complex methods, missing data imputation via statistical parameters may lead to the monotony of the completed missing spaces which conduce toward inaccurate assignments (Yozgatligil et al, 2013). Moreover, the research community has further put forward the well-known statistical approach, i.e., regression, to impute the missing values (Bashir and Hua-Liang Wei, 2015). Through the regression, the missing values can be assigned using the regression curve if the missing data is randomly distributed in the original dataset. However, regression models tend to yield satisfactory accuracy once the sample size is large enough. Also, a single regression curve is followed for all imputed values and natural variance is not represented in the dataset. In this sense, the regression model estimates the most probable value of the missing data; however, it does not provide adequate information regarding the uncertainty of the imputed samples (Osman et al, 2018). Along with the traditional methods, more convenient strategies are presented in the pertinent literature for missing data imputation, such as

hot deck imputation (HDI) (Little and Rubin, 2002). The HDI is described as a method that is based on handling missing data in which each missing value is replaced with an observed response from a “similar” unit. In practice, the HDI scans the entire dataset to round the similar units to the missing values up and determines the observed value of one of the candidates, creating a complete data set instead of missing data samples. Different from the regression model, the HDI constitutes the missing value imputation using the inherent nature of the original dataset rather than employing a statistical representation manner (Andridge and Little, 2010). Therefore, the HDI is defined as an implicit strategy and regarded as robust against misidentification and interaction effects since the model does not require an explicit model from which imputations are generated.

The main motivation for the missing data imputation implementations is to specify a method that is not only as close as possible to the original data but also requires the least theoretical assumptions (Gao et al, 2018). Since most of the classical algorithms, i.e., regression analysis, principal component analysis, expectation maximization, etc., have significant prerequisites (Hron et al, 2010) that need to be satisfied, recently, investigators have examined the modern methods (such as machine learning algorithms) that are capable of producing accurate estimations without statistical restrictions and/or assumptions. To date, several researches comprising modern missing data imputation techniques’ applications have been proposed in handling meteorological phenomena (Sattari et al, 2020; Wang and Shi, 2021). In general, machine learning (ML) algorithms are classified into four groups as supervised, semi-supervised, unsupervised, and reinforcement (Bagherzadeh and Asil, 2019). Among them, supervised learning algorithms relying upon the estimation of dependent variable(s) by means of the independent variable(s) are the most frequently used tools. These algorithms work through the principle of introducing pre-defined samples to the system for training at first, and then, the imputation of missing values is performed in this regard. Studies over the past two decades have provided vital insights regarding the fact that modern modeling techniques produce more accurate estimations compared to traditional imputation methods (Gautam and Ravi, 2015; Silva-Ramírez et al, 2015; Zhang et al, 2015). Despite the practicality and convenience of ML strategies revealed in the existing literature, there are still limited attempts for the imputation of missing values in solar radiation datasets. Such that, García Acevedo et

al. (2016) utilized the neural networks, while Zainudin et al. (2015) employed the k-nearest neighbor (KNN) algorithm to accurately model the missing values in solar radiation. Scholars have long debated the effectiveness of the classical techniques on solar radiation's missing values. Accordingly, the linear regression method has widely been adopted by several researchers (Dimas et al, 2011; Turrado et al, 2014; Žliobaitė et al, 2014). For instance, Mohammadi and Moazenzadeh (2021) employed multiple linear regression to estimate the solar radiation values using various meteorological variables, including minimum, average, and maximum temperatures, as well as relative humidity and precipitation. They further benchmarked their analysis results with the traditionally adopted empirical equations' findings and highlight the superiority of the proposed regression-based model. In addition, Ogunsola and Song (2014) compared the performance of three different methods, including singular spectrum analysis (SSA), statistically adjusted solar radiation (SASR), and temperature-based approach (TBA), to impute the missing solar radiation instances. They chose two different study domains, i.e., Oklahoma City North and Albuquerque, for 2012 and 2005, respectively, and imputed the missing solar radiation instances artificially generated up to 25 days. According to their results, SSA is the best-performed model to impute solar radiation values for up to five days, while SSA and TBA yielded more accurate estimations for up to 20 days, and finally, for more than 20 days, TBA produced the most prominent results. In addition, Demirhan and Renwick (2018) constituted a total of 36 models to carry out missing data imputation for solar radiation measurements using the different combinations of Interpolation, Kalman Filters, Persistence, Weighted Moving Average, and Random Sample. Recently, Moazenzadeh et al. (2022) integrated the support vector machine with the nature-inspired Cuckoo search optimization algorithm to conduct solar radiation predictions. They used two different datasets, in which one includes the meteorological variables (such as maximum, minimum, dew point, wet, and dry temperatures, relative humidity, actual sunshine hours, maximum possible sunshine hours, cloud cover, and extraterrestrial radiation), while the other covers geographical inputs (such as day of the year, day of the month, number of months, as well as station's elevation, longitude, and latitude). Based on comprehensive analysis, they concluded that the model constructed using the meteorological variables outperformed its counterpart.

Recent developments in computer applications have heightened the implementation of ML techniques in all research domains. Accordingly, a considerable amount of algorithms using neural networks-based (i.e., artificial neural networks), distance-based (i.e., KNN), statistic-based (i.e., Naïve Bayes), and tree-based (i.e., random forest, adaptive boosting, gradient boosting, etc.) approaches have been introduced in the pertinent literature. In recent years, there is a growing body of literature that recognizes the utility of tree-based approaches in various research domains. Most of the tree-based algorithms are regarded as white-box models elaborating the interpretability of the obtained results distinct from the black-box models, such as ANNs and support vector machines. However, the exhaustive literature survey performed within the scope of this study revealed that the applicability of the tree-based algorithms has not been questioned yet for the imputation of missing values that exist in solar radiation measurements. Therefore, this study adopted one of the recently proposed robust tree-based ML techniques, i.e., eXtreme Gradient Boosting (XGBoost), to model missing values in solar radiation measurements made in stations located in one of the largest basins of Türkiye, the Konya closed basin. Despite the XGBoost method has successfully been used in most of the meteorology-related fields, including pollutant concentration (Li et al, 2022), air pollution (Dai et al, 2022), and predicting actual vapor pressure (Qiu et al, 2022), this study is the initial attempt of its utilization for solar radiation missing data imputations. In addition, in ML applications, correct adjustment of the hyperparameters of the utilized algorithms is another vital issue since they may have drastic impacts on the attained results. The relevant literature introduced several techniques to tune the hyperparameters of the ML algorithms. For instance, Feng et al. (2021) and Bendavid et al. (2022) considered applying traditional grid search and random search approaches to configure the XGBoost, respectively. However, these classical hyperparameter tuning approaches are described as quite time-consuming since they follow exhaustive scans within the entire candidates existing in the search space (Abbaszadeh et al, 2022). In addition, increasing the step size to overcome the computational burden may potentially lead to a decrease in the performance of the optimizations. To cope with the above-mentioned challenges, meta-heuristic optimization algorithms have become extremely useful tools, especially in light of the recent developments in computer science. To exemplify, different meta-heuristics, i.e., particle swarm optimization (Dias Júnior et al, 2021; Gu et al, 2021), genetic algorithm (Liu and Sun, 2023; Pan et al, 2022), artificial bee colony

optimization (Luo et al, 2021; Wang et al, 2022), and ant colony optimization (Zhao et al, 2022), are hybridized with the XGBoost algorithm not only increase the robustness of the model but also stabilize the attained predictions.

Overall, given the vitality of the missing data imputations having a substantial impact on the determination of solar energy potential, this study proposed a comprehensive comparison scheme by incorporating not only conventional interpolation techniques but also modern machine learning algorithms into missing value imputations for solar radiation measurements. The major objectives and corresponding contributions of the current research are provided as follows:

- The machine learning algorithms were employed for imputing missing solar radiation values as an initial attempt in the literature; such that the vast majority of the existing attempts are based on the utilization of traditional techniques. This research also enhances the robustness of the predictions by incorporating the meta-heuristic differential evolution into the extreme gradient boosting algorithm.
- The imputations regarding the missing solar radiation values were conducted on a daily resolution, in which the present literature heavily relies on the minute or hourly resolutions. In addition, the validity of the proposed models was assessed based on four different gap sizes, i.e., 5%, 10%, 20%, and 30%, with taking the MCAR mechanism into account. Limited studies have introduced such a detailed framework.
- In terms of the utilized datasets to perform missing data imputations, this research recognized the role of both other meteorological variables measured in the target station and the solar radiation measurements attained through the neighboring stations. Such a comprehensive analysis has not been acknowledged in the literature.

3.2 Study Area And Data

This study was conducted for imputing the missing values that exist in solar radiation measurements performed in Konya closed basin extending between 36°51' N and 39°29' N latitudes and between 31°36' E and 34°52' E longitudes. The basin is surrounded by Sakarya and Kizilirmak from the north, Kizilirmak and Seyhan from the east, the Eastern Mediterranean from the south, and Antalya and Akarcay Basins from the west. The total area of the river basin, which covers approximately 6.4% of Türkiye's surface area, is 49,786 km² (Figure 3.1). A flat plain in the Konya closed

basin covers the vast majority of the basin and forms the main part of the Central Anatolian Plateau. This plain, called the Konya Plain, is surrounded by volcanic mountains and the mountains surrounding the plain prevent the water from draining into the sea from the Basin. For this reason, this basin is characterized as a closed basin that could not discharge its waters into the sea making it among a few examples across the globe. Precipitations are generally observed in the winter and spring seasons; however, Konya closed basin is one of the regions receiving the least precipitation across Türkiye due to its geographical location (Orhan et al, 2020). In addition, semi-arid climate characteristics are dominant in the basin and there has been a decreasing trend in the precipitation regimes leading to 1.2 -11.6 mm subsidence over the last three decades (Ekercin et al, 2013). Since the area of the basin is large, the climate types are diverse. Southern and western parts of the basin are dominated by Csa and Csb climate classes based on the Köppen-Geiger climate classification. Northern and eastern parts of the basin are under Csb and BSk climates, respectively. The central part of the basin mainly covering the city of Konya is characterized by the BSk class which reflects hot and dry summers and cold and snowy winters (Canan et al, 2020). According to the long-term averages, the annual temperature is 11.1 °C across Konya closed basin, and the coldest month is January and the hottest month is July with -0.3 °C and 23.4 °C, respectively.

Furthermore, brief statistics regarding the meteorological measurements conducted in Aksaray station are presented in Table 3.1. What stands out in the table is that solar radiation is the parameter that shows the closest characteristics to the normal distribution according to the skewness statistics. This study also examined the solar radiation measurements conducted in the meteorological stations located in the vicinity of Aksaray station. Here, Table 3.2 provides the descriptive statistics of the solar radiation measurements attained from the Nevsehir, Nigde, Kulu, and Karatay stations. According to Table 3.2, Nigde is the station having the highest values in terms of maximum daily total solar radiation, and the lowest minimum values are observed in Nigde and Kulu stations. In addition, the average total solar radiation values take similar values at all stations. When the skewness coefficients are examined, one can conclude that the solar radiation values measured in all neighboring stations are close to the normal distribution characteristics but the values of the Nevsehir station that showed perfect consistency with the normal distribution with a skewness of 0.02192.

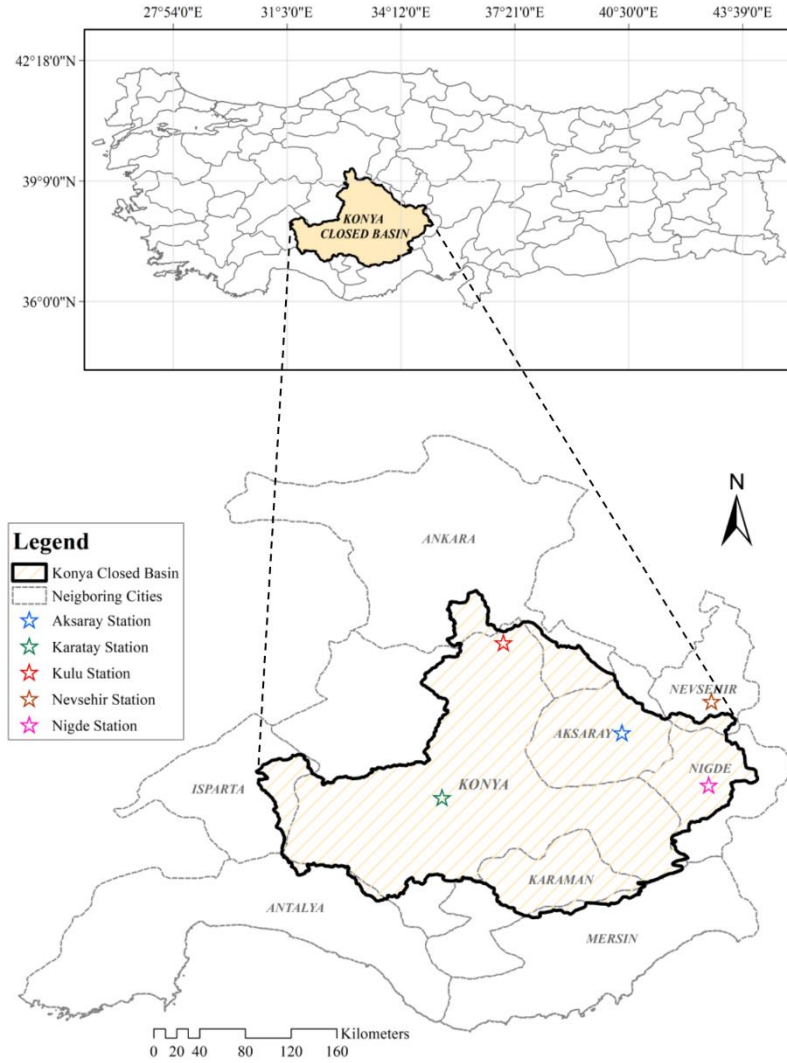


Figure 3.1 : Study area.

Table 3.1 : The descriptive statistics of the dataset acquired from Aksaray station.

Parameters	Unit	Minimum	Mean	Maximum	Standard Deviation	Skewness
Maximum Temperature	°C	-5.5	20.28	38.7	9.83	-0.32
Minimum Temperature	°C	-16.5	8.09	24.8	7.76	-0.23
Average Temperature	°C	-11.1	13.91	31	8.91	-0.21
Maximum Relative Humidity	%	24	71.12	96	15.95	-0.41
Minimum Relative Humidity	%	5	29.91	91	16.44	0.98
Average Relative Humidity	%	14.1	50.6	94.9	17.42	0.26
Sunshine Duration	Hour	0	7.31	14	4.1	-0.31
Wind Sped	m/s	0.4	1.75	4.9	0.69	1.04
Total Solar Radiation	Watt/m ² /day	6000	282429	532578	144874	-0.07

Table 3.2 : The descriptive statistics of the solar radiation data acquired from neighboring stations.

Stations	Unit	Maximum	Minimum	Mean	Standard Deviation	Skewness
Nevsehir	Watt/m ² /day	522000	3600	266742	143376	0.02192
Nigde		565263	600	292953	128432	-0.15122
Kulu		535217	600	280240	136034	-0.10381
Karatay		536505	1920	288117	132806	-0.12998

3.3 Research Framework

The present research seeks to assess the applicability of various missing data imputation strategies for the gaps possibly encountered in the meteorological stations. To accomplish the overarching aim of this study, three scenarios (namely Scenario 1, Scenario 2, and Scenario 3) were taken into account. The accuracy of the proposed models is benchmarked with the solar radiation values measured in Aksaray station that locates in Konya closed basin, Türkiye. Hence, synthetic gaps were created randomly in accordance with the MCAR at four different gap sizes, i.e., 5%, 10%, 20%, and 30%, in order to manifest the robustness and reliability of the missing data imputation processes. Within the scope of Scenario 1, only the time series of measured solar radiation values are analyzed. Missing values were imputed by means of the interpolation techniques for all gap sizes. The first scenario results in Model 1 and Model 2 which contain the implementation of linear interpolation and spline interpolation techniques, respectively. Furthermore, Scenario 2 and Scenario 3 cover the incorporation of the machine learning algorithms into the missing data imputation strategies. For Scenario 2, the meteorological parameters that are influential on solar radiation values and measured at the Aksaray station were used to fulfill the imputation of randomly generated gaps with machine learning models. To establish the input-output selection procedure, Pearson correlation coefficients (R) between all meteorological parameters and solar radiation were computed (Figure 3.2). Accordingly, the models belonging to Scenario 2 were specified based on the inclusion of the parameter from having the highest correlation with solar radiation to one that has the lowest correlation, respectively. Table 3.3 underpins that the sunshine duration (N) having a correlation of 0.86 constitute the only input variable to estimate missing solar radiation values (Model 3), whereas the last model (called Model 10) within Scenario 2 was formed with the inclusion of maximum temperature (Tmax), average

temperature (Tavg), minimum relative humidity (RHmin), minimum temperature (Tmin), average relative humidity (RHavg), maximum relative humidity (RHmax), and wind speed (WS), respectively, according to the correlations in descending order. Also, relevant literature covering studies that attempted to model the solar radiation values considered similar meteorological variables, i.e., Tmin, Tavg, and Tmax (He et al, 2020), and (RHmin), (RHavg), (RHmax) (Bounoua et al, 2021), as well as N (Zhang et al, 2021) and WS (He et al, 2020).

Table 3.3 : Model combinations constituted within Scenario 2.

Models		Input Combination						Output
Model 3	N							Aksaray Stations's Solar Radiation
Model 4	N	T _{max}						
Model 5	N	T _{max}	T _{avg}					
Model 6	N	T _{max}	T _{avg}	RH _{min}				
Model 7	N	T _{max}	T _{avg}	RH _{min}	T _{min}			
Model 8	N	T _{max}	T _{avg}	RH _{min}	T _{min}	RH _{avg}		
Model 9	N	T _{max}	T _{avg}	RH _{min}	T _{min}	RH _{avg}	RH _{max}	
Model 10	N	T _{max}	T _{avg}	RH _{min}	T _{min}	RH _{avg}	RH _{max}	WS

Correlation Matrix

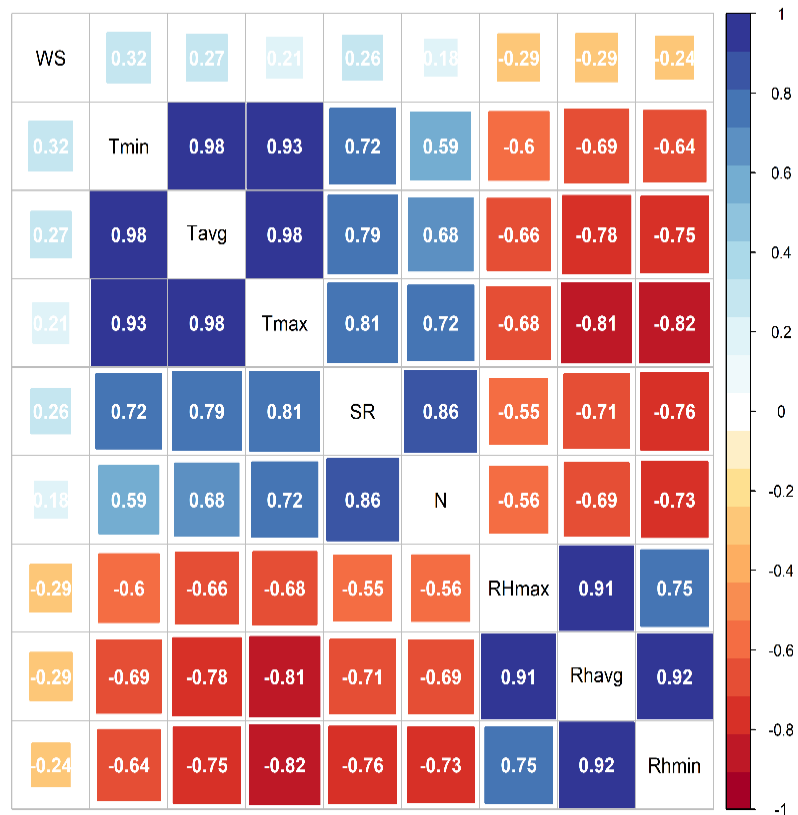


Figure 3.2 : Pearson correlation matrix of the meteorological parameters and solar radiation measurements.

For Scenario 3, solar radiation values that were simultaneously measured in the stations (i.e., Karatay, Kulu, Nevsehir, and Nigde) located in the vicinity of Aksaray station in order for performing missing data imputations. The corresponding scenario also covers the implementation of machine learning algorithms (namely XGBoost-DE, MARS, and RF). Similar to Scenario 2, Pearson correlation coefficients were taken into account in determining the input-output combinations in Scenario 3. Hence, R values between the neighboring stations' solar radiation measurements and those measured in the Aksaray stations were computed (Figure 3.3). As seen in Figure 3.3, the stations with the strongest relationship with the solar radiation values of Aksaray were Nevsehir, Karatay, Kulu, and Nigde stations, respectively. Considering the corresponding relationships, Table 3.4 presents the input combinations. Nevsehir's solar radiation values were only used in Model 11 as it showed the strongest correlations, while the remaining stations' (i.e., Karatay, Kulu, and Nigde) data were appended into the input-output combinations based on the correlations in a descending order to generate Model 12, Model 13, and Model 14, respectively.

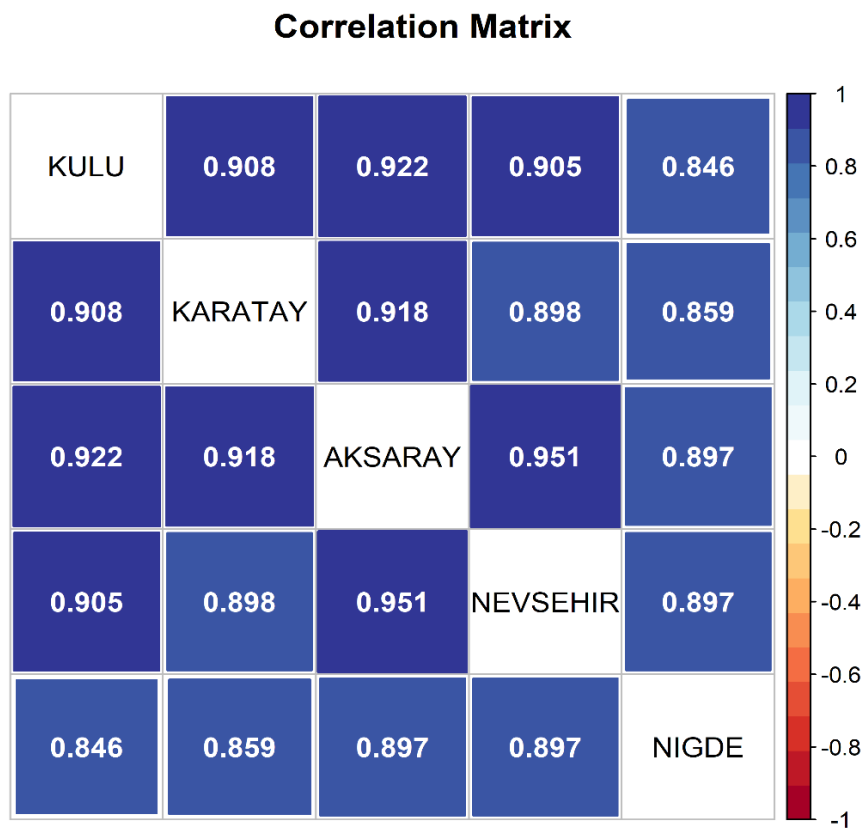
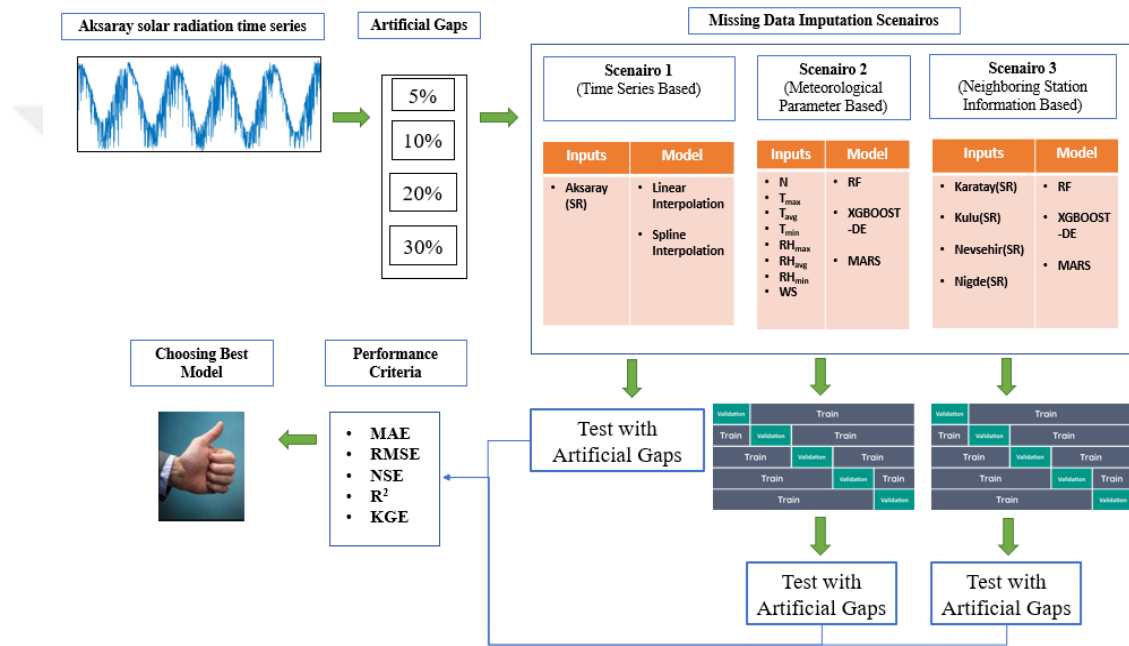


Figure 3.3 : Pearson correlation matrix of the neighboring stations' solar radiation measurements and Aksaray station's.

Table 3.4 : Model combinations constituted within Scenario 3.

Models	Input Combination				Output
Model 11	Nevsehir SR				Aksaray Stations's Solar Radiation
Model 12	Nevsehir SR	Karatay SR			
Model 13	Nevsehir SR	Karatay SR	Kulu SR		
Model 14	Nevsehir SR	Karatay SR	Kulu SR	Nigde SR	

Overall, the analysis steps followed for achieving the study objective are provided in Figure 3.4.

**Figure 3.4 :** Flowchart of the current research.

3.4 Methods

In this study, two different traditional missing data imputation strategies, namely linear and spline interpolations, and three different ML methods, i.e., XGBoost, MARS, and RF, were utilized to perform a comprehensive comparison. It is especially worth mentioning that this study chiefly design to propose the applicability of the XGBoost model, while the MARS and RF were employed to benchmark the attained predictions. In addition, state-of-the-art differential evolution was adopted to the XGBoost algorithm to accurately and effectively tune its hyperparameters. Therefore, this section includes a brief explanation of interpolation methods (Section 3.4.1) followed

by the introduction of XGBoost (Section 3.4.2), MARS (Section 3.4.3), RF (Section 3.4.4), and the DE (Section 3.4.5), as well as the performance evaluation indicators (Section 3.4.5), respectively.

3.4.1 Interpolation methods

Interpolation methods are described through the identification of the function, i.e., $f(x)$, corresponding to an intermediate value of x based on the list of independent variables specified as x_i ($i = 0, 1, 2, \dots, N$) (Luo and Lu, 2022). Interpolation techniques further form the basis of various methodologies such as numerical derivative and integration, as well as the numerical solution of ordinary and partial differential equations. One can employ the interpolation methods by simply fitting curve(s) with respect to the data instances designated as x_i, f_i . These functions are defined as interpolation functions. Additionally, polynomials regarding miscellaneous orders are often used as interpolation functions. However, in some cases, more specific functions such as logarithmic, exponential, hyperbolic, and trigonometric functions are taken into account for periodic data values. Finite difference-based interpolation methods are generally adopted once the data points are evenly dispersed, while different interpolation techniques, i.e., linear interpolation, Lagrangian interpolation, etc., are regarded as more appropriate for vice versa. Instead of solving a set of linear equations to interpolate by assigning a polynomial function with respect to a set of points, the relevant literature recognizes the essentialness of appropriately transforming the acquired formulas into suitable forms for direct interpolations. In this study, missing solar radiation data imputation was accomplished by means of the linear and spline interpolation strategies via R packages called “imputeTS” library existing in the R software.

3.4.2 eXtreme gradient boosting (XGBoost)

XGBoost is one of the commonly used tree-based ML methods that adopt the gradient boosting rationale to reach more precise approaches for generating the optimum predictive model (Chen and Guestrin, 2016). The most prevailing feature of the XGBoost is its computational efficiency compared to other boosting algorithms, such as adaptive boosting, gradient boosting, and stochastic gradient boosting. Decision trees are trained sequentially based on the pre-defined training dataset and the algorithms adjoin a newly generated decision tree to the previous decision trees at each

iteration to improve the accuracy of the model with regard to the objective function. The objective function that needs to be minimized is composed of the loss function and regularization term. The loss function controls the accuracy of the predictive model based on training data, while the regularization term is used as a penalty factor in order to reduce the complexity of the constructed model. The XGBoost further implements the second-order Taylor expansion to approximate the loss function value. In addition, XGBoost ensures performing the cross-validation which is a commonly applied technique to acquire the error of the validation set minimizing the over-fitting of the model at each iteration of the boosting process (Serencam et al, 2022).

3.4.3 Multivariate adaptive regression splines

The Multivariate Adaptive Regression Splines (MARS) algorithm which is a form of non-parametric regression analysis was initially proposed by (Friedman, 1991). The MARS is one of the practical tools for adapting non-linear multivariate functions to specific problems as it avoids assumptions regarding any functional relationship among the dependent and independent variables. The model is constituted with regard to the base functions and their associated coefficients within the boundaries of data availability and accessibility. In addition, one of the strengths of the MARS algorithms comes from the fact that it estimates the contributions of the base functions constructed both over the explanatory variables and their allowable interactions to determine the response variable. The MARS is mainly grounded on two different algorithmic concepts, i.e., forward stage and backward stage (Friedman, 1991). Since the model obtained with the forward stage has an inherently complex structure, the optimum model is reached by extinguishing the redundant base functions through the backward stage, also called the pruning stage.

3.4.4 Random forest

The Random Forest (RF) contains many decision trees and is based on collecting the predictions of the corresponding decision trees into different sub-samples of datasets (Breiman, 2001). Accordingly, the RF is a combination of tree estimators such that each tree is sampled independently and depends on the values of a random vector with the same distribution for all trees in the forest. A large number of decision tree structures are constituted through the new data samples derived from the original data set by means of the bagging method. In order to create decision trees, the classification

and regression trees (CART) algorithm is utilized. While generating these decision trees, a multifarious number of randomly selected variables are used via the random subspace method. Thus, the best-branched variable in each leaf node is determined through the random sub-space technique. The RF introduces the prediction results based on considering the individual results of the decision trees; such that the overall results are acquired by taking the majority voting into account for classification problems, while for the regression problems, the overall estimations are refined based on the average predictions of the decision trees.

3.4.5 Differential evolution

Differential Evolution (DE) which is an evolutionary-based meta-heuristic was first proposed by Storn and Price (1997) to handle complex optimization problems. The nature of the DE is similar to the genetic algorithm (GA); such that as in the GA application, DE is based on several evolutionary operators, i.e., cross-over, mutation, selection, and reproduction. Although the cross-over, mutation, and selection operators are adopted in the DE, each operator is not applied sequentially to the entire population. Thus, chromosomes are handled one by one, and a new individual is obtained by using the other three randomly selected chromosomes. The new chromosome and the current chromosome are compared in accordance with the objective function, and the selected new individual is transferred to the next generation. This process is pursued by repeating these steps until the pre-determined stopping condition is satisfied to reach the global optimum. Hence, the mutation operator improves the performance of the algorithm and provides new solution values in the search space. At last, the chromosomes are compared according to their fitness values and for the minimization problem, the chromosome with the smaller fitness value is transferred to the next population. Mutation, crossover, and selection steps are repeated until the stopping criterion is satisfied. Owing to its several advantages, the DE algorithm was coupled with several widely adopted machine learning tools, such as ANFIS (Emamgholizadeh et al, 2023), ANN (Singh et al, 2023), SVM (Panahi et al, 2022), and RF (Bazzaz Abkenar et al, 2021).

3.4.6 Performance evaluations

The current research took several performance metrics into account to assess the accuracy of the models adopted for the missing data imputations. Hence, mean

absolute error (MAE), root mean square error (RMSE), Nash–Sutcliffe efficiency index (NSE) (Nash and Sutcliffe, 1970), coefficient of determination (R^2), and Kling-Gupta efficiency index (KGE) (Gupta et al, 2009) were utilized. Table 3.5 illustrates the minimum, maximum, and best values of the performance metrics as well as their corresponding equations.

Table 3.5 : Details of the performance indicators considered in the models' evaluation.

Performance indicators	Abbreviation	Lowest	Highest	Best Value	Equation
Root Mean Square Error	RMSE	0	∞	0	$\sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$
Determination Coefficient	R^2	0	1	1	$1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$
Mean Absolute Error	MAE	0	∞	0	$\frac{1}{N} \sum_{i=1}^N y_i - \bar{y} $
Nash–Sutcliffe Efficiency	NSE	$-\infty$	1	1	$1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y}_i)^2}$
Kling-Gupta Efficiency	KGE	$-\infty$	1	1	$1 - \sqrt{(r - 1)^2 + (\alpha - 1)^2} + \beta$

Note: Variables given in the equations are; y_i = actual values, $\hat{y}_i$ = imputed values, $\bar{y}_i$ = average of actual values, $\bar{\hat{y}}_i$ = average of imputed values, n = number of observations, r is the linear correlation between observations and simulations, α a measure of the flow variability error, and β a bias term.

It is worth mentioning that the lower the MAE and RMSE, the higher the accuracy of the models, while the increase in the NSE and KGE towards 1 reveals the increase in model performance. R^2 refers to the proportion of the variation in the dependent variable(s) that is predictable from the independent variable(s). Once the R^2 takes a value of 1, the corresponding model perfectly estimates the actual targets.

3.5 Results

3.5.1 Interpolation-based results (scenario 1)

In this study, linear and spline interpolation rationales were used as interpolation techniques to impute the missing solar radiation values within the scope of Scenario 1. Interpolations were conducted through the “imputeTS” package in the R software.

Imputation performances attained with the interpolation analysis for four different gap sizes (i.e., 5%, 10%, 20%, and 30%) are given in Table 3.6. The results clearly demonstrate that the linear interpolation technique (i.e., Model 1) has achieved more accurate imputation results in all gap sizes compared to its counterpart (i.e., Model 2). In accordance, NSE values of 0.783, 0.820, 0.820, and 0.796 were obtained with Model 1 once the gap sizes are at 5%, 10%, 20%, and 30%, respectively, while Model 2 yields 0.705, 0.745, 0.708, and 0.683 of NSE values for the corresponding gap sizes. Similar conclusions can further be made with respect to other performance indicators as presented in Table 3.6.

Table 3.6 : Test performances of Interpolation-Based models.

Model	Performance Indicator	Gap Size (%)			
		5	10	20	30
Linear (Model 1)	MAE*	45205	41735	44349	45583
	RMSE*	65002	60484	63945	66134
	NSE	0.783	0.820	0.802	0.796
	R ²	0.787	0.824	0.809	0.799
	KGE	0.874	0.904	0.897	0.883
Spline (Model 2)	MAE*	52013	49430	53749	55649
	RMSE*	75797	72005	77660	82461
	NSE	0.705	0.745	0.708	0.683
	R ²	0.730	0.769	0.742	0.715
	KGE	0.852	0.872	0.853	0.844

The bold values represent the results of the best-performed model.

* The units are in Watt/m²/day.

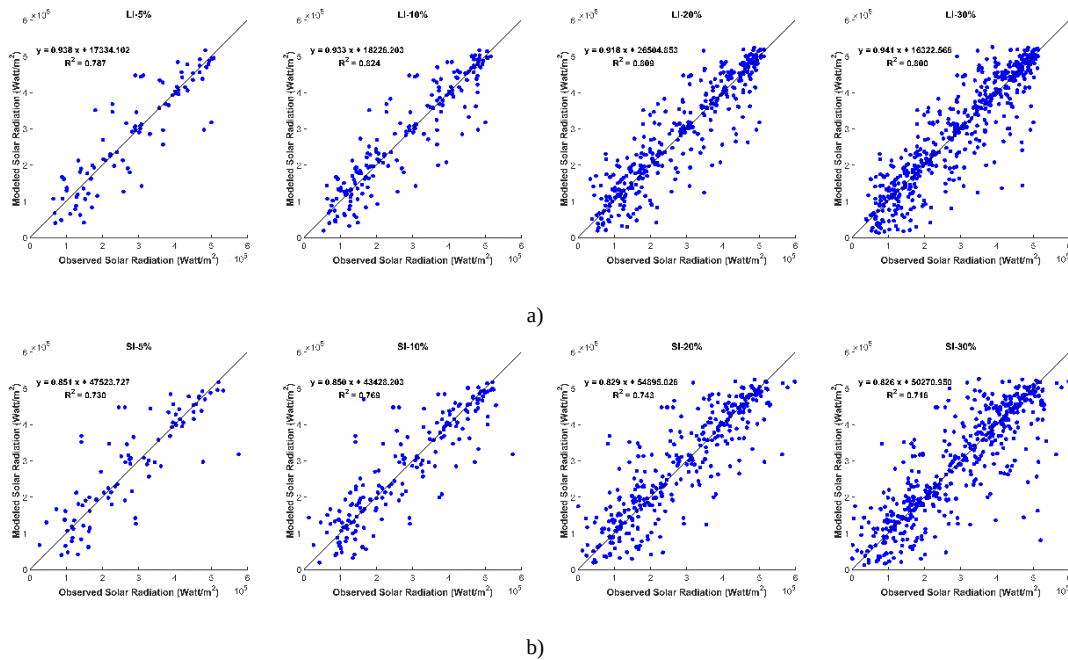


Figure 3.5 : Scatters plots of interpolation-based missing data imputations for the testing set a. Model 1, b. Model 2.

For a visual interpretation of the missing data imputation performances of the interpolation-based models, scatter plots are given in Figure 3.5. The graphs are quite revealing in several ways; such that the imputations carried out with the linear interpolation are more scattered on the perfect model line than those obtained by the spline interpolation technique. As can be seen from the figure that there is a plethora of points diverging from the 1:1 perfect model line making the spline interpolation method's imputed solar radiation values within the test sets quite biased.

3.5.2 Machine learning-based results

This research further embraces Scenario 2 and Scenario 3 for evaluating the applicability of machine learning algorithms to impute the solar radiation missing values. Associatively, we used three different machine learning algorithms, i.e., random forest (RF), multi-variate regression splines (MARS), and eXtreme Gradient Boosting (XGBoost) to holistically draw the performance comparison frame. For constituting the machine learning-based analysis, the pre-determined gap sizes to be desired to impute were considered test sets, while the remaining parts of the dataset were used as training and validation sets. Note that the 10-fold cross-validation, which is the most widely adopted technique, was used in the training sets and the hyperparameter tuning processes were fulfilled via the grid search algorithm for RF and MARS methods, whereas the differential evolution (DE) searching algorithm was used for the XGBoost. As a result of the optimizations, the “number of trees (n)” and the “optimum number of variables (mtry)” for the RF algorithm, and the “number of pruning (nprune)” and “maximum degree of interaction (ndegree)” values were identified for the MARS algorithm. Hyperparameter values vary recursively for each training set in all models implemented in Scenario 2 and Scenario 3. After the optimum values are determined based on training and validation sets, missing value imputations were carried out with respect to blinded testing sets in order to conduct fair performance assessments. It is worth mentioning that the DE algorithm also contains several hyperparameters that need to be diligently determined (Emamgholizadeh et al, 2023). Thus, this study selected the maximum number of iterations as 100, population size as 20, and cross-over probability as 0.3, while the lower and upper bound of the scaling factor were identified as 0.2 and 0.8, respectively.

3.5.3 Scenario 2 results

As mentioned earlier, Scenario 2 covers the implementation of machine learning algorithms (i.e., RF, MARS, and XGBoost) to model missing solar radiation values in Aksaray station using other meteorological variables (i.e., N, Tmax, Tavg, RHmin, Tmin, RHavg, RHmax, and WS) measured in the same stations. The performance results regarding the testing set are provided in Table 3.7. Table 3.7 compares all input-output alternatives ranging from Model 3 to Model 10 simulated with three machine learning methods under four different gap size conditions. According to the performance indicators attained for Model 3, in which only sunshine duration (N) is considered input, the XGBoost optimized with the DE was found as the best-performed model (0.750, 0.762, 0.759, and 0.747 for 5%, 10%, 20%, and 30%, respectively), followed by the MARS and RF algorithms.

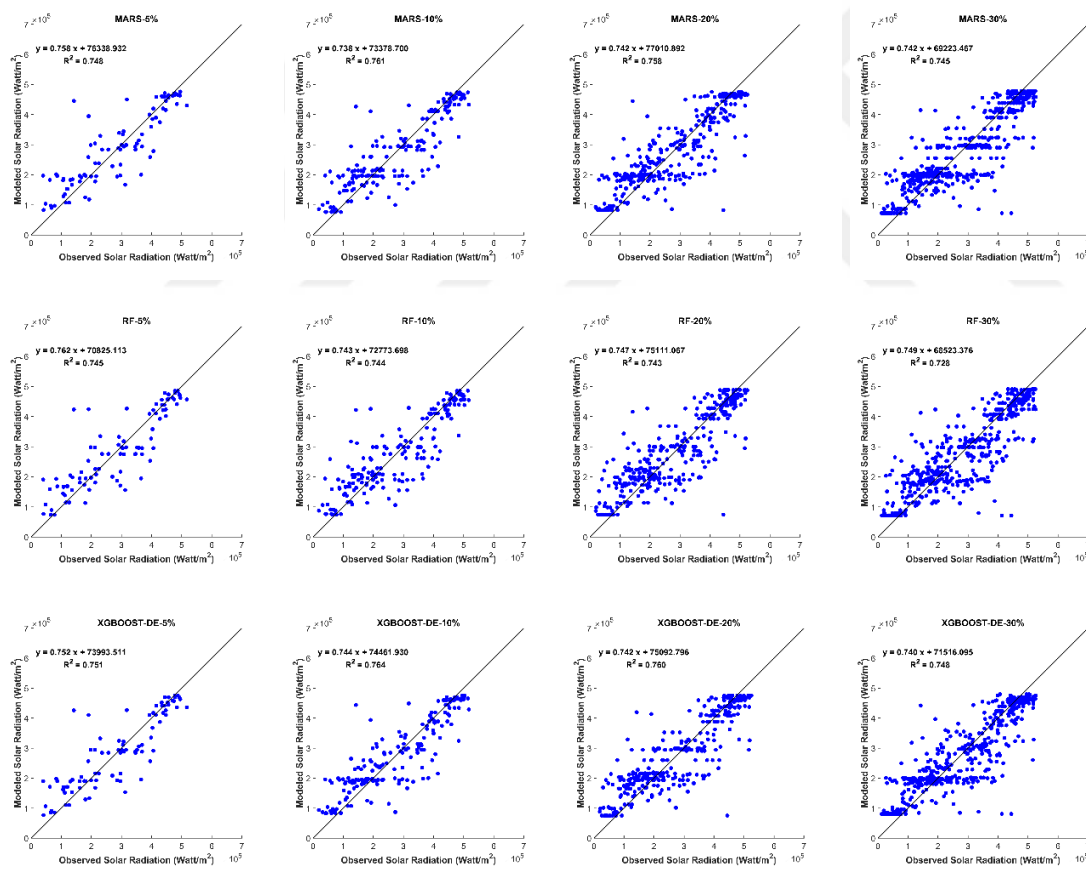


Figure 3.6 : Scatters plots of missing data imputations for the testing set with Model. 3.

The scatter diagrams regarding the most parsimonious model containing only one input are presented in Figure 3.6. Once the maximum temperature (Tmax) was

appended into the input variables (i.e., Model 4), the hybrid XGBoost-DE was the most accurate model for the first three gap size alternatives (NSE of 0.803, 0.826 ve 0.828 for 5%, 10%, and 20%, respectively). However, the RF outperformed its counterparts with an NSE of 0.841 for the largest gap size (30%) within Model 4. The overall results regarding Model 4 underpinned that the inclusion of Tmax increased the model performances in each gap size. Likewise, considering the average temperature (Tavg) as an additional input (Model 5) has yielded a slight increase in imputation performances; such that the NSE of 0.828, 0.833, 0.842, and 0.845 were obtained through the XGBoost-DE for 5%, 10%, 20%, and 30%, respectively. Similar to the Tavg, the inclusion of the minimum relative humidity (RHmin) depicted in Model 6 has augmented the accuracy of the machine learning models. These results are in accord with recent studies indicating that the temperature has a significant impact on acquiring high accuracies in solar radiation modelling attempts (Bounoua et al, 2021; He et al, 2020).

What is striking about the results is that considering the minimum temperature (Tmin) did not have a significant impact on the modelling accuracies, once the Model 7 results were examined (Table 3.7). On the other hand, appending the average relative humidity (RHavg) into the input variables within Model 8 has enabled significant improvements in the XGBoost-DE model's accuracies resulting in NSE values of 0.866, 0.844, 0.851, and 0.862 in each gap size alternative, respectively. Similar to the T min but contrary to the RHavg, the inclusion of RHmax (Model 9) did not enhance the accuracy of the proposed predictive framework. This result can be explained by the fact that RHmax was one of the variables having lower correlations with the solar radiation values. This conclusion further is in agreement with Tao et al. (2019)'s finding, in which the authors followed similar input selection steps (i.e., according to the Pearson correlation) and found that the inclusion of the RHmax did not make any improvement in model predictions. Finally, once the wind speed (WS) is incorporated into the input set through Model 10, a slight improvement was observed in the imputation performances with respect to each performance indicator. Still, the best results within the scope of Scenario 2 were obtained via Model 10 which comprised all meteorological inputs to estimate the missing solar radiation values under each gap size. The corresponding result corroborates with the existing literature, in a way that using all inputs to model solar radiation values yields the most accurate outcomes. For

instance, Moazen-zadeh et al. (2022) included various meteorological-related variables that help predict the solar radiation values and concluded with the best predictions. Based on the entire results, scatter plots derived for Model 10 are presented in Figure 3.7. On the one hand, closer inspection of the figure highlights that the MARS and XGBoost-DE methods exhibit a symmetrical scattering on the perfect (1:1) model line at all gap sizes. On the other hand, one can observe that the RF method produced over-estimated results at 5% and 10% gap sizes, while it slightly underestimated the missing solar radiation values at 20% and 30% gap sizes (Figure 3.7). Overall, according to the comparison of scatter diagrams presented for the models having only one input (Figure 3.6) and the one containing all meteorological variables (Figure 3.7) within the scope of Scenario 2, one can conclude that the imputation performance transparently enhances with respect to the increasing number of inputs.

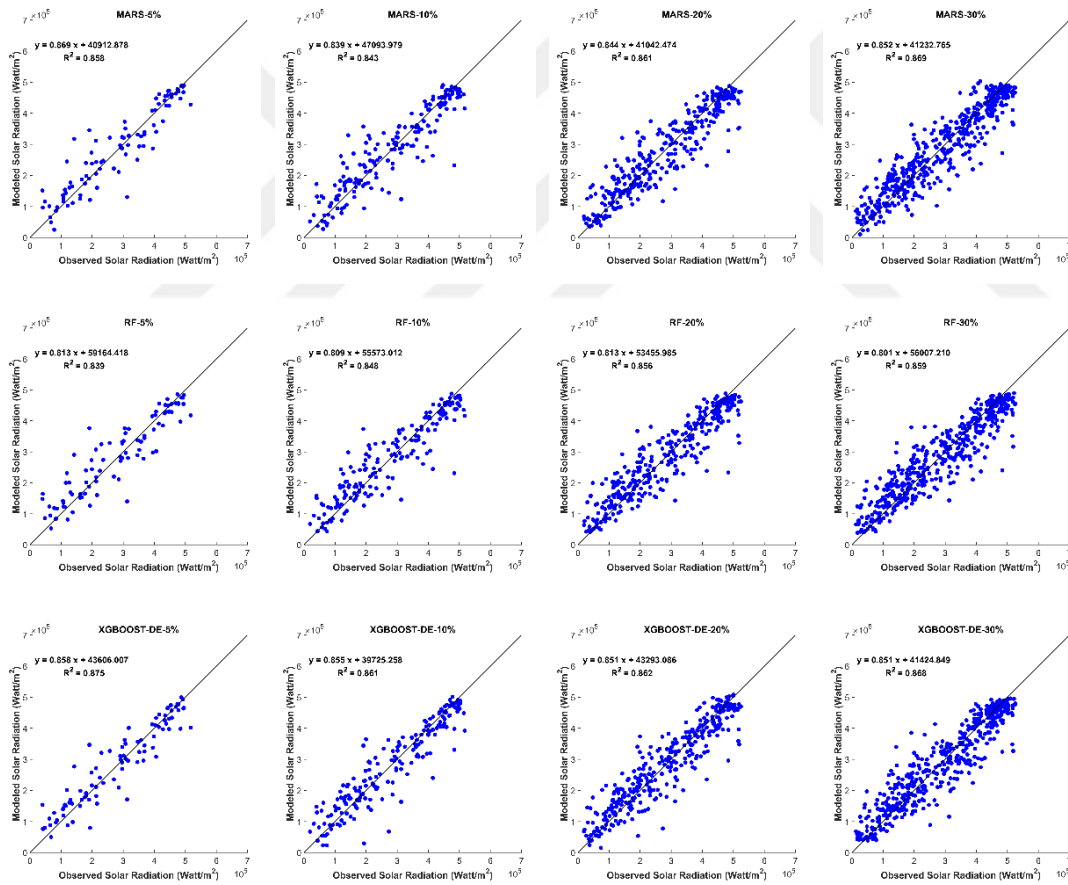


Figure 3.7 : Scatters plots of missing data imputations for the testing set with Model 10.

Table 3.7 : Test performances of the models within Scenario 2.

Model	Algorithm	Performance Indicator	Gap Size (%)				Model	Algorithm	Performance Indicator	Gap Size (%)			
			5	10	20	30				5	10	20	30
Model 3	XGBOOST-DE	MAE	49350	50593	51016	52578	Model 7	XGBOOST-DE	MAE	41823	41971	42220	42926
		RMSE	69880	69447	70504	73743			RMSE	56228	59570	56233	56691
		NSE	0.750	0.762	0.759	0.747			NSE	0.838	0.825	0.847	0.850
		R ²	0.750	0.763	0.760	0.747			R ²	0.840	0.825	0.847	0.851
	MARS	KGE	0.811	0.804	0.803	0.801		MARS	KGE	0.875	0.869	0.884	0.873
		MAE	49934	50823	51260	52889			MAE	42887	41657	41624	43179
		RMSE	70602	69777	70958	74167			RMSE	58409	55957	56518	57819
		NSE	0.744	0.760	0.756	0.744			NSE	0.825	0.846	0.845	0.844
	RF	R ²	0.748	0.761	0.757	0.745		RF	R ²	0.827	0.846	0.845	0.845
		KGE	0.814	0.799	0.803	0.803			KGE	0.879	0.873	0.875	0.871
		MAE	49540	52502	53262	54527			MAE	45062	43739	43192	44154
		RMSE	70774	72101	73019	76650			RMSE	60429	59688	57541	58497
		NSE	0.743	0.744	0.742	0.726			NSE	0.813	0.824	0.839	0.840
		R ²	0.744	0.744	0.742	0.728			R ²	0.816	0.826	0.841	0.844
		KGE	0.819	0.804	0.807	0.808			KGE	0.842	0.844	0.852	0.837
Model 4	XGBOOST-DE	MAE	43113	42873	44015	43856	Model 8	XGBOOST-DE	MAE	36901	40167	40898	41226
		RMSE	61947	59451	59591	59018			RMSE	51020	56271	55360	54395
		NSE	0.803	0.826	0.828	0.838			NSE	0.866	0.844	0.851	0.862
		R ²	0.805	0.826	0.828	0.839			R ²	0.867	0.844	0.851	0.862
	MARS	KGE	0.861	0.861	0.8686	0.854		MARS	KGE	0.902	0.889	0.896	0.886
		MAE	46384	43671	43785	43531			MAE	38477	40366	42772	41774
		RMSE	64040	61111	58929	59073			RMSE	54284	58175	58499	56721
		NSE	0.790	0.816	0.832	0.837			NSE	0.849	0.833	0.834	0.850
	RF	R ²	0.794	0.816	0.832	0.838		RF	R ²	0.849	0.833	0.834	0.851
		KGE	0.849	0.863	0.863	0.862			KGE	0.894	0.883	0.884	0.874
		MAE	6412	1056	1666	1014			MAE	42889	41581	41028	41806
		RMSE	62223	59458	59866	58394			RMSE	58591	57874	55836	55980
		NSE	0.801	0.826	0.826	0.841			NSE	0.824	0.835	0.849	0.854
		R ²	0.804	0.826	0.827	0.842			R ²	0.826	0.836	0.850	0.856
		KGE	0.871	0.870	0.864	0.860			KGE	0.863	0.863	0.869	0.856

Table 3.7 (continued) : Test performances of the models within Scenario 2.

Model	Algorithm	Performance Indicator	Gap Size (%)				Model	Algorithm	Performance Indicator	Gap Size (%)			
			5	10	20	30				5	10	20	30
Model 5	XGBOOST-DE	MAE	41216	41816	42729	42365	Model 9	XGBOOST-DE	MAE	37476	38553	39765	40097
		RMSE	57884	58182	57137	57582			RMSE	51866	55604	53450	53695
		NSE	0.828	0.833	0.842	0.845			NSE	0.862	0.848	0.861	0.865
		R ²	0.830	0.833	0.843	0.846			R ²	0.863	0.848	0.862	0.866
	MARS	KGE	0.883	0.873	0.863	0.871		MARS	KGE	0.890	0.887	0.890	0.885
		MAE	44451	43008	42483	43203			MAE	41284	41656	40984	41185
		RMSE	63299	60243	57408	58697			RMSE	59106	59002	54538	55128
		NSE	0.794	0.821	0.840	0.839			NSE	0.821	0.828	0.856	0.858
	RF	R ²	0.799	0.821	0.841	0.840		RF	R ²	0.824	0.828	0.857	0.859
		KGE	0.861	0.869	0.877	0.858			KGE	0.896	0.874	0.877	0.874
		MAE	44656	42748	43770	43453			MAE	42049	40651	40862	42183
		RMSE	61690	59053	59271	58440			RMSE	58073	56487	55727	56309
		NSE	0.805	0.828	0.830	0.841			NSE	0.827	0.843	0.849	0.852
		R ²	0.807	0.828	0.830	0.842			R ²	0.830	0.844	0.851	0.855
		KGE	0.862	0.862	0.874	0.852			KGE	0.860	0.862	0.861	0.847
Model 6	XGBOOST-DE	MAE	41123	41309	42642	42743	Model 10	XGBOOST-DE	MAE	36346	39040	39765	39574
		RMSE	55115	55689	56383	56680			RMSE	49528	53214	53450	53271
		NSE	0.844	0.847	0.846	0.850			NSE	0.874	0.860	0.861	0.868
		R ²	0.845	0.847	0.846	0.851			R ²	0.875	0.860	0.862	0.868
	MARS	KGE	0.889	0.875	0.879	0.865		MARS	KGE	0.893	0.893	0.890	0.889
		MAE	41370	41137	42600	42897			MAE	37436	40282	40031	40368
		RMSE	58213	56801	56504	57658			RMSE	52899	56519	53813	53248
		NSE	0.826	0.841	0.845	0.845			NSE	0.856	0.842	0.860	0.868
	RF	R ²	0.829	0.842	0.845	0.845		RF	R ²	0.857	0.843	0.860	0.868
		KGE	0.876	0.864	0.878	0.872			KGE	0.90285	0.88081	0.8841	0.890
		MAE	45912	43605	42738	43602			MAE	41790	40556	40555	41987
		RMSE	61193	59270	57201	57802			RMSE	56525	55910	54933	55819
		NSE	0.808	0.827	0.841	0.844			NSE	0.836	0.846	0.854	0.855
		R ²	0.810	0.828	0.843	0.847			R ²	0.839	0.848	0.856	0.859
		KGE	0.847	0.851	0.859	0.845			KGE	0.857	0.854	0.857	0.845

3.5.4 Scenario 3 results

The analysis performed within the scope of Scenario 3 covers the application of machine learning algorithms for the imputation of missing solar radiation values in the Aksaray meteorological station using the solar radiation values of the closest four stations in terms of air distance. To establish the order of models (Model 11, Model 12, Model 13, and Model 14) in Scenario 3, correlations between the solar radiation values measured in Aksaray and those attained from the other four stations were considered. Hence, the models are ordered in accordance with the Pearson correlation from the highest correlated station (i.e., Nevsehir) to the lowest correlated station (i.e., Nigde), resulting in Model 11 and Model 14, respectively.

The performance results regarding the testing set are provided in Table 3.8. Table 3.8 compares all input-output alternatives within Scenario 3 ranging from Model 11 to Model 14 simulated with three machine learning methods under four different gap size conditions. According to the performance indicators attained for Model 11, which regarded only solar radiation values of Nigde as an input, the hybrid XGBoost-DE model outperformed its counterparts (i.e., RF and MARS) with NSE values of 0.901, 0.894, 0.910 and 0.904 at 5%, 10%, 20%, and 30% gap sizes, respectively (Figure 3.8). Once the solar radiation values of 2nd highly correlated station (i.e., Karatay) were incorporated into the input variables (Model 12), the hybrid XGBoost-DE was the most accurate model for the three gap size alternatives (NSE of 0.913, 0.929 ve 0.925 for 10%, 20%, and 30%, respectively). However, the MARS outperformed its counterparts with an NSE of 0.923 for the smallest gap size (5%) within Model 12. The corresponding results underline that the inclusion of the Karatay station's solar radiation values has a positive impact on model accuracies. The reason for this increase can be attributed to the fact that the stations are positioned symmetrically at the southeast, and northwest ends of the Aksaray station. In terms of reflecting the latitude effect, one can argue that these two stations (i.e., target: Aksaray and independent variable: Karatay) obviate each other's deficiencies. Furthermore, different from previous models, Model 13 which contains the solar radiation measurements of Kulu station in addition to those taken from Aksaray and Karatay yielded varying testing performance with respect to the utilized machine learning methods. Such that, the RF algorithm was found as the best-performed model (NSE: 0.937) once the gap size is

5% and the hybrid XGBoost-DE had the highest accuracies at 10% and 30% gap sizes with NSE values of 0.925 and 0.937, respectively, whereas the MARS model outperformed (NSE: 0.933) its counterparts considering the gap size as 30% (Table 3.8). One can further conclude that the inclusion of Kulu station's solar radiation data did not significantly augment the imputation performances, in which this conclusion is probably associated with the fact that Kulu station is located far north of Aksaray station. Ultimately, once all the solar radiation measurements made in the stations that locate in close vicinity of Aksaray station are incorporated into input space (Model 14), the highest accuracies were acquired. It is worth mentioning that the run time of the corresponding analysis took 235 seconds using the workstation configuration of the AMD Ryzen 9 5900X 12-Core Processor (Win-10, 3.70 GHz, and 64 GB RAM).

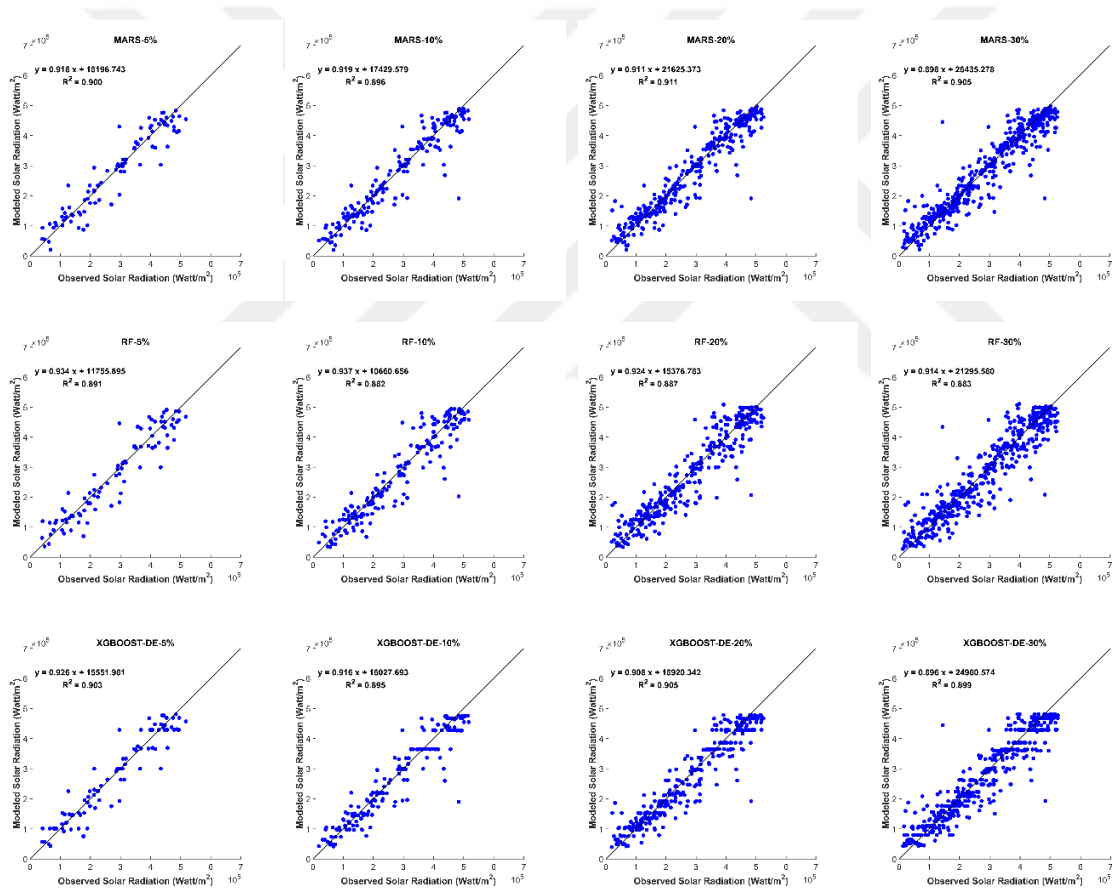


Figure 3.8 : Scatters plots of missing data imputations for the testing set with Model 11.

This performance can be regarded as fair since all the data taken from close meteorological stations contributed to the predictions. Among Model 14, the hybrid XGBoost-DE model provided the most successful imputation modelling at all gap sizes but the 20% which is obtained by the RF. Given the increase in accuracies, the

XGBoost-DE yielded NSE values of 0.950, 0.934, and 0.939 at 5%, 10%, and 30%, respectively, and the NSE was found as 0.944 for the RF at the gap size of 20%. Overall, scatter plots derived for Model 14 are presented in Figure 3.9. The figure is quite revealing that the number of extreme errors is considerably low and almost every model exhibits a symmetrical scattering on the perfect (1:1) model line at all gap sizes. One can also deduce from the scatter plots that the MARS model clearly overestimated the missing solar radiation values at 20% gap size, while all the machine learning methods utilized in the current study produced extreme error values, especially between 4 and 6 Watt/m² solar radiation regardless of the gap size considered.

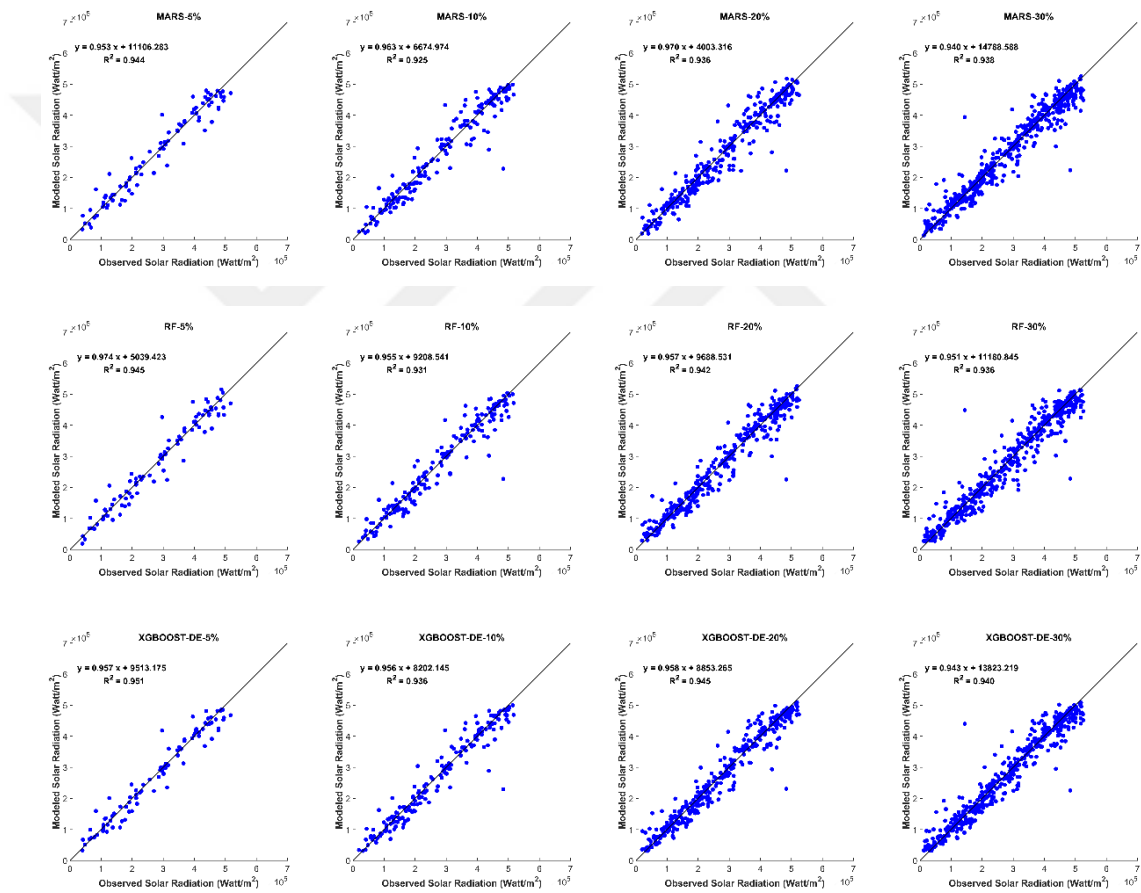


Figure 3.9 : Scatters plots of missing data imputations for the testing set with Model 14.

The reason for these results may be related to the challenges arising from data structures. Hence, it would be possible to perform not only more efficiently but also enhanced imputation processes by filtering these values in a way that does not result in an information loss. Overall, according to the comparison of scatter diagrams presented for the models having only the information of one station, i.e., Nevsehir,

(Figure 3.8) and the one containing the information of all neighboring stations (Figure 3.9) within the scope of Scenario 3, the imputation performance of the adopted models significantly increased based on the introduction of additional stations' solar radiation values.

Table 3.8 : Test performances of the models within Scenario 3.

Model	Algorithm	Performance Indicators	Gap Size (%)				Model	Algorithm	Performance Indicators	Gap Size (%)			
			5	10	20	30				5	10	20	30
Model 11	XGBOOST-DE	MAE	31818	29861	28735	29954	Model 13	XGBOOST-DE	MAE	26140	24545	26015	24928
		RMSE	43970	46370	42991	45384			RMSE	36072	38932	38771	37842
		NSE	0.901	0.894	0.910	0.904			NSE	0.933	0.925	0.927	0.933
		R ²	0.903	0.896	0.911	0.904			R ²	0.933	0.926	0.928	0.933
		KGE	0.940	0.936	0.934	0.924			KGE	0.961	0.960	0.961	0.956
	MARS	MAE	32361	31343	30966	32138		MARS	MAE	25752	26191	23740	26636
		RMSE	44680	46892	44843	46915			RMSE	36606	40374	36011	39953
		NSE	0.897	0.891	0.902	0.897			NSE	0.931	0.919	0.937	0.925
		R ²	0.899	0.895	0.905	0.898			R ²	0.933	0.922	0.937	0.926
		KGE	0.936	0.931	0.928	0.922			KGE	0.964	0.958	0.965	0.954
	RF	MAE	34651	34537	34821	35783		RF	MAE	24619	24730	24498	25511
		RMSE	47024	50120	49088	50550			RMSE	35002	38825	36869	38611
		NSE	0.886	0.876	0.883	0.881			NSE	0.937	0.925	0.934	0.930
		R ²	0.891	0.882	0.886	0.882			R ²	0.938	0.927	0.935	0.931
		KGE	0.938	0.934	0.935	0.932			KGE	0.966	0.961	0.964	0.955
Model 12	XGBOOST-DE	MAE	28174	27583	25560	27055	Model 14	XGBOOST-DE	MAE	21073	22739	23558	23171
		RMSE	39227	41849	38330	39962			RMSE	31229	36472	34918	35928
		NSE	0.921	0.913	0.929	0.925			NSE	0.950	0.934	0.941	0.939
		R ²	0.922	0.916	0.929	0.926			R ²	0.950	0.935	0.941	0.940
		KGE	0.959	0.953	0.959	0.948			KGE	0.967	0.962	0.966	0.957
	MARS	MAE	27463	28246	27157	27942		MARS	MAE	23591	23726	24407	24283
		RMSE	38637	42513	40669	42328			RMSE	33153	37805	36933	36520
		NSE	0.923	0.911	0.920	0.916			NSE	0.943	0.929	0.934	0.937
		R ²	0.924	0.914	0.921	0.917			R ²	0.944	0.930	0.936	0.938
		KGE	0.957	0.948	0.954	0.944			KGE	0.964	0.961	0.963	0.955
	RF	MAE	27183	27655	27656	28589		RF	MAE	22972	25633	22607	24421
		RMSE	38849	42733	40821	42258			RMSE	33270	39534	33959	37404
		NSE	0.922	0.910	0.919	0.916			NSE	0.943	0.923	0.944	0.934
		R ²	0.924	0.913	0.921	0.917			R ²	0.944	0.925	0.944	0.935
		KGE	0.959	0.950	0.953	0.944			KGE	0.970	0.959	0.966	0.961

3.5.5 Overall results

Comprehensive performance comparisons are made with the Taylor Diagram in order to shed light on the research results and to evaluate all models from a wider perspective. Therefore, Figure 3.10, Figure 3.11, Figure 3.12, and Figure 3.13 are generated for assessing models' performances conjunctly for gap sizes of 5%, 10%, 20%, and 30%, respectively. The Taylor diagram graphically summarizes how closely a set of prediction patterns match the observations (Taylor, 2001). The similarity

between the prediction outcomes and actual targets is measured in terms of their correlations I , RMSEs, and standard deviations. Therefore, the Taylor diagram provides a promising visual tool due to its ability to “graphically” display the prediction performances of different models (Zeynolabedin et al, 2021). In a typical Taylor plot, the lower positioned the signature representing the focalized model, the lower the RMSE, and in a similar vein, the closer positioned the signature to the x-axis, the higher the correlation coefficient meaning that the model performed accurate predictions.

Considering the information provided above, Figure 3.10 demonstrates that the models applied within Scenario 1 (Model 1 to Model 3) have the lowest imputation performance for the gap size of 5%, while Scenario 2 (Model 4 to Model 10) results are more accurate. It can also be seen that the most successful predictive performances are obtained through the models belonging to Scenario 3 (Model 11 to Model 14) and, despite slight differences, Model 14 outperformed its counterparts. In a similar vein, the models established within Scenario 3 revealed the best modelling ability for a gap size of 10%, followed by those constituted in Scenario 2 and Scenario 1, respectively (Figure 3.11). At both 5% and 10% gap sizes, Model 14 consisting of the implementation of the XGBoost-DE algorithm had the highest accuracies. For the gap size of 20% (Figure 3.12), although the models introduced in Scenario 3 performed better compared to the Scenario 2- and Scenario 1-based models and Model 14 had the highest accuracy among Scenario 3-based models, Model 14 comprised the application of RF algorithm outperformed its counterparts (namely XGBoost and MARS). What is also important regarding Figure 3.12 is that all models had correlation coefficient values greater than 0.9 but the models exist in Scenario 1. In addition, similar to other gap size-based results, models pertaining to Scenario 3 illustrated higher predictive performances, and especially Model 14 with the hybrid XGBoost-DE algorithm yielded the best accuracy (Figure 3.13).

In general, the models belonging to Scenario 1 (i.e., interpolation-based models) have the lowest imputation success and the correlation coefficients are below 0.9 for these models. The models constituted within Scenario 2 (i.e., machine learning-based models with meteorological variables of the same station) improved prediction accuracies and generated quite close results with each other ranging correlation coefficient of between 0.9 and 0.95. The missing data imputation performance of the

models that exist in Scenario 3 (i.e., machine learning-based models only with solar radiation variable of the closer stations) gave the highest predictive success. One can observe that the models distributed close and parallelly to the x-axis that represents standard deviation values close to those calculated for the actual values, and still, there are significant differences in the correlation coefficients. This is related to the improvement of model prediction outcomes by including solar radiation values from neighboring stations in the input sets. When the models that exist within Scenario 2 are examined, it can be concluded that the meteorological parameters do not have notable impacts on the targets since there is no significant difference between the adopted models' correlation coefficients.

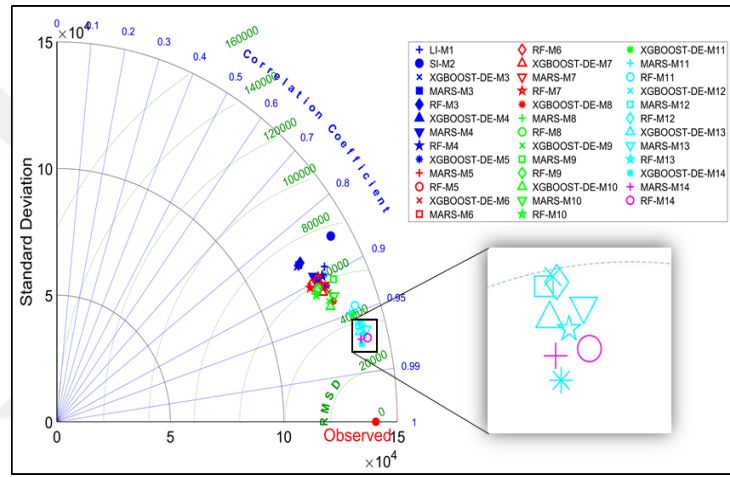


Figure 3.10 : Holistic evaluation of adopted models at 5% gap size with Taylor diagram.

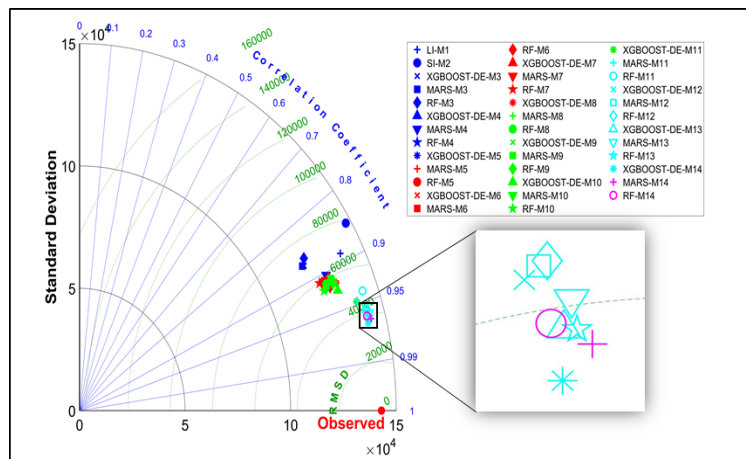


Figure 3.11 : Holistic evaluation of adopted models at 10% gap size with Taylor diagram.

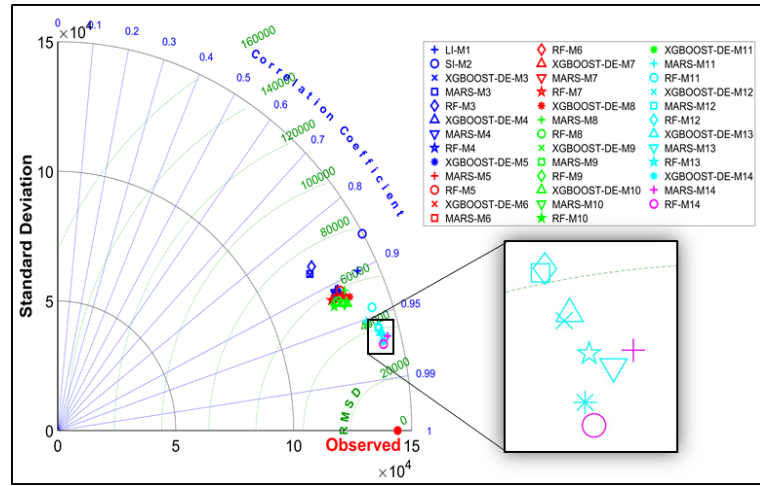


Figure 3.12 : Holistic evaluation of adopted models at 20% gap size with Taylor diagram.

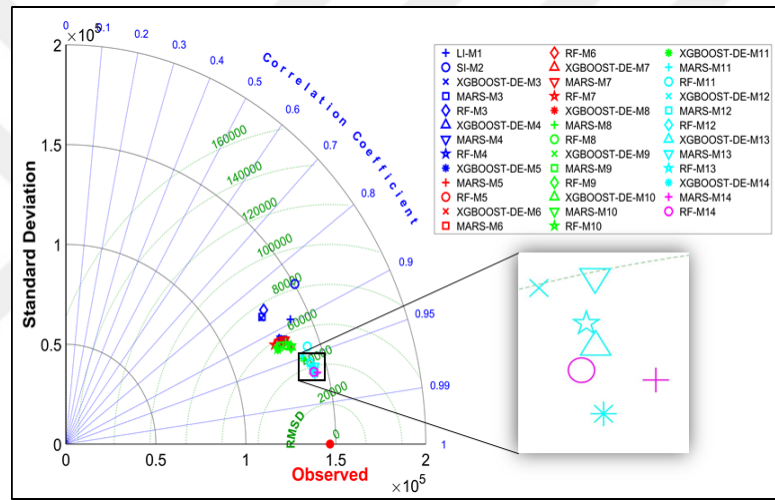


Figure 3.13 : Holistic evaluation of adopted models at 30% gap size with Taylor diagram.

3.6 Discussions, Limitations, And Insights For Future Research

The results of the current research have clearly highlighted that the interpolation techniques can be regarded as the least effective methods in imputing solar radiation values. Still, in some of the studies, the interpolation techniques yielded satisfactory performances, especially for the minute or hourly resolutions (Demirhan and Renwick, 2018; Pereira et al, 2018). The reason for this can be related to the fact that the reduction of the measurement steps as well as the increase in the serial dependence. Besides, the mechanisms regarding the missing values are of critical importance in missing value imputation attempts. However, a few of the existing literature, i.e., Demirhan and Renwick (2018) and Denhard et al. (2022), has explored the

mechanisms of the missing values as presented in Table 3.9 containing the recent studies devoted to the missing value imputations. In the present work, the artificially generated gaps were attained based on the well-known MCAR mechanism. In the vast majority of recent studies that imputed missing solar radiation data, only the incomplete time series comprising inherent gaps have been used (Demirhan and Renwick, 2018; Denhard et al, 2022; Hoyos-Gómez et al, 2022; Saaban et al, 2016; Turrado et al, 2014). Contrarily, various models with different input combinations under three different scenarios were established in order for estimating the missing values in the solar radiation series. These models are constituted based on i) missing solar radiation time series values (with interpolation techniques), ii) meteorological variables' values (with machine learning methods), and iii) solar radiation values of neighboring stations (with machine learning methods). Therefore, this study differentiated from the pertinent literature through its multi-perspective input selection strategies.

Furthermore, none of the existing attempts has implemented machine learning algorithms in imputing solar radiation values. Rather, this research incorporated the state-of-the-art XGBoost algorithm to estimate the missing values in solar radiation measurements. The XGBoost can be regarded as flexible, differentiable, as well as compatible with the user-defined objective function(s) and parallel processing. In addition, the algorithm contains regularization parameters to control the model complexity and avoids over-fitting (Liu et al, 2022). This study also utilized ensemble tree-based techniques, i.e., random forest, for benchmarking the proposed model. Among ensemble models, there are several reasons for attaining better prediction power using the XGBoost over other tree-based ensembles. The XGBoost embraces an integrated algorithm with respect to the gradient lifting decision tree optimization which is its unique advantage in reducing the variance of the estimations (Yan et al, 2022). Despite its strengths, there are some points that need to be acknowledged in using the XGBoost algorithm; such that the corresponding method is composed of several hyperparameters which might increase the computational burden, especially once the sample size increases. Yet, in this research, the hyperparameters of the XGBoost were tuned via the meta-heuristic differential evolution method which is one of the robust stochastic evolution algorithms (Rwigema et al, 2019).

Table 3.9 : Benchmarking the proposed comprehensive framework with the existing attempts.

Study	Imputation Methods	Inputs	Resolution				Missing Data Values (%)	Missingness Mechanism	Performance Criteria
			Minute	Hour	Day	Week			
Schwandt et al. (2014)	DF	GHI, DHI, DNI		✓			Ranging from 0.6 to 31	N/A	MB
Song (2014)	SSA, SASR, TBA	T, SR			✓		Ranging from 1 to 25 day	N/A	ME, MAE, RMSE
Turrado et al. (2014)	MICE, IDW, MLR	SR	✓				33	N/A	RMSE, MAE
Saaban et al. (2015)	PPCCBC	SR		✓			10,30,50	N/A	RMSE, R ²
Demirhan and Renwick (2018)	IT, KF, P, WMA, RS, CSD, SS	SR	✓	✓	✓	✓	5, 10, 25, 50	MCAR	MAE, RMSE, MASE
Pereira et al. (2018)	DF, IT	GHI, DHI, DNI, SR		✓			0.0085	N/A	MBE
Denhard et al. (2022)	KF, ARIMA, LI, SI, STI, MA, WMA, EWMA	SR		✓			N/A	MCAR	RMSE, MAE, MBE
Gomez et al. (2022)	MI, HS	SR		✓	✓		Ranging from 54.7 to 77.1	N/A	N/A
This Study	RF, MARS, XGBOOST, LI, SI	SR, T, RH, WS, N			✓		5, 10, 20, 30	MCAR	RMSE, MAE, NSE, R ² , KGE

What is also striking about Table 3.9 is that none of the studies presented for missing data imputations have considered a performance evaluation criterion taking the variances of the observed solar radiations into account. This study further assessed the predictive accuracy of the proposed models based on reliable and robust performance indicators, such as NSE and KGE, that are frequently used in statistical studies in addition to the well-known metrics, such as the determination coefficient. The reason for considering these additional performance metrics comes from the fact that they overcome some of the drawbacks of the performance metrics, e.g., determination coefficient, providing information for only linear dependence between observations and corresponding estimates (Kisi, 2010).

From a methodological aspect, although this study performed a comprehensive analysis regarding the accurate missing value imputations, it is still not free of limitations. For instance, we generated missing values only based on the MCAR mechanism, and therefore, it is planned to conduct wider applications with additional

mechanisms, such as MAR and MNAR. Also, investigating the efficiency of the deep learning techniques in increasing the accuracy of the estimations would be a valuable research direction for the follow-up attempts. Another point that needs to be considered during such analysis is the computational cost. It is worth mentioning that complex problems can be solved with exhaustive experimental efforts in laboratories (either physical or virtual) and the corresponding outcomes result in empirical evidence to be used in future research. Likewise, in our case, since we dealt with the imputation of solar energy which requires diligent investigations in various fields (such as energy production, industrial processes, environmental assessments, and understanding the Earth's energy balance and weather patterns, etc.) (Barba et al, 2019; Stephens and L'Ecuyer, 2015; Wittstadt et al, 2017), accuracies should be maintained with a high priority over other factors such as run time. Still, the computational burden should not be compromised, and therefore, additional studies devoted to optimizing the trade-off between accuracy and computational cost would be undertaken to develop a full picture of missing solar radiation imputations.

Regarding the conceptual facet, the best-performed model was found as the machine learning based imputations that used the solar radiation values of the neighboring stations. The reason behind this conclusion is mostly related to the interdependence between the solar radiation values of the neighboring stations and the target station. Yet already, the correlations between the meteorological variables and the solar radiations measured at the target station were relatively low compared to the correlations attained among the solar radiation values. Hence, examining the applicability of the proposed model and enhancing the prediction accuracies in larger scale regions encompassing more stations' information would be fertile for future research direction. Likewise, evaluating the applicability of the proposed imputation scheme for different study domains and/or meteorological variables could be a valuable research direction in order for certifying the generalizability of the introduced model. Finally, based on our results, the incorporation of solar radiation measurements of more neighboring stations into the input sets (if applicable) is expected to increase the predictive performance of the models, in which exploring the optimum number of neighboring stations to perform most accurate missing data imputations for regions covering wider surface area would result in rewarding outcomes in the following studies.

3.7 Concluding Remarks

The current research was undertaken to be a remedy to impute the potential gaps taking place in solar radiation measurements which are vital for exploring the climate and weather dynamics as well as energy technologies. To accomplish the overarching aim of this study, we proposed a machine learning based novel strategy for the missing data imputations in solar radiation measurements. Hence, a state-of-the-art extreme gradient boosting algorithm was used to fill the gaps and the applicability and efficiency of the corresponding model were examined for four different gap sizes, i.e., 5%, 10%, 20%, and 30%. In addition, since the proposed model is complex due to its several hyperparameters, we incorporated to widely applied and robust differential evolution technique for configuring the XGBoost model. This study further highlighted the predictive ability of the proposed hybrid XGBoost-DE model through benchmarking with the traditional random forest and multivariate adaptive regression splines algorithms. Besides the machine learning algorithms, we also applied two basic interpolation techniques, namely linear and spline interpolation, to impute the solar radiation missing values.

Thus, a total of three different scenarios, i.e., Scenario 1, Scenario 2, and Scenario 3, were proposed to model solar radiation gaps within the scope of the current research. Scenario 1 covers the application of interpolation techniques by using the own solar radiation measurements of the target station. Furthermore, sensitivity analysis was performed in machine learning applications in order for highlighting the added value of newly appended inputs during the estimations. Hence, the meteorological variables have been introduced to the models one by one based on their correlations with the solar radiation values through a descending order in Scenario 2, while the same rationale was embraced in Scenario 3 by adding the neighboring solar radiation values as inputs from the station having the highest correlation to the one which has the lowest correlation, respectively.

The overall results indicated that the machine learning algorithms performed better than the interpolation techniques as expected. In addition, among the machine learning models, the XGBoost-DE model that was constituted with the solar radiation measurements of all neighboring stations outweighed its counterparts with respect to the many performance indicators in all gap sizes but 20%. Furthermore, among the

models configured with the meteorological-based variables (Scenario 2), this study suggests using Model 8 which comprises N, Tmax, Tavg, RHmin, Tmin, and RHavg since appending more variables did not make any significant change in model accuracies. Considering the models established with the neighboring stations-based variables (Scenario 3), the present research recommends implementing the model having all neighboring stations' solar radiation values as inputs (Model 14 in our case).



4. ESTIMATION OF DAILY REFERENCE EVAPOTRANSPIRATION BY HYBRID SINGULAR SPECTRUM ANALYSIS-BASED STOCHASTIC GRADIENT BOOSTING ³

4.1 Introduction

Evapotranspiration (ET) - the combination of transpiration from plants and evaporation from non-stomatal surfaces – is a critical term in the hydrologic cycle (Muhammad et al, 2021). ET plays a crucial role in the management of water resources, including agricultural planning and modeling of hydrologic processes (Kim et al, 2023). There are many measurement techniques including lysimeters (Dong and Hansen, 2023), eddy covariance (Zhou et al, 2014), and Bowen ratio energy balance methods (Xiong et al, 2022), that are used to determine ET. However, these techniques are site-specific, expensive, and require continuous calibration. In recent years, various models have been developed to estimate the reference ET (ET_o), which refers to the ET of a reference crop under well-watered conditions for the purpose of estimating how much ET could occur under ideal conditions. The FAO-56-Penman-Monteith method (Allen et al, 1998) is the most widely used of these methods, but accurately estimating ET_o across space and time is still a challenge due to the complex nonlinear relationships between evapotranspiration as well as the factors that control ET_o (Adarsh et al, 2018).

In recent years, many prediction models have been developed to estimate ET_o using classical linear models (Anda et al, 2015), artificial neural networks (Ferreira et al, 2019), fuzzy logic (Mehdizadeh et al, 2017), support vector machines (Mehdizadeh et al, 2017), multivariate adaptive regression splines (Ferreira, 2022), and random forests (Wang et al, 2019). Incorporating signal processing methods into the prediction rationale tends to enhance the accuracies in ET_o predictions. For instance, Kisi and

³ This chapter is based on the paper “Başakın, E. E., Ekmekcioğlu, Ö., Stoy, P. C. and Özger, M. (2023). Estimation of daily reference evapotranspiration by hybrid singular spectrum analysis-based stochastic gradient boosting, *MethodsX*, 10, 102163. doi: 10.1016/j.mex.2023.102163.”

Alizamir (Kisi and Alizamir, 2018) used the wavelet transform to divide the ETo series into its sub-bands at different frequencies and estimated ETo values using the hybrid wavelet artificial neural network (WANN). They also benchmarked the WANN model with the results of the standalone extreme learning machine (ELM) in order to highlight the superiority of the WANN model. In addition, Lu et al. (2023) hybridized the backpropagation neural network model (BPNN) with the empirical mode decomposition (EMD) and variational mode decomposition (VMD) methods to simulate ETo series. They also compared the attained results with the outcomes of two widely embraced ML algorithms, namely support vector regression (SVR) and gradient boosting regression tree (GBRT). The VMD-BPNN model had a prediction performance of 0.405 mm/day RMSE, while the SVR and GBRT single models were 1.119 mm/day and 1.132 mm/day, respectively. In addition, Zheng et al. (2023) proposed an approach to predict 1-day ahead ETo values. They utilized a multivariate variational mode decomposition (MVMD) technique along with a soft feature filter (SSF) method in the pre-processing stage to decompose the original time series into its sub-series. For the predictions, the authors used Gated Recurrent Units (GRUs) due to its promising performance in capturing long-term dependencies within time series data and benchmarked the proposed hybrid model with standalone models. Their results showed that hybrid MVMD-SoFeFilterGRU model outperformed them in terms of prediction accuracy. Similarly, Araghi et al. (2020) aimed to predict the daily evapotranspiration in three distinct climates in Iran using three different machine learning techniques, namely ANN, Adaptive neuro-fuzzy inference system (ANFIS), and multiple linear regression (MLR). The authors applied both standalone machine learning techniques and their hybridized variants with the integration of discrete wavelet transform (DWT). The results pointed out that the models developed using the DWT had better prediction performance than the models developed without a pre-processing technique. In another study, Kang et al. (2021) investigated the effects of different preprocessing methods on the prediction of daily ETo values using four meteorological variables, including maximum temperature, minimum temperature, sunshine duration, and relative humidity. They further decomposed them into subseries using the VMD method by normalizing sub-series with the Box-Cox (BC) transformation method and compared the performance of different models, such as the standalone SVM model, the hybrid VMD-SVM model, the hybrid BC-SVM model,

and the hybrid VMD-BC-SVM model. According to the results obtained, the hybrid VMD-BC-SVM model performed the best in predicting daily ETo values.

From these results, it is clear that signal processing and machine learning methods can improve ETo predictions. Here, this study combined both signal processing (i.e., singular spectrum analysis (SSA)) and machine learning technique (i.e., stochastic gradient boosting (SGB)) techniques to predict the ETo values. The ETo values were decomposed into different sub-bands having divergent frequency and time scales using the novel SSA approach. Then, the SGB model was adopted to perform the predictions. It is worth noting that the present research is unique as it combines the SSA method with SGB methods to predict ETo values for the first time (to our understanding) in ET estimation literature.

4.2 Study Area And Data

In this study, the meteorological parameters observed in the Adiyaman Province in southeastern Anatolia, Türkiye were used to calculate the FAO-56-PM ETo values (Figure 4.1). According to the Macro Climate Region Map of Türkiye, Adiyaman is located in a continental climate zone with very hot and dry summers and very cold winters. The average temperature is 17°C, total annual precipitation is 935 mm, and the average relative humidity is 49% (Bozali, 2020). For the prediction model, daily maximum temperature, minimum temperature, wind speed, relative humidity, sunshine duration and solar radiation values measured between 2009 and 2021 in Adiyaman meteorological observation station were used. The descriptive statistics of the meteorological variables used in the study are given in Table 4.1.

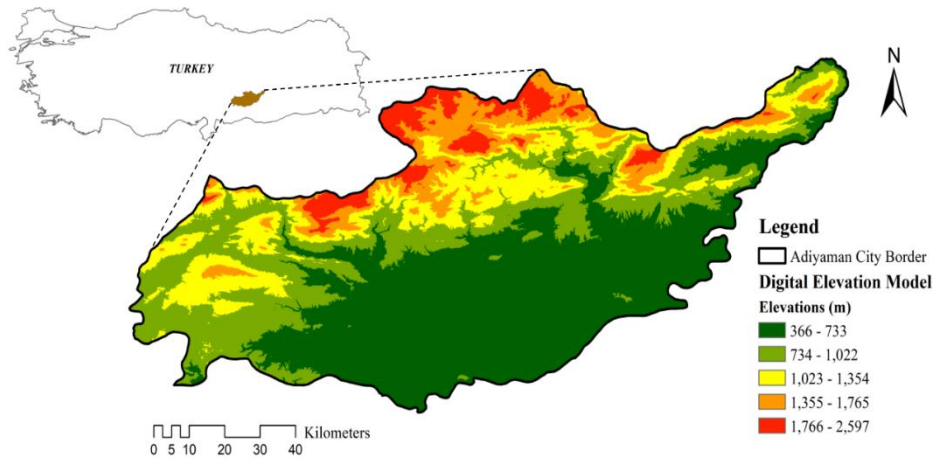


Figure 4.1 : Study area.

Table 4.1 : Descriptive statistics of meteorological variables.

	$T_{\max}$ (°C)	$T_{\min}$ (°C)	RH (%)	WS (m/s)	N (hour)	R_s (MJ/m ² /day)
Maximum	43.8	29.8	98.8	6.7	14	29.33
Minimum	-0.5	-6.3	9.4	0.5	0	0
Average	25.9	12.33	49.87	1.53	6.93	15.79
Standard Deviation	11.59	8.06	19.15	0.85	3.59	7.42
Skewness	-0.022	-0.016	-0.510	2.349	-0.652	-0.244

$T_{\max}$ = Maximum Temperature, $T_{\min}$ = Minimum Temperature, RH = Relative Humidity, WS = Wind Speed,
N = Sunshine Duration, R_s = Solar Radiation

4.3 FAO-56 Penman-Monteith Reference Evapotranspiration Model

The FAO-56 Penman-Monteith model is recognized as the most accurate empirical approach for calculating reference ETo in the absence of ETo measurements through lysimeters or other measurement techniques (Allen et al, 1998). The Penman-Monteith FAO-56 (ETo) is given as:

$$ET_0 = \frac{0.408(R_n - G)\gamma \frac{900}{T + 273} u_2 (e_s - e_a)}{\Delta + \gamma(1 + 0.34u_2)} \quad (4.1)$$

where ETo is the reference crop evapotranspiration (mm d⁻¹), Δ is the slope of saturation vapor pressure versus air temperature curve (kPa °C⁻¹), R_n is the net solar radiation at the crop surface (MJ m⁻² d⁻¹), G is the soil heat flux (MJ m⁻² d⁻¹), γ is the psychrometric constant (kPa °C⁻¹), T is the mean air temperature at 2 m height (°C), u_2 is the wind speed at 2 m height (m s⁻¹), e_s is the saturation vapor pressure (kPa), and e_a is the actual vapor pressure (kPa).

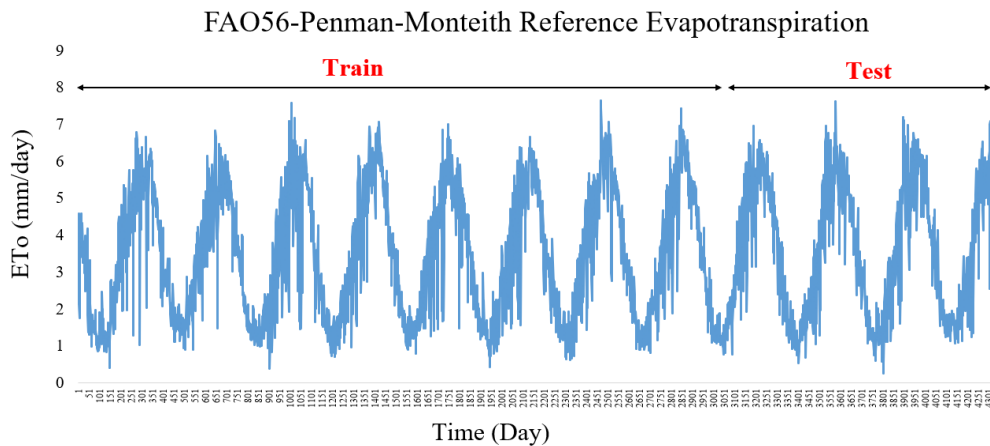
**Figure 4.2** : FAO-56-PM ETo time series starting from January 1, 2009.

Figure 4.2 illustrates the time series of the calculated ETo values. To train the proposed model, this study used 70% of the data, whereas the remaining 30% (called testing set) was used for validating the accuracy of the predictions. The distribution of the training and testing parts of the time series is shown in Figure 4.2. The minimum, average, maximum ETo values in the corresponding time series are calculated as to be 0.261 mm d⁻¹, 3.46 mm d⁻¹, and 7.63 mm d⁻¹, respectively, while the standard deviation is 1.73 and the skewness coefficient is 0.199 mm d⁻¹.

4.4 Method Details

4.4.1 Singular spectrum analysis

Singular Spectrum Analysis (SSA) is a non-parametric statistical signal processing method used to analyze and decompose multivariate time series data. It is used to extract the intrinsic components of a time series, such as trend, seasonality, and noise, by singular value decomposition (SVD) of the data matrix. The main steps in SSA are:

- **Embedding:** The first step in SSA is to embed the time series data into a 2-dimensional matrix. This is done by constructing a time-lagged matrix of the original time series, where each column is a time-lagged version of the original time series. For instance, if the original time series is of length N , the embedding matrix would have $N - m + 1$ rows and m columns, where m is the embedding dimension (or the number of lags). This embedding matrix captures the temporal dependencies within the time series data. $X = (X_1, \dots, X_N)$ is represented as the L -dimensional series, $2 \leq L \leq N$. The $Y_i = (X_i, \dots, X_{i+L-1})$ can be generated by converting $X = (X_1, \dots, X_N)$ into a trajectory matrix $Y = (Y_1, \dots, Y_L)$ as equation 4.2 (Mi and Zhao, 2020).

$$Y = \begin{pmatrix} X_1 & X_2 & \dots & X_K \\ X_2 & X_3 & \dots & X_{K+1} \\ \dots & \dots & \dots & \dots \\ X_L & X_{L+1} & \dots & X_N \end{pmatrix} \quad (4.2)$$

where, $K = N - L + 1$, and all elements along the diagonals of the Y are equal.

- **Singular Value Decomposition (SVD):** SVD is a matrix factorization technique that decomposes a matrix into its orthogonal singular vectors and singular values. The singular vectors and values provide information about the structure

and importance of the underlying components of the matrix. Then the matrix Y is rewritten as follows:

$$Y = \sum_{i=1}^d Y_i \quad (4.3)$$

$$Y_i = U_i \sqrt{YY^T} V_i \quad (4.4)$$

where U and V are the left eigenvector and right eigenvector of the matrix YY^T , respectively.

- **Grouping of Singular Vectors:** After obtaining the singular vectors and values from SVD, the singular vectors are grouped into several components that describe the intrinsic structure of the time series data. The number of components depends on the number of singular values, and the objective is to identify the components that capture the important structures in the time series data. This is usually done by choosing a certain number of largest singular values (corresponding to the largest singular vectors) to represent the underlying components. Then the matrix Y is rewritten as equation 4.5:

$$Y = \sum_{i=1}^m Y_i \quad (4.5)$$

where $m \in [1, d]$, i varies from 1 to m , and $\{Y_1, Y_2, \dots, Y_m\}$ are the useful components.

- **Reconstruction:** In the last step of SSA implementation, the original time series is reconstructed using a combination of the selected components. This stage is performed by multiplying the selected singular vectors with the corresponding singular values and summing the resulting products. The reconstructed time series provides a representation of the original time series that captures the important structures and patterns in the data.

SSA is a flexible and powerful method for analyzing time series data, as it can handle non-stationary data and missing data, and can extract underlying structures in the data (Wang et al, 2021). Additionally, it can be used to perform various tasks such as time series forecasting (Sulandari et al, 2020), de-trend

analysis (Wei et al, 2022), and anomaly detection (Guo et al, 2019), making this technique a valuable tool for many applications.

4.4.2 Stochastic gradient boosting

Stochastic Gradient Boosting (SGB) is a machine learning algorithm that operates by combining multiple weak decision tree models to create a stronger, more accurate predictive model proposed by Friedman (2001). The algorithm utilizes an iterative optimization process to continually improve the accuracy of the model. The optimization process begins by training a simple base decision tree model on the data. The model is then used to make predictions on the data, and the prediction errors are calculated. Subsequent decision tree models are then trained on the prediction errors made by the previous model, with each model focusing on minimizing the prediction errors. This process iteratively continues until a stopping criterion is met, such as a pre-defined number of iterations or when the improvement in accuracy reaches a specified threshold. The final SGB model is created by combining the predictions of all the decision tree models in the sequence using a weighted combination, where the weights are determined based on the performance of each individual decision tree model. The result is a single, more accurate model that is capable of making accurate predictions on new data. One of the key advantages of the SGB algorithm is its ability to identify important features in the data (Pasha and Mohamed, 2022). The iterative optimization process allows the algorithm to continually refine its understanding of the relationships between the features and the target variable, leading to a final model that accurately captures the key drivers of the target variable. Additionally, SGB has been shown to be effective in preventing overfitting since the iterative optimization process helps to control the complexity of the model. Another advantage of SGB is its ability to model non-linear relationships between data points (Forkuor et al, 2017). The algorithm can handle complex systems by combining multiple decision tree models, each of which is capable of modeling non-linear relationships. This makes SGB a valuable tool for a variety of predictive modeling tasks, especially when the relationship between the features and the target variable is complex.

4.5 Method Implementation

The flowchart of the study is presented in Figure 4.3. In this study, a machine learning framework comprising the hybridized implementation of SSA and SGB was utilized to predict the ETo time series. Before establishing the predictive model, the ETo time series was decomposed into sub-series using the SSA signal processing, which resulted in 8 sub-bands (Figure 4.4). Upon examination of Figure 4.4, the first band represents the approximations sub-band which reflects the long-term trend indication of multi-annual components, the second band represents the seasonal component, and the eighth band denotes the white noise. The other bands contain various detailed components of the time series.

Each sub-series was predicted using the SGB method. During the model setup, 70% of the data was used for training and the performance of the model was tested with remaining 30% (Figure 4.2). The flowchart of the present study containing all the stages to perform the predictions is given in Figure 4.3. It is worth mentioning that the partial autocorrelation function (PACF) was used to determine the input variables. Thus, both short-term and long-term predictions covering 1, 7, 14, 30, 45, and 60-days ahead were carried out based on the inputs determined through the PACF analysis. It is also important to note that the hyperparameters of the SGB algorithm were determined using a well-known 10-fold cross-validation approach (Xu et al, 2017).

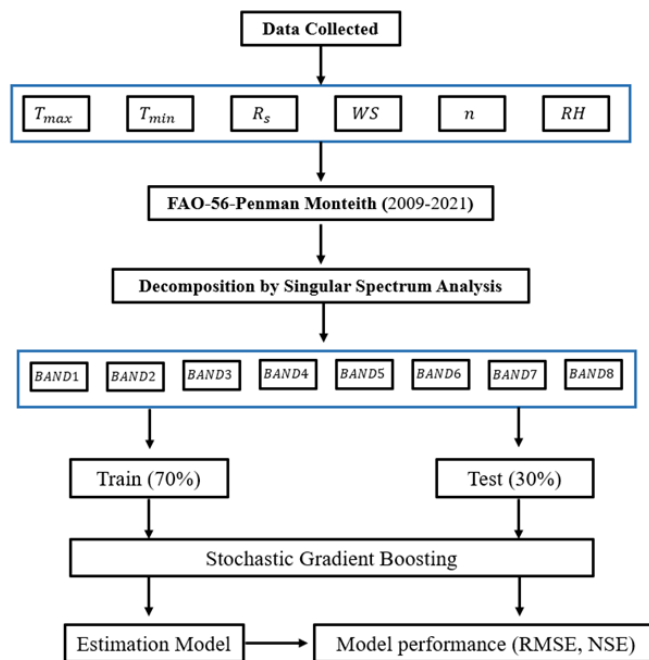


Figure 4.3 : Flowchart of the current research.

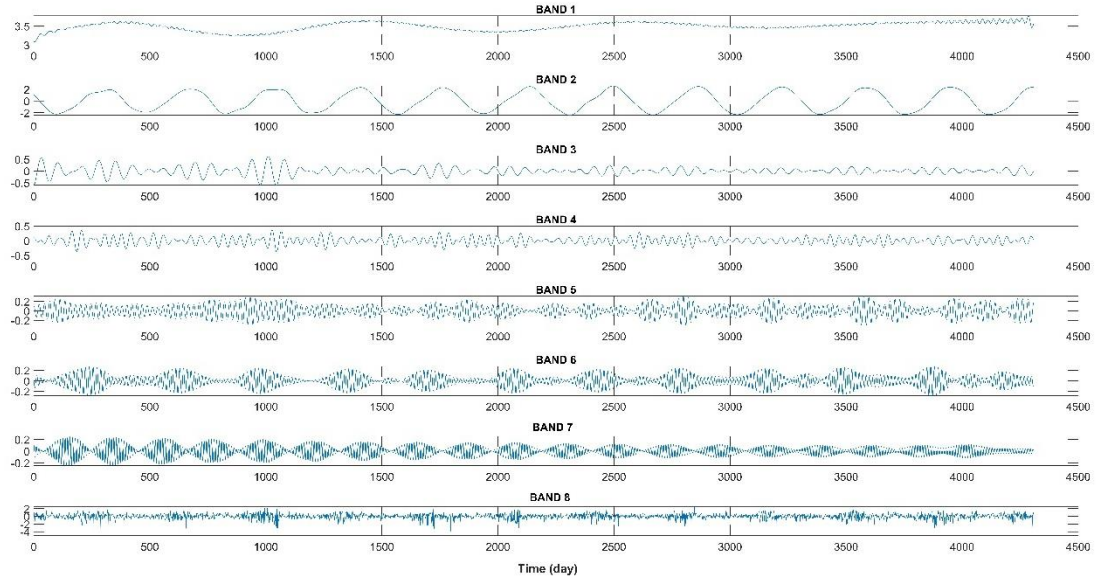


Figure 4.4 : Components of FAO-56-PM time series that decomposition by SSA.

The root mean square error (RMSE) and Nash-Sutcliffe efficiency (NSE) (Nash and Sutcliffe, 1970) indicators were used to evaluate the accuracy of the predictions. RMSE is a commonly used performance measure in regression attempts. The smaller the RMSE values, the higher the predictive power of the model. In addition, the NSE is another performance index used to evaluate the accuracy of ML models and is commonly used in the hydrology domain. NSE can also be considered as the normalized version of the coefficient of determination (R^2) and ranges between $-\infty$ and 1. Here, a NSE of 1 indicates perfect agreement between the observed and predicted values, while values closer to 0 indicate worse performance. The equations regarding the utilized performance indicators are presented in equation 4.6 and 4.7.

$$RMSE = \sqrt{\sum_{i=1}^n \frac{(\hat{y}_i - y_i)^2}{n}} \quad (4.6)$$

$$NSE = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y}_o)^2} \quad (4.7)$$

where y_i is the actual ETo values, $\hat{y}_i$ is the modeled ETo values, $\bar{y}_o$ is the average of actual ETo values, and n denotes the number of observations.

4.6 Methods Validation

In this study, two different models were developed for the estimation of reference ET values. In the first model, only the SGB method was used (called ‘STD’ to confirm the differentiation), whereas the second model comprises the hybrid use of SSA and SGB. The predictions were made for six different lead times by utilizing 3 lag values defined as inputs. It is worth mentioning that the original time series was subjected to the partial autocorrelation function (PACF) analysis and the inputs were determined in this regard. Figure 4.5 illustrates the PACF analysis results. As can be seen from the figure that the first three lag times had the highest correlation with the original time series, and therefore, we included $t-2$, $t-1$, and t as inputs in order to perform the predictions. In addition, the predictions were conducted for six different lead times under the umbrella of two horizons, i.e., short-term and long-term. In accordance with the study of Lu et al. (2023), predictions made for 1, 3, 5 and 7 days ahead were regarded as short-term, whereas the others (i.e., 30, 45, and 60 days ahead) were referred to as long-term predictions.

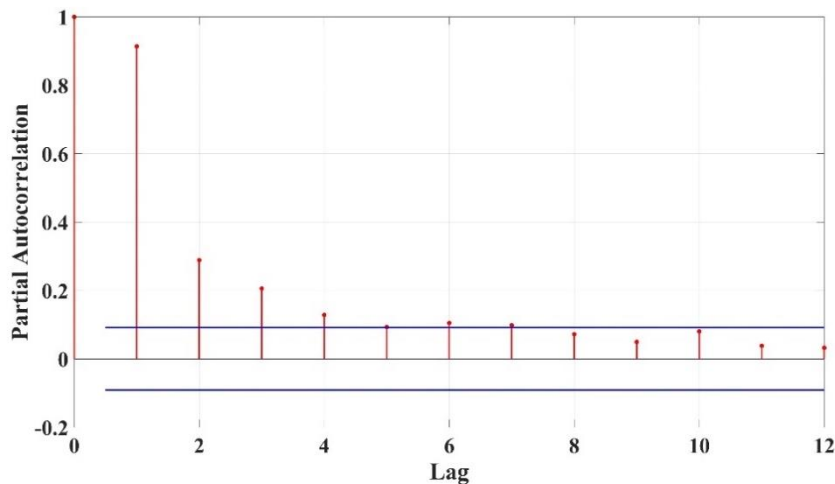


Figure 4.5 : Partial autocorrelation function results.

The train and test prediction performance results of the models are provided in Table 4.2. Upon examining Table 4.2 with respect to the test set results, acceptable predictions were achieved for 1, 7, and 14 lead times (i.e., short-term) in the STD models with NSE values of 0.877, 0.791, and 0.762, respectively. Compared to the short-term predictions, lower performances were obtained regarding 30 (0.633), 45 (NSE:0.442) and 60 (NSE: 0.264) lead-time values in the long-term predictions. As expected, overall results illustrated that the highest performance was achieved for the

1-day ahead prediction, while the lowest accuracy was obtained for 60-day ahead predictions. For the training set, one can also conclude with similar outcomes as the model showed decreasing accuracy trend when the horizon is extended.

Table 4.2 : Model performance results.

Model	Inputs	Outputs	Train		Test	
			RMSE (mm/day)	NSE	RMSE (mm/day)	NSE
STD	t-2, t-1, t	t+1	0.623	0.868	0.629	0.877
STD	t-2, t-1, t	t+7	0.817	0.773	0.816	0.791
STD	t-2, t-1, t	t+14	0.834	0.764	0.870	0.762
STD	t-2, t-1, t	t+30	1.023	0.647	1.078	0.633
STD	t-2, t-1, t	t+45	1.223	0.499	1.321	0.442
STD	t-2, t-1, t	t+60	1.389	0.356	1.511	0.264
SSA	t-2, t-1, t	t+1	0.576	0.887	0.510	0.919
SSA	t-2, t-1, t	t+7	0.611	0.873	0.608	0.884
SSA	t-2, t-1, t	t+14	0.675	0.845	0.660	0.852
SSA	t-2, t-1, t	t+30	0.828	0.769	0.859	0.767
SSA	t-2, t-1, t	t+45	0.987	0.674	1.020	0.667
SSA	t-2, t-1, t	t+60	1.094	0.601	1.143	0.579

STD= Standalone Stochastic Gradient Boosting model, SSA= Hybrid Singular Spectrum Analysis and Stochastic Gradient Boosting model, t=lag time 1, t-1= lag time 2, t-2= lag time 3, t+1=lead time 1, t+7=lead time 7, t+14=lead time 14, t+30=lead time 30, t+45=lead time 45, t+60=lead time 60, RMSE= Root Mean Square Error, NSE= Nash-Sutcliffe Efficiency

Once the SSA analysis was incorporated into the predictions, significant performance improvements have been achieved for all lead time values. Short-term predictions demonstrated promising performance as the NSE values were obtained as 0.919, 0.884, and 0.852 for 1-day, 7-day, and 14-day lead times. Despite unsatisfactory performance of the STD model with regard to the long-term assessments, the proposed hybrid model yielded promising improvements. Such that, the NSE results for long-term lead time values (e.g. 30, 45 and 60 days) were calculated as 0.767, 0.667, and 0.579, respectively. The table shows a remarkable similarity between the achieved results for the training and testing sets, indicating that the proposed model is highly resistant to overfitting. This finding is particularly striking, as it confirms the model's robustness and reliability in generalizing to new data. Furthermore, scatter plots are shown in Figure 4.6 to visually evaluate the prediction performance of the models. When examining the relationship between observed values and predicted values, it is expected that the best predictions will be scattered on the 1:1 line. Based on this graph, it can be observed that the distribution of the points generated using the SSA method are closer to the 1:1 line compared to those attained by the STD model, which is the

clearest indication that the incorporation of the SSA approach into the SGB method improves the prediction performance. As with the statistical performance evaluation metrics, the scatter plots also indicate that the model's performance tends to decrease as the lead time increases as they have greater variations concerning the 1:1 perfect line and contain more underestimated and overestimated predictive outcomes. This observation underscores the importance of carefully considering the impact of temporal dependencies when modeling time series data. Moreover, while the standalone SGB method performed a relatively uniform prediction for all lead time values, the monotony in the predictions of low values was eliminated by the SSA-SGB method.

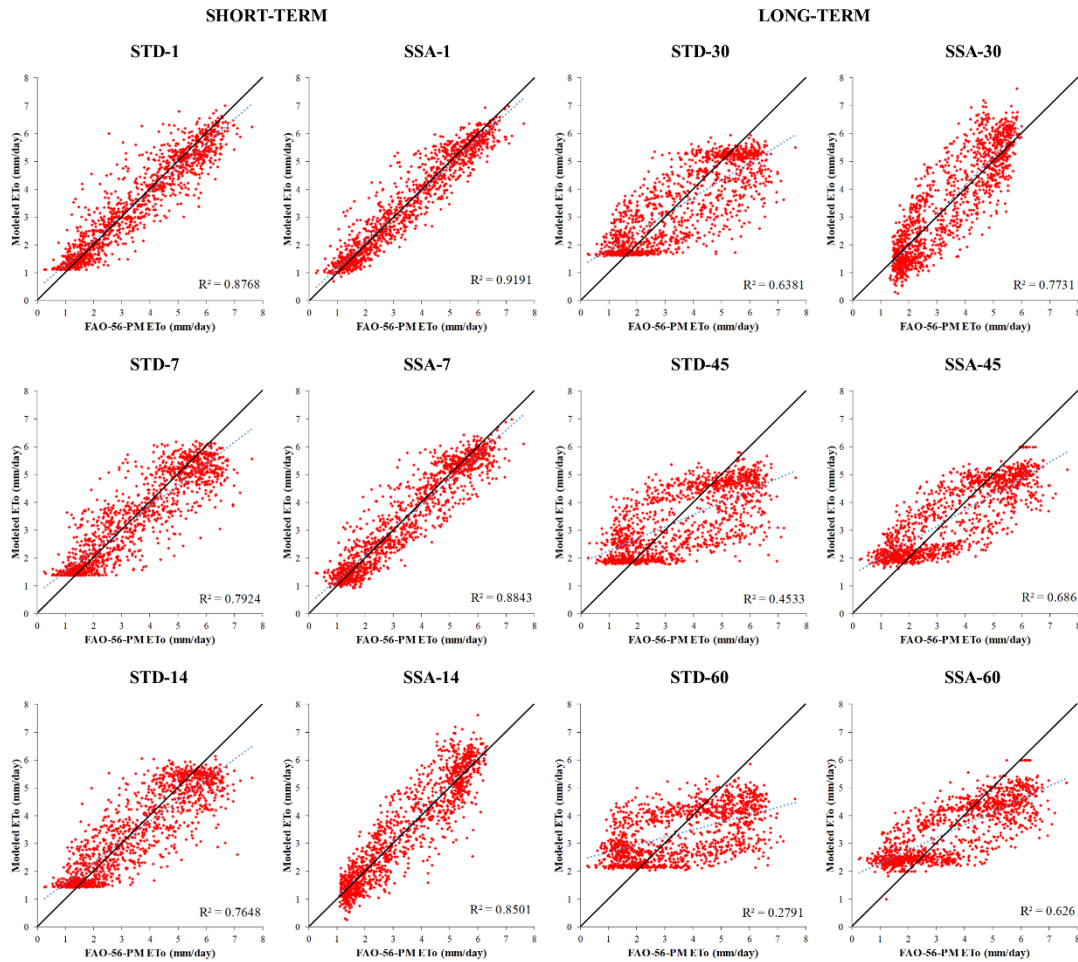


Figure 4.6 : Models scatter plots.

4.7 Conclusion

In this study, the ETo values were estimated using the SGB, which is one of the robust tree-based ensemble machine learning methods. To acquire higher performance, the

SGB further reinforced with the time series decomposition technique, called SSA. The original time series was decomposed into different frequency components in the study, and the SGB method was used for the predictions. Compared to the predictions performed with the stand-alone SGB method, the hybrid SSA-SGB method yielded more accurate outcomes. It is worth mentioning that these results provide insightful contributions regarding the determination of future ETo values in promoting the effective solutions regarding the water resource management. Hence, the current research provided a new perspective to the body of knowledge with improved prediction accuracies. Despite this study attempted the perform improved predictions, this study has still some limitations. For instance, the models were built using data from a single station, and hence, future research new models using data from a large number of stations over a larger area can be built in order to validate the applicability of the proposed hybrid prediction framework. In addition, to explore the utility of other techniques, different machine learning algorithms and signal processing techniques can be adopted to the predictions in the follow-up attempts.



5. GROSS PRIMARY PRODUCTIVITY RESEARCH: ONGOING TRENDS AND FUTURE TRAFECTORIES⁴

5.1 Introduction

Plants, the silent engines of our planet, breathe in carbon dioxide through photosynthesis, fueling life on Earth (Lees et al, 2018; Schimel, 1995). Gross Primary Productivity (GPP), the total amount of carbon dioxide (CO₂) absorbed by plants through photosynthesis, plays a critical role in the global carbon cycle and Earth's climate system (Bonan, 2016; Schimel, 1995). Terrestrial ecosystems are the largest global carbon sink, and terrestrial GPP is the largest flux in the global carbon budget (Schimel et al, 2015; Walker et al, 2021). Understanding GPP is therefore essential for quantifying the global carbon cycle.

The heightened scientific focus on the terrestrial carbon cycle in recent decades stems from the urgent need to predict climate change and its cascading effects. Remote sensing has emerged as a transformative tool for studying the carbon cycle, enabling researchers to estimate ecosystem carbon fluxes and stocks across vast spatial scales (Running et al, 2004; Saatchi et al, 2011; Xiao, 2014). This capability allows us to understand how global change drivers, such as climate change itself, land-use alterations, and atmospheric pollution, influence carbon dynamics within ecosystems (Li et al, 2018; Nemani et al, 2003). Understanding these dynamics is essential for predicting the resulting feedbacks between the carbon cycle and climate that can further amplify or mitigate climate change (Ollinger et al, 2008; Xiao et al, 2014). The availability of continuous satellite observations since the 1970s, particularly from the early 1980s onwards, has revolutionized our ability to assess the magnitude, spatial patterns, temporal variability, and long-term trends of GPP across landscapes, regions, and the entire globe. Summarizing all aspects of GPP research exceeds the scope of

⁴ This chapter is based on the paper “Başakın, E.E., Stoy, P.C., Ranjbar, S., Demirel M.C. (2024). Gross Primary Productivity research: Ongoing trends and future trajectories, Carbon Fluxes and Biophysical, ESLEVIER.”

any single analysis. Instead, we briefly summarize the major approaches that have been used to estimate and model GPP before undertaking a literature review that describes the state of GPP research and its current trajectory with an emphasis on opportunities to further improve global collaboration, predictive capabilities, and mechanistic understanding.

5.2 Background

5.2.1 Estimating GPP from ecosystem scale measurements

GPP represents the total amount of carbon dioxide taken up by plant canopies and cannot be measured directly in its entirety at the ecosystem scale. Various methods, including, flux tower networks (eddy covariance) (Desai et al, 2008; Owen et al, 2007), remote sensing (Xiao et al, 2019), chamber-based measurements (Cheng, 2024), stable isotope techniques (Bernacchi and VanLoocke, 2015), modeling approaches, and monitoring of stem respiration and growth rates (Yu-E et al, 2007), are employed to estimate GPP, each with benefits and limitations.

Traditional methods include biomass tracking, providing an indirect measure of GPP through total vegetation C gain at different time points (Malhi, 2012). Because total vegetation C gain represents the balance between GPP and autotrophic respiration, GPP is not measured directly in biomass tracking methods and must be inferred.

One can also make multiple measurements of leaf-level carbon dioxide uptake, from which net photosynthesis can be measured and gross photosynthesis inferred. These can be used to ‘scale up’ from plant-scale to canopy-scale GPP estimates, ideally by measuring leaves in different positions in the canopy – i.e. sunlit and shaded leaves – and leaves from different species and across the diurnal cycle and seasons. This method is labor-intensive and remains subject to uncertainties in the canopy scaling procedure.

The eddy covariance technique is increasingly used to estimate GPP at the canopy scale, but also cannot measure it directly. Rather, eddy covariance measures the net ecosystem exchange of CO₂ (NEE), from which GPP can be inferred after applying a model for ecosystem respiration (RECO). Multiple approaches to estimate GPP from eddy covariance observations have been described in existing syntheses (Reichstein et al, 2012). These include “nighttime” approaches which create RECO models based on

nighttime NEE observations in the absence of GPP (in ecosystems without CAM plants) (Reichstein et al, 2005), “daytime” methods which infer RECO from the intercept of the light response curve for NEE (Lasslop et al, 2010), approaches which account for the strong suppression of turbulence in the early evening when ecosystem temperatures remain relatively warm (van Gorsel et al, 2009; Van Gorsel et al, 2007), approaches which seek to infer the inhibition of leaf respiration by light (Keenan et al, 2019), isotopic approaches (Wehr and Saleska, 2015) and more. All approaches benefit from established methods for inferring the uncertainty of both eddy covariance observations (Richardson et al, 2006) and the GPP and RECO estimates which arise from flux partitioning algorithms (Pastorello et al, 2020).

5.2.2 Estimating GPP from remote sensing

Multiple remote sensing approaches have been indispensable for estimating GPP at ecosystem to global scales (Figure 5.1) and we briefly describe them here. Light use efficiency (LUE) models have been foundational for GPP research by estimating the efficiency of plants in converting absorbed photosynthetically active radiation (APAR) into biomass (Landsberg and Waring, 1997; McMurtrie et al, 1994). LUE models often assume a linear relationship between GPP and APAR, with adjustments made for temperature, vapor pressure deficit, soil water content, or other environmental constraints. However, while GPP tends to exhibit a linear trend with APAR over monthly to annual time scales, it can be highly nonlinear on a daily basis (Medlyn et al, 2003; Turner et al, 2003), making them less suited to understand the sub-daily processes that can determine GPP. Moreover, the parabolic response of GPP to temperature in traditional LUE models may lead to an underestimation of photosynthetic production, particularly during warmer seasons, emphasizing the complexities of effectively measuring and modeling canopy temperatures. LUE models may fail to capture more complex controls on photosynthesis. For example, in boreal and temperate conifers, the primary impact of temperature on photosynthesis is believed to occur through acclimation to the annual temperature cycle, characterized by a temperature-dependent dynamic delay process (Mäkelä et al, 2004; van Dijk et al, 2005). In other words, the complexities of canopy thermodynamics are difficult to accurately observe and simulate to effectively model GPP. These challenges have led to ongoing improvements to the LUE approach (Pei et al, 2022), including new insights from machine learning (Kong et al, 2023).

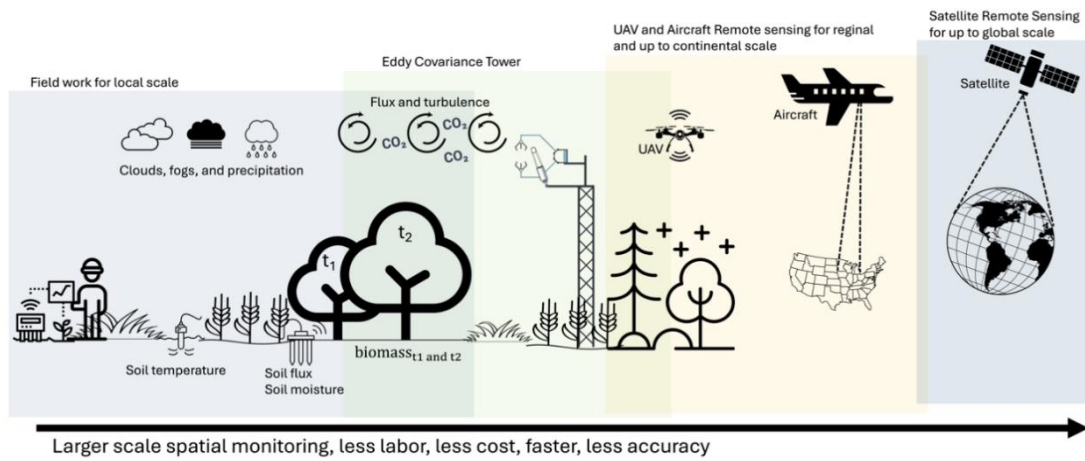


Figure 5.1 : Schematic view of Gross Primary Productivity (GPP) and its measurements methods.

Vegetation indices (VIs), such as the normalized difference vegetation index (NDVI) and the enhanced vegetation index (EVI), are another tool frequently used to approximate or directly estimate GPP (Rahman et al, 2005; Xiao and Moody, 2004; Zhou et al, 2001) through their relationship with leaf area index (LAI) and thereby APAR (Potter et al, 1993; Running et al, 2004). Although NDVI is widely used because of its simplicity, the kernel-NDVI (kNDVI) method, which better represents nonlinear relationships, also gives very good results for GPP estimation (Camps-Valls et al, 2021). The most important feature that highlights kNDVI is its performance in capturing the nonlinear relationships of spectral bands (Pabon-Moreno et al, 2022). In a study conducted by Wang et al. (2023), the kNDVI approach was validated over several Ameriflux sites for its features of generalization, boundedness, robustness to noise, and low error propagation. Ongoing innovations in VIs seek to capture additional complexities of canopy structure and function. New research has discovered that the near-infrared reflectance of vegetation (NIRv) – the NDVI multiplied by infrared reflectance – is strongly related to GPP (Badgley et al, 2019), especially when multiplied by photosynthetically active radiation (P) to obtain the NIRvP (Dechant et al, 2022), leveraging the reflectance properties of vegetation to map a proxy of GPP based on spectral data (Dechant et al, 2022; Khan et al, 2022).

5.2.3 Model-data fusion and machine learning

Emerging methods offer advancements in GPP estimation by combining machine learning, remote sensing and ground-based flux measurements from eddy covariance towers (Campos-Taberner et al, 2024; Dechant et al, 2022; Jung et al, 2020; Qiu et al,

2020; Ranjbar et al, 2024; Reyes-Muñoz et al, 2024; Zhu et al, 2023). Using ground based GPP measurements from eddy covariance towers to train machine learning models improves the accuracy of GPP prediction (Xiao et al, 2019; Zhu et al, 2024). These newer methods present promising avenues for enhancing our understanding of GPP dynamics and their implications for ecosystem functioning and climate change mitigation.

To address these critical questions – the contemporary status of the field, the evolution of research methodologies, and the existing research framework – this chapter delves into the ongoing saga of GPP research. We will explore the challenges of measuring this vital metric and the diverse methodologies employed to predict it. We will show how efforts to measure GPP have evolved from traditional techniques like biomass measurement and the LUE models to newer modeling of GPP like combination of remote sensing and machine learning that promise a deeper understanding of GPP dynamics. This analysis will not only illuminate the current landscape of GPP research but also pave the way for identifying potential research opportunities that can propel the field forward.

5.3 Methodology For Literature Review

We conducted a bibliometric analysis to analyze the past and current trends of GPP research and use this analysis to infer future trajectories. Bibliometric analysis is used to systematically and quantitatively examine the scientific literature and describes the network of research materials based on citations to provide information about the conceptual and social structures of a specific field, along with its evolution and dynamic characteristics (Akın et al, 2023; Pritchard, 1969). Additionally, it aims to demonstrate how particular research areas are structured and how they progress over time, as well as to explore research gaps (Fergnani, 2019), as we investigate here for GPP. Relevant publications on a specific topic or field are first collected from academic databases. Publications are selected within a specific time frame using predefined keywords. During the analysis phase, various metrics are employed. These metrics generally include the measurement of information related to authors, institutions, journals, and countries found in the studies, as well as various analyses of the words contained in titles, keywords, and references. These analyses typically include basic statistical information (mean, frequency, etc.).

Bibliometrix is one of the most frequently utilized bibliometric analysis tools (Aria and Cuccurullo, 2017) and has been used to analyze research in important aspects of carbon cycle science including blue carbon (Xu et al, 2023; Zhong et al, 2023), the low carbon economy (Yao et al, 2023), and sustainable agriculture (Sánchez et al, 2022). We used bibliometrix v4.2.3 in R (R Core Team, 2023) for the present analysis. In addition to the bibliometrix package, we used the VOSviewer software developed by the Centre for Science and Technology Studies at Leiden University in the Netherlands. This tool supports the generation of visual bibliometric knowledge maps through network data, allows for the examination of literature keywords, subject terms, authors, among other elements, and provides text mining features (van Eck and Waltman, 2010).

Information about some of the terms and parameters used in bibliometric analysis is shared under this title. There are some important parameter values that should be selected by the expert, especially for journal, co-occurrence network of keywords and thematic map analyses. These values are subject to expert judgment and there are no set rules for their selection. For the journal analyses, the ‘h’, ‘g’, and ‘m’ indices include quantitative information measuring the number of articles with at least h citations over time, the maximum number of times a given set of articles has been cited more than that number, and the continuous research activity of a journal since its first publication. The h-index tends to increase with career length, while the m-index can be used in situations where this is a limitation, such as when comparing journals in a field with very different career lengths. The m-index measures a journal’s continuous research activity since its first publication. These measures are used to minimize the drawbacks of examining the number of publications or the number of citations alone. In the co-occurrence network of keywords analysis, the number of links and the number of clusters between the elements that make up these links must be determined by the user in order to obtain the graphs. In this study, the minimum number of occurrences was set to 3, while the minimum cluster size was set to 50. When preparing a thematic map, it is important to pay attention to several key points. The selection of minimum cluster frequency (per thousand documents), number of words, and number of labels can lead to significant changes on the map. The literature does not define a precise number for the selection of these numbers. The choice of these parameters is up to the user and is the most in need of development in bibliometric analysis. In this

study, min cluster frequency was chosen as 5, number of words as 350, and number of labels as 5.

The analysis procedure adopted within this study encompasses three stages. In the first stage, general statistical information regarding the current literature, based on the screening results, is shared. In the second stage, the bibliometric analyses obtained through bibliometrix is presented, and in the final stage, we conduct in-depth look at studies in potential areas based on current data and the discussion of future promising research topics. As a result of all these analysis, comprehensive future-oriented results and perspectives are formulated.

5.4 Database And Search Strategies

The Scopus database was utilized to conduct a holistic literature review. Scopus is exceptionally rich in terms of the number of journals and articles (Mongeon and Paul-Hus, 2016) and is one of the most frequently used databases for bibliometric analysis, alongside the Web of Science (WoS) (Singh et al, 2021). The preference for Scopus in our search operation stems from its comprehensive coverage of both fundamental and applied research, in contrast to WoS, which prioritizes articles detailing primary research (Stahlschmidt and Stephen, 2020). The primary focus of the search process was on “gross primary productivity” and “GPP”, and keywords such as “estimate”, “prediction”, “forecasting”, and “monitoring” were employed to scan for studies likely to involve GPP. The studies analyzed in this work range from publication dates of 2000 to 2023 and were limited to research papers, review articles, book chapters, and conference papers written in English. All searches were selected among titles, abstracts, and keywords. The following search string was used to extract bibliographic data from the Scopus database: TITLE-ABS-KEY (“gross primary producti*” AND “estimat*” OR “predict*” OR “forecas*” OR “monitor*”) AND PUBYEAR > 1999 AND PUBYEAR < 2024 AND (LIMIT-TO (DOCTYPE , “ar”) OR LIMIT-TO (DOCTYPE , “cp”) OR LIMIT-TO (DOCTYPE , “re”) OR LIMIT-TO (DOCTYPE , “ch”)) AND (LIMIT-TO (LANGUAGE , “English”)).

5.5 Results and Discussion

5.5.1 Basic information of field research

The screening resulted in a total of 3447 publications. Of these, 3208 are research articles, 53 are review articles, 24 are book chapters, and 162 are conference papers. General statistical information pertaining to the documents retrieved from the search results is provided in Table 5.1. Publications have an average age of 7.71 years over the 24-year study period. The number of documents authored by a single individual is considerably low (58), and the average number of co-authors per document exceeds 7, demonstrating the collaborative nature of most GPP research. To facilitate more effective analyses, Keywords Plus (ID) keywords were included in the analyses in addition to the author-provided keywords (DE). ID is a feature provided by some academic databases like Web of Science where additional words or phrases are generated to an article in addition to DE on the frequently occurring words and phrases in titles of the article's references. Hence, these are not provided by the authors but through the use of the most relevant and repeated terms in the cited literature. ID serves to capture the essence of the research in a broader context, which can influence the discoverability of such articles that might not be immediately captured through the author's own keywords. This can help to identify work that could become relevant through various challenges related to a specific publication. Trends in the number of publications per year are presented in Figure 5.2. The publications are divided into 3 groups temporally. These groups are the “early stage”, where the total number of publications is relatively low (Wang et al, 2024), then the “activation period” (Tan et al, 2024), where the number of publications increases but there is a more stable increase compared to the first period, and finally the acceleration period, where the number of publications increases exponentially also called rapid growth stage (Wang et al, 2024). During the early stage (2000 and 2007), a total of 240 publications (6.96%) were identified. Some notable publications from this period have begun to present the first examples of the use of remote sensing products and have analyzed GPP values in the North American ecosystems (Sims et al, 2006), Italian forests (Maselli et al, 2006), and tropical monsoon regions (Gebremichael and Barros, 2006). The period that can be referred to as the “activation period”, covering the years 2007 to 2016, has been identified with a total of 1254 publications (36.38%). The last period, marked by an exponential increase in the number of publications, is termed the “acceleration period”

covering the years 2017 to 2023. This period includes a total of 1953 publications, accounting for 56.66% of the total number of publications. Detailed analyses of these three periods are provided under the heading Bibliometric Analysis.

Table 5.1 : Results of a Scopus database search on GPP research.

Description	Results
Timespan	2000:2023
Sources (Journals, Books, Conference Papers)	493
Documents	3447
Annual Growth Rate %	14.12
Document Average Age (year)	7.71
Average citations per document	44.54
References	181775
Keywords Plus (ID)	9400
Author's Keywords (DE)	6287
Authors	9615
Authors of single-authored documents	53
Single-authored documents	58
Co-Authors per document	6.93
International co-authorship %	47.29
Articles	3208
Book chapters	24
Conference papers	162
Reviews	53

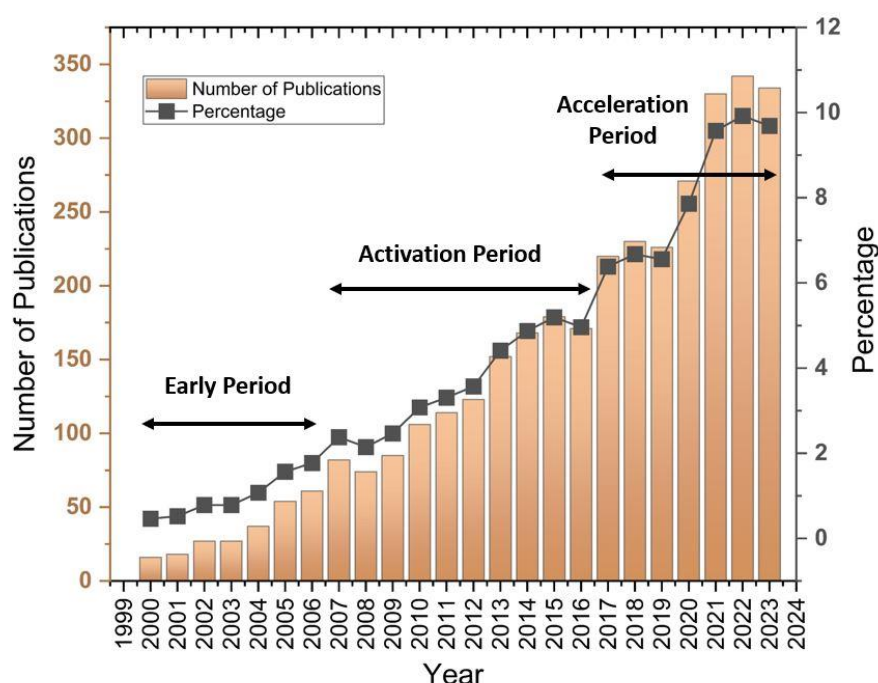


Figure 5.2 : Annual publications in GPP research listed in Scopus from 2000 to 2023.

5.5.2 Top countries

The most productive 10 countries within the search criteria are listed in Table 5.2, considering only publications that are classified as articles. China has the highest number of Single Country Publications (SCPs), but has the third lowest Multi Country Publication (MCP) ratio. MCP signifies the quantity of publications featuring at least one international co-author per country, highlighting the level of global cooperation intensity each country exhibits. This indicates that publications originating from China have a lower number of collaborative studies with different countries. The USA ranked second in articles and similar to China, also falls below the top ten countries' average MCP ratio (40%). Of the top ten countries for GPP research, Germany has the highest MCP ratio (76%) and Japan the lowest (34%) (Table 5.2). European countries, like United Kingdom, Germany, and France, have a lower number of publications compared to China and the USA, but significantly higher MCP. This is attributed to Europe's relatively smaller geographical size hosting several countries, and the presence of neighboring countries with flux towers, and historical collaborations with the CarboEurope and ICOS networks of eddy covariance towers, which makes collaboration more feasible.

Table 5.2 : Publication numbers of the top 10 corresponding author's countries.

Country	Articles	SCP	MCP	Frequency (%)	MCP (%)
China	820	470	350	24%	42 %
USA	739	443	296	21%	40%
Germany	138	33	105	4%	76%
Canada	133	53	80	4%	60%
France	123	43	80	4%	65%
Japan	122	80	42	3%	34%
United Kingdom	111	41	70	3%	63%
Australia	87	42	45	3%	51%
Italy	83	44	39	2%	47%
Spain	71	33	38	2%	53%

SCP: Single Country Publications, MCP: Multiple Country Publications

Additionally, the publication map, which is prepared by considering the affiliations of all authors, is presented in Figure 5.3, where it is clearly shown that the areas with a high concentration of publications are developed and developing countries. From most African countries, there appear to be no studies conducted in the field of GPP estimation. Furthermore, another insight that can be derived from the map is that

Western European countries have conducted more studies compared to Eastern European countries.

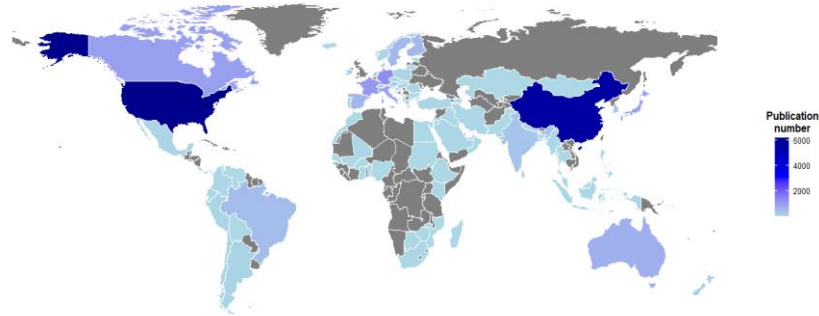


Figure 5.3 : The geographical distribution of papers based on all authors' affiliations.

5.5.3 Top journals

Agricultural and Forest Meteorology had the greatest number of GPP-related publications (NP), whereas *Remote Sensing of Environment* published manuscripts with the highest h-index and the highest g-index (Table 5.3). Papers published in *Global Change Biology* have the highest number of total citations (TC), and *Remote Sensing* has the highest m-index. Examining the year of first publication (PY start) numbers, *Remote Sensing of Environment* and *Agricultural and Forest Meteorology* have historically led the publications within the search criteria, while *Science of the Total Environment* and *Remote Sensing* have gained prominence more recently.

Table 5.3 : The top 10 journals by source impact for the search criteria related to GPP research.

Sources	h- index	g- index	m- index	TC	NP	PY start
Remote Sensing of Environment	70	124	3.18	16067	164	2003
Global Change Biology	65	122	2.6	15391	150	2000
Agricultural and Forest Meteorology	62	103	2.58	13438	279	2001
Biogeosciences	46	85	2.3	8532	185	2005
Journal of Geophysical Research: Biogeosciences	40	72	2.1	5712	143	2006
Remote Sensing	34	56	2.12	4355	196	2009
Limnology and Oceanography	31	59	1.24	3551	59	2000
Ecosystems	30	44	1.3	2808	44	2002
Global Biogeochemical Cycles	29	48	1.16	3898	48	2000
Ecological Modelling	28	46	1.16	2239	63	2001

TC: total number of citations, *NP:* number of publications, *PY start:* year of first publication

5.5.4 Top institutions

The identification of productive institutions is crucial for potential authors who wish to conduct research in a specific field, as it facilitates the establishment of possible collaborations by highlighting institutions where such cooperative efforts can be fostered. This accelerates the formation of potential collaborations and allows researchers to initiate partnerships with scholars from productive institutions. Chinese institutions are the most productive overall, Max Planck Institute for Biogeochemistry (4th) is the most productive European institution and the University of California (7th) is the most productive U.S.-based institution (Table 5.4).

Table 5.4 : The top 15 productive institutions.

Institution	Country	Number of Publications
Institute of Geographic Sciences and Natural Resources Research	China	353
Nanjing University	China	286
Beijing Normal University	China	285
Max Planck Institute for Biogeochemistry	Germany	275
Lund University	Sweden	253
University of Helsinki	Finland	251
University of California	United States	208
California Institute of Technology	United States	196
University of Oklahoma	United States	190
Chinese Academy of Sciences	China	173
University of Chinese Academy of Sciences	China	172
Peking University	China	162
University of British Columbia	United States	160
University of New Hampshire	United States	122
University of Montana	United States	121

5.6 Bibliometric Analysis

5.6.1 World cloud analysis

Word cloud analyses provide insights into the progression of work in any field, and Figure 5.4 demonstrates how concepts are utilized in order of frequency. Upon reviewing found studies in GPP methodology, the most frequently used word (9%) is

“photosynthesis”, followed by “primary productivity” itself, “remote sensing” (7%), “carbon dioxide” (8%), and “climate change” (6%). When examining the word cloud, the presence of terms related to remote sensing, such as “MODIS”, “satellite data”, and “vegetation index”, are among the most recurrently repeated 50 words.

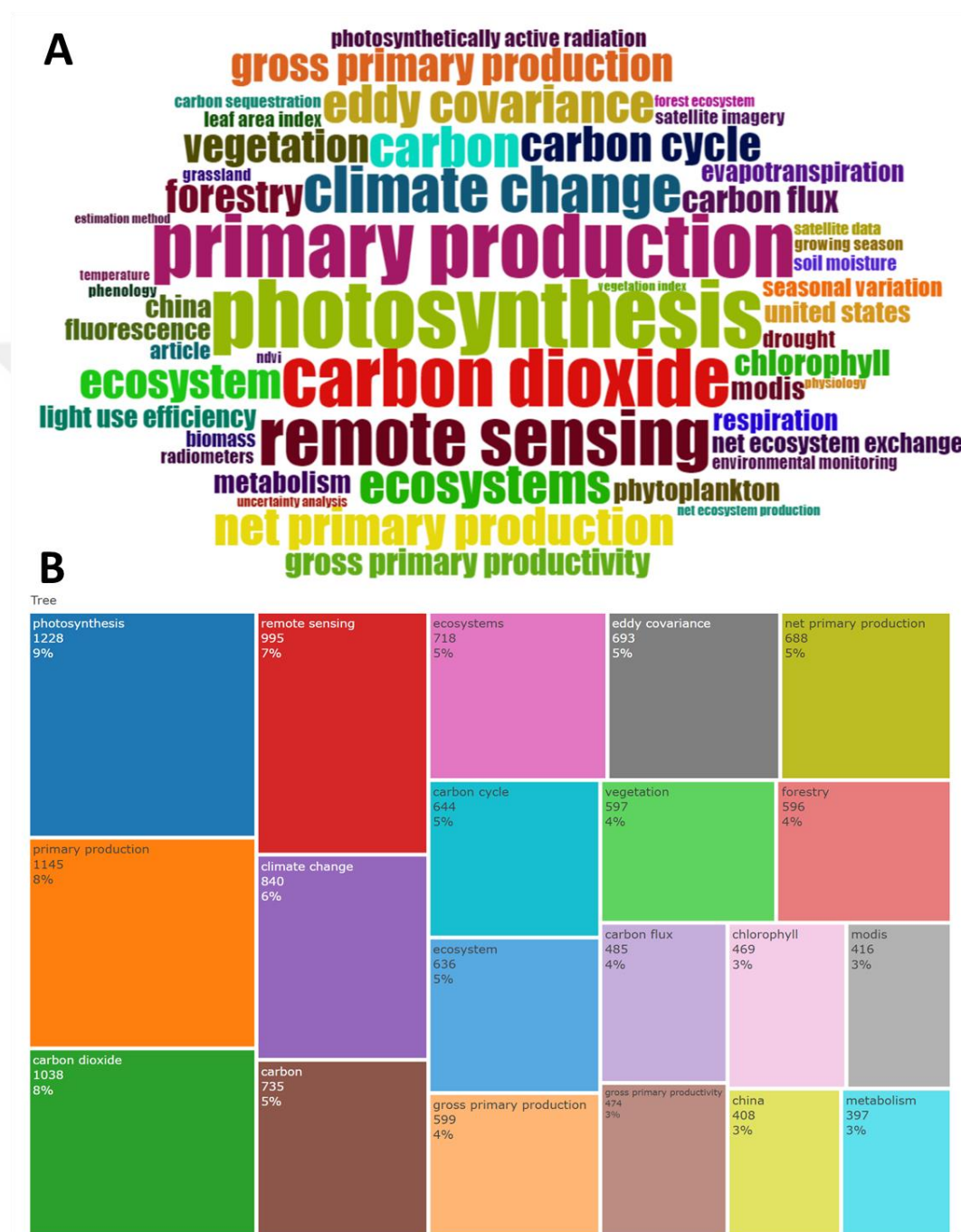


Figure 5.4 : The word cloud representing the 50 most frequently repeated words (A), the frequencies of the 20 most recurrent words, and their representative areas within the search space (B).

5.6.2 Co-occurrence network of keywords

We also conducted a network analysis among the most frequently used keywords to identify the strength of the connections between keywords and keyword clustering (Table 5.5). “gross primary productivity” is the most frequently used, followed by “eddy covariance”, then “gpp” (which of course should be clustered with “gross primary productivity”) and “remote sensing”.

Table 5.5 : The top 20 most frequently repeated words following the VOSviewer Co-Occurrence Network analysis.

Keyword	Occurrence	Total Link Strength	Keyword	Occurrence	Total Link Strength
gross primary production	445	1340	light use efficiency	120	423
gross primary productivity	343	1026	primary production	116	298
eddy covariance	317	1108	net ecosystem exchange	113	390
gpp	242	760	evapotranspiration	106	359
remote sensing	207	682	respiration	100	364
modis	193	690	drought	97	322
climate change	177	580	phenology	82	272
ecosystem respiration	151	577	ndvi	66	236
photosynthesis	147	549	solar-induced chlorophyll fluorescence	63	194
carbon cycle	131	478	water use efficiency	59	192

The structure of the keyword network is illustrated in Figure 5.5. The size of the circular shapes represents the frequency magnitude of the corresponding keyword and connections between circles provide insights into the co-usage of the words. Additionally, different colors in the graph represent different clusters formed based on the usage frequency and connections of the keywords identified using VOSviewer. Following these parameters, a network graph of 825 studies was obtained. The graph illustrates that the red cluster features terms related to vegetation and crops, such as “vegetation index” and “crop yield”. The green cluster focuses on ecology and carbon

metrics, with words like “gross primary production” and “ecosystem processes”. The yellow cluster groups together carbon and measurements, including “carbon cycle” and “fluxnet”. Lastly, the purple cluster pertains to remote sensing and technology, with “remote sensing” and “machine learning” among its terms. The keyword network analysis conducted further allows for the determination of the currency of specific keywords using the VOSviewer. In Figure 5.6, blue-colored keywords represent older terms, green-colored words are those approaching more recent dates, and yellow-colored words are the most current. Among the most “current” keywords are “machine learning”, “random forest”, “NIRv”, “sif”, “agricultural drought”, “crop yield”, and “vegetation indices”.

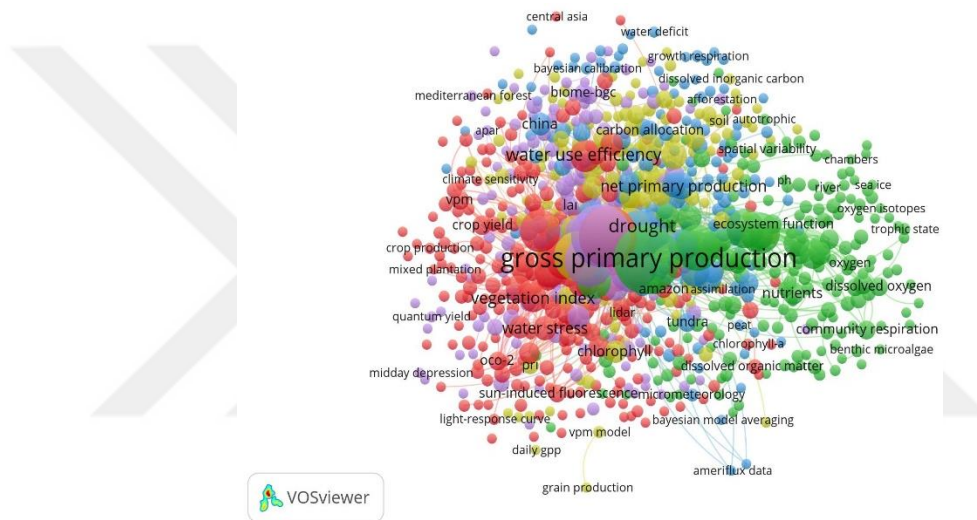


Figure 5.5 : The network visualization analysis of keywords.

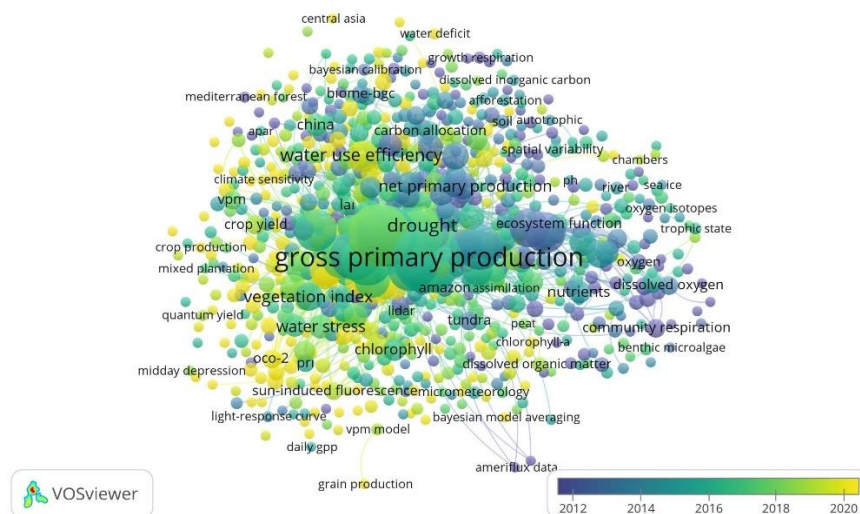


Figure 5.6 : The overlay visualization analysis of keywords.

5.6.3 Institution network analysis

Identifying collaborations among institutions is crucial to lay the groundwork for future joint efforts. Following the network analysis of the top 30 productive institutions, three significant clusters were identified (Figure 5.7). The red cluster includes institutions from Asia, highlighting primarily Chinese origin with key exceptions (e.g. the University of Oklahoma), whereas the blue cluster comprises institutions located largely in Europe. The green cluster largely contains institutions from North America. Additionally, some institutions that have strong collaborations with other clusters. Within the red cluster, Beijing and the Institute of Geographic Sciences and Natural Resources Research, both located in China, have collaborations with many centers from the other two clusters. In the green cluster, the California Institute of Technology and the University of California, and in the blue cluster, Lund University and the University of Helsinki have established numerous collaborations with centers from other clusters. This collaboration analysis highlights partnerships developed by countries in close proximity that own flux towers. The relative lack of flux tower systems in Africa and the Middle East is reflected in academic research and further emphasizes the need for more observation systems and carbon cycle science from Africa, the Middle East, Latin America, and other underrepresented regions.

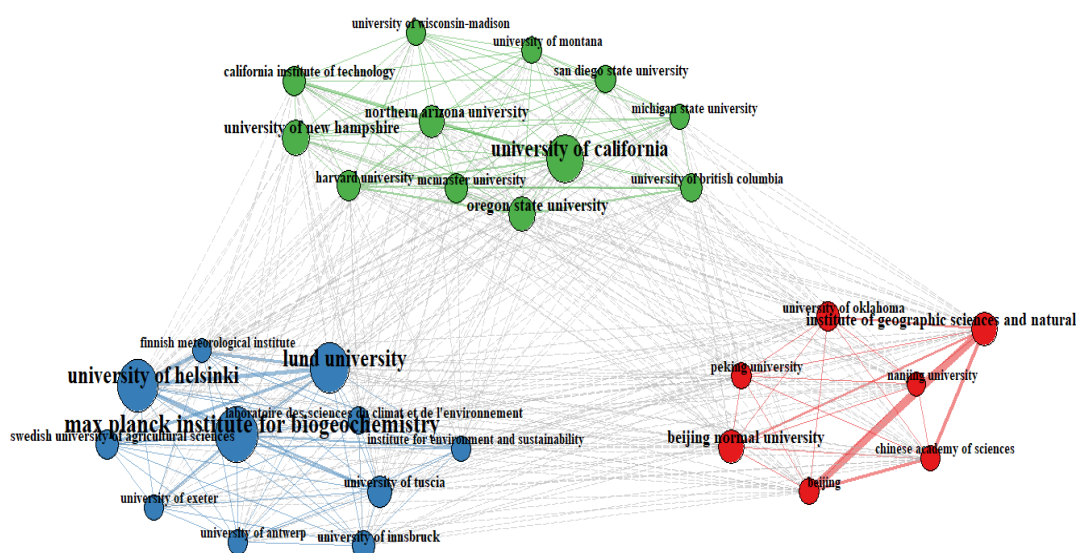


Figure 5.7 : International collaboration network.

5.6.4 Journal network analysis

Figure 5.8 represents the network among the journals through total citation counts (represented by the diameter of the circles) and their interconnections. The selection

criteria for the graphic included at least one document in each journal, with each document having received at least one citation. *Remote Sensing of Environment* has the largest circle diameter proportional to the number of total citations (TC) as listed in Table 5.3, indicating a significantly high interaction within the network. A total of 5 journal-related clusters have been formed, each containing a minimum of 35 journals. Upon examining the cluster represented in red, studies published in *Nature* stands out with the highest total number of citations, with a publication Ciais et al. (2005) receiving over 3000 citations. The blue cluster contains a larger total number of journals including *Science of the Total Environment*, *Remote Sensing*, *Biogeosciences*, *Ecological Indicators*, *Geophysical Research Letters*, and *Journal of Environmental Management*, suggesting an aggregation around earth sciences. The red cluster consists of journals specializing in agriculture and ecology including *Agricultural and Forest Meteorology*, *Agriculture, Ecosystems and Environment*, *Ecological Modelling*, and *Ecosphere*. The journal with the highest citations within the yellow cluster is *Atmospheric Chemistry and Physics* and includes *Global Biogeochemical Cycles* suggesting a chemistry cluster. Finally, the green cluster features *Limnology and Oceanography*, *Journal of Marine System*, *Journal of Sea Research*, *Hydrobiologica*, *Journal of Marine Research*, and *Freshwater Biology*, which cover diverse topics in ecosystem-water interactions.

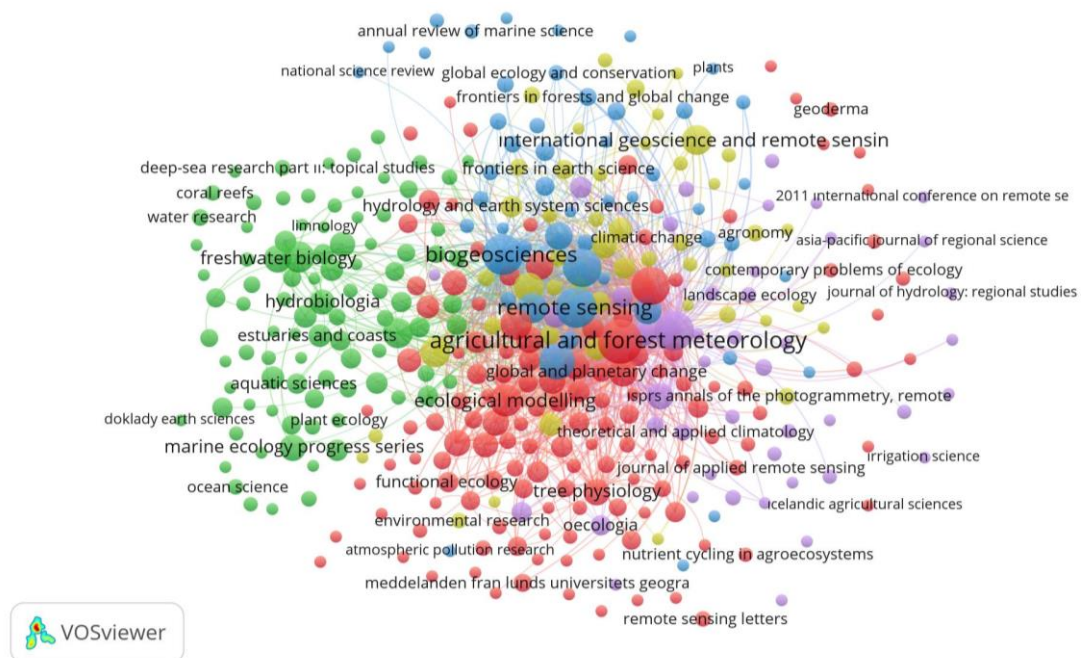


Figure 5.8 : The visualization map of journals citation network.

Figure 5.9 evaluates the average citation counts of journals involved in the citation network analysis over the years. Purple-colored journals have citations from earlier periods, whereas yellow-colored journals represent those with more recent citations, on average. The top three journals with the highest citation counts have average citation ages represented in green centered around an average publication year in 2016. *Remote Sensing* stands out both for its significant citation count and its possession of current citations, making it a likely venue for future research in this field and a potential target for prospective authors.

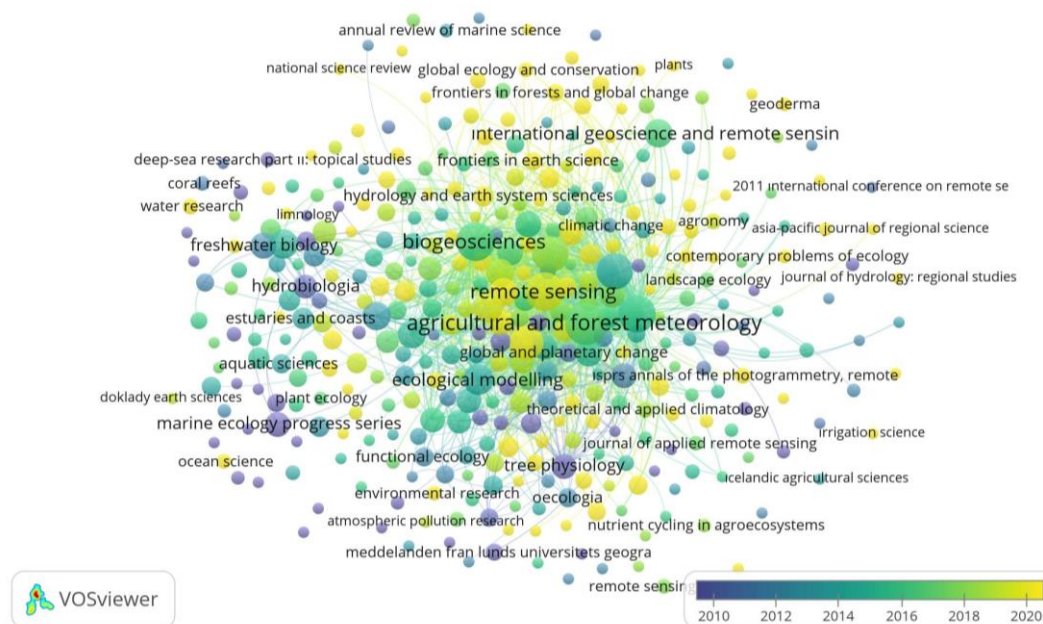


Figure 5.9 : The visualization map of journals citation overlay.

5.6.5 Thematic map

In bibliometric analyses, one of the most important tools for effectively predicting future directions from current information is “thematic mapping”. Thematic maps are divided into four different quadrants, categorizing research themes using two primary measurements: centrality and density (Figure 5.10). Centrality measures how strong the connections between clusters are, indicating a theme's importance in the research field and its level of attention. Density measures the intensity of connections within a cluster, showing how a theme has developed over time and its significance. In the upper right quadrant, “Motor Themes” have a high research volume and importance. Researchers and institutions often focus on Motor Themes due to their interests and potential impacts on the industry or discipline, representing the established and central aspects of the research area. Key terms in this quadrant include “gpp”, “remote

sensing”, “MODIS”, and “photosynthesis”. Located in the upper left quadrant, “Niche Themes” have lower research activity compared to Motor Themes. These themes may not have garnered widespread interest yet but are still important, typically characterized by specialized interest areas. The first group of keywords in this quadrant, with higher centrality and density, includes “climate change”, “primary production”, “carbon”, “temperature”. A second group of keywords in the Niche Themes, with lower centrality and intensity, includes “carbon cycle”, “drought”, “soil moisture” and “data assimilation”. “Fundamental Themes” in the lower right quadrant represent themes of importance that are in the development stage, where research interest is increasing and may become more prominent over time. Keywords identified in this quadrant are “gross primary production”, “gross primary productivity”, “eddy covariance”, “light use efficiency”, “ecosystem respiration”, “net ecosystem exchange”, “carbon flux” and “carbon dioxide”. “Emerging or Declining Themes” in the lower left quadrant are themes with low research activity that are either gaining momentum or declining. It is important for researchers to follow “Emerging Themes”, as they may become more significant in the future. Keywords in this quadrant include “respiration”, “net ecosystem production”, “metabolism”, and “ecosystem metabolism”.

Some results from the thematic map analysis are particularly striking. Among these, the emergence of the relationship between GPP and agricultural drought in the Niche Themes section stands out as an area of development. Although drought is assessed based on numerous criteria, its impact on GPP is becoming a topic of current research. When examining current literature on drought studies, Chen et al. (2023) noted that more than 70% of global regions show a time-lagged GPP response to drought for future periods. Relationships between delayed months and water balance gradient suggests that future studies would benefit from establishing long-term monitoring networks to track vegetation responses, especially as they interact with hydrology. This could help validate model predictions and enhance the understanding of vegetation ecosystem dynamics. In a study proposing preliminary definitions for the recovery phase of sudden drought, Wang et al. (2023) emphasized the importance of understanding the evolution of sudden drought and uncovering the underlying mechanism for recovery from sudden drought based on soil moisture change and GPP change. Lv et al. (2023) examined how Light Use Efficiency (LUE) methods respond

to droughts but show significant differences in both amplitude and phase for seasonal anomalies. These studies indicate the widespread use of GPP in drought research and ongoing needs to better understand the coupling between the carbon and water cycles at multiple scales in time and across global ecosystems.

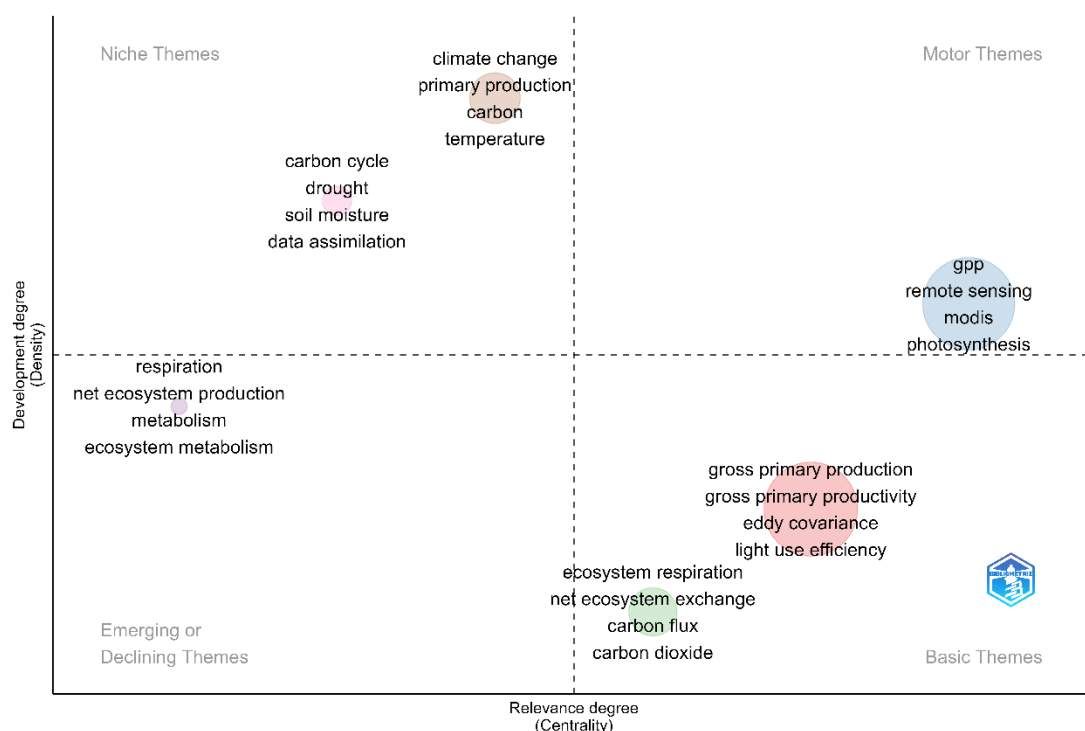


Figure 5.10 : Thematic map of author's keywords.

5.6.6 Trend topics

A simple and effective analysis to understand the growing importance of certain topics over time can be achieved through the generation of a trend topics graph. In Figure 5.11, the two most important keywords trending each year are indicated by circles. The size of these circles represents their frequency, while the horizontal lines on which the circles are placed indicate the years the keywords were influential. In the initial years covered by the search, “gross primary productivity (gpp)” and “mathematical models” were the standout keywords. The word “Asia” began to trend in 2008, influencing during the years 2006 and 2015, a phenomenon that can be attributed to the advent of and ongoing research from China. Subsequently, another significant group of keywords that emerged in 2012 were “algae” and “estimation”. By 2012, “Europe” had become a trending word, remaining influential between 2007 and 2018, similar to “China”, which trended in 2020. In 2015, the utilization of the keyword “MODIS”

showed tendency, indicating a slightly higher frequency compared to previous keywords, suggesting a focus shift towards remote sensing studies. Similarly, in 2018, “remote sensing” emerged as a trending word, possessing one of the highest frequencies among the keywords, along with “photosynthesis”, which became the most repeated keyword that year. By 2019, “climate change” entered the list of trend keywords, signifying an increase in the frequency with which current issues are addressed in climate change.

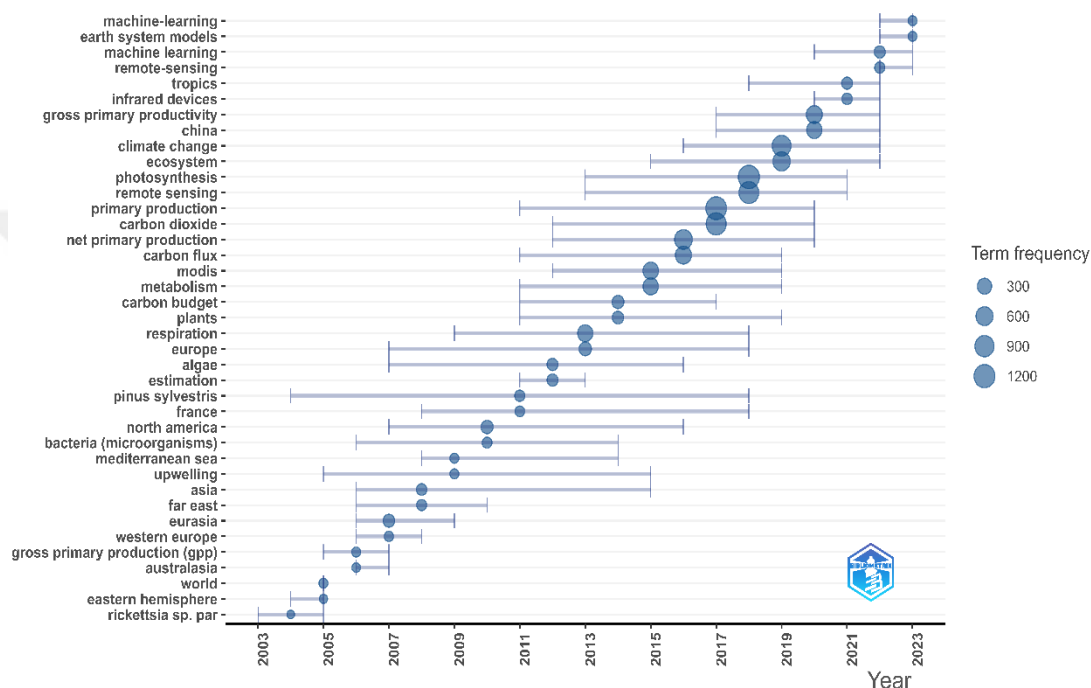


Figure 5.11 : Trend topics.

The presence of “machine learning” (ML) in 2022 and 2023 is noteworthy. Developments in artificial intelligence and improved data acquisition capabilities have led to an increase in data availability and research. (Alemohammad et al, 2017; Prakash Sarkar et al, 2022; Wang et al, 2023; Zhu et al, 2023). Alongside the development of ML methods, advancements in remote sensing systems have also occurred, with ML frequently used in conjunction with remote sensing. In the literature, after flux tower measurements, many GPP products based on ML models are among the most frequently used (Tramontana et al, 2016; Sun et al, 2019). These models seek to estimate GPP values, often at any point on the Earth's surface, without the need for expensive and maintenance-intensive eddy covariance systems. Some studies utilize these models not only for estimating GPP but also for predicting inputs used in GPP estimation. For example, Jiang et al. (2021) managed to achieve high

accuracy in GPP values by predicting photosynthetically active radiation (PAR) using an ML method.

5.7 Conclusion, Limitations And Future Research Opportunities

Estimating GPP lays a crucial foundation for many ecological studies and is critical for understanding our changing carbon cycle. Our bibliometric analysis demonstrates that machine learning and remote sensing methods continue to increase in importance for understanding GPP. However, “deep learning” and “interpretable artificial intelligence” are scarcely utilized in GPP studies to date despite ongoing innovation in Deep Learning models suitable for the utilization of large-scale data, such as Convolutional Neural Networks, Long Short-Term Memory (LSTM), and more. Deep Learning models, therefore, have significant potential in GPP estimation and researchers have begun to use interpretable artificial intelligence models to overcome the limitations of black box models. Whether black box or white box models, the body of knowledge recognizes the critical role played by divergent approaches such as Local Interpretable Model-agnostic Explanations (LIME), SHapley Additive exPlanations (SHAP), and more, facilitating the interpretation of model predictions. The use of the corresponding techniques in future studies is strongly recommended for potential researchers.

This study also shows that research may be hampered by a lack of collaboration across geographical bounds. For example, drought is a global phenomenon yet researchers tend to cluster within geographical bounds. These localized studies, when integrated with international efforts, can significantly enhance the potential for generating more comprehensive and globally applicable solutions. This synergistic approach leverages detailed insights from specific regional analyses to contribute to a broader understanding of the global patterns and mechanisms underlying the relationship between GPP and drought. Such collaborations pave the way for the creation of advanced predictive models, the formulation of resilient agricultural practices, and the implementation of sustainable water management policies. Ultimately, this integrated approach not only enriches our understanding of the Earth’s complex ecological systems but also fosters the development of innovative solutions that are vital for mitigating the impacts of climate change on a global scale.

Despite our efforts to conduct the bibliometric analysis on GPP estimation as comprehensively as possible, this study still has some limitations. Firstly, we analyzed only publications in the Scopus database, thereby missing some publications in other databases. Secondly, we analyzed only publications in English, excluding some works in languages other than English, particularly Chinese publications. Future work could therefore extend their analyses to include other databases and languages, and translation services could help international researchers obtain information from global sources. This expansion would provide a more holistic scanning alternatives and allow for the collection of more diverse information. We chose search terms and analysis parameters designed to be comprehensive but not exhaustive, and future bibliometric analyses may seek to include additional literature sources as noted as well as sensitivities to parameter selection.

In conclusion, this bibliometric analysis sheds light on the evolving environment of GPP estimation, highlighting the increasing importance of machine learning and remote sensing methods while also drawing attention to the underutilization of deep learning and interpretable artificial intelligence in current research. The integration of localized studies with international efforts suggests a promising path for developing global solutions to understand and mitigate the effects of drought on GPP.



6. SPATITEMPORAL VARIABILITY OF GROSS PRIMARY PRODUCTIVITY IN TÜRKİYE: A MULTI-SOURCE AND MULTI-METHOD ASSESMENT⁵

6.1 Introduction

Gross primary productivity (GPP) represents the carbon taken up by the biosphere that can be used for biotic processes. Understanding its changes across space and time is crucial to understand our changing planet as it is the largest component of the global carbon budget (Friedlingstein et al, 2023). GPP has been increasing across much of the globe due to in large part to global greening (Piao et al, 2019; Zhu et al, 2016) and CO₂ fertilization (Keenan et al, 2023), and is predicted to increase further throughout the 21st century (Lu et al, 2024). Simultaneously, water limitation is playing an increasing role in controlling GPP at interannual to decadal time scales (Liang et al, 2024) and often dampening its response to otherwise favorable climatic conditions (Madani et al, 2020).

Satellite remote sensing has revolutionized our ability to estimate GPP (Anav et al, 2015; Hilker et al, 2008; Song et al, 2013) and is particularly important for understanding its changes in areas where ground observations are sparse. Interpreting GPP from space can be difficult because different interpretations can result from different GPP data products (Jiang et al, 2021; Zhang and Ye, 2022; Zhou et al, 2014), and from different statistical methods for time series analysis (Huang et al, 1998; Mann, 1945; Sen, 1968; Şen, 2012). Multi-method comparisons can help reduce uncertainties, critique methods, and lend confidence that results are robust (Tucker et al, 2018). Doing so across diverse regions can further help understand how, why, and where GPP is changing, and where our interpretations of changes are insensitive to a chosen satellite product or methodological approach, thereby requiring additional study for more robust inference.

⁵ This chapter is based on the paper “Başakın, E. E., Stoy, P. C., Demirel, M. C., and Pham, Q. B. (2024). Spatiotemporal Variability of Gross Primary Productivity in Türkiye: A Multi-Source and Multi-Method Assessment, Remote Sensing, 16(11), 1994. doi: 10.3390/rs16111994.”

The calculation of trends over time for any natural phenomenon can employ parametric or non-parametric methods (or both). Among parametric methods, linear regression is the most frequently used. Despite numerous studies predicting the trend of GPP values using linear regression (Gupta et al, 2023; Liu et al, 2023; Ma et al, 2019; Moore et al, 2018), researchers tend to prefer non-parametric methods such as the Mann-Kendall, Sen's Slope, and Spearman's Rho due to the preconditions of linear regression, including the assumption of normal distributions, homoscedasticity, and the absence of multicollinearity (He et al, 2022; Keenan et al, 2023; Tian and Zhang, 2020). Certain details, which conventional methods struggle to elucidate, can be examined through graphical methods such as Innovative Trend Analysis (ITA) and Empirical Mode Decomposition (EMD). Contrary to conventional methods, ITA and EMD provide detailed information on the fluctuations of the trend over time and, in our case, the tendencies of trends at low, medium, and high GPP values. Utilizing oversimplified methods to deduce the trend in GPP values may omit significant details.

GPP varies dynamically across multiple scales in time, but studies tend to conduct trend analyses of annual GPP values. For example, the annual trend of GPP values has been investigated in studies where the entire earth is considered as the study area (Tang et al, 2015; Yang et al, 2022), in studies for China (Cai et al, 2023; Liu et al, 2023; Lv et al, 2023; Shi et al, 2024), in studies for India (Gupta et al, 2023; Sarkar et al, 2024) and in studies for Australia (Hutley et al, 2022). While this may provide acceptable results for the study at hand the trend of monthly GPP values over time is important in natural areas and especially cultivated agricultural areas to understand trends in productivity, and to investigate changes in seasonal patterns that may have management implications (O and Park, 2024). Annual GPP values represent the whole year after a single value has been obtained by averaging or summing, and do not explicitly consider variations during periods that are sensitive to plant growth including agricultural activity. The second innovation in this study, in addition to the multi-trend analysis, is the trend analysis of monthly GPP values.

Here, we compare multiple satellite-based GPP estimates and multiple approaches for time series analysis to interpret decadal-scale GPP trends, using the diverse geography and vegetation of the Republic of Türkiye as a case study. Türkiye has a relatively sparse observations from eddy covariance of surface-atmosphere carbon dioxide exchange from which GPP can be inferred (Evrendilek et al, 2011) Şaylan et al, 2018;

Yesilkoy et al, 2018), emphasizing the importance of satellite remote sensing to interpret carbon cycling dynamics (Gulbeyaz et al, 2018). It also has a relatively large and growing population and a changing climate with increases in temperature and atmospheric heat content (Stoy et al, 2022). Precipitation has been increasing in parts of the north and east and decreasing in the west (Tayanç et al, 2009; Türkeş et al, 2016), with an expansion of semi-arid regions (Selek et al, 2018) and predicted future declines in precipitation throughout the country (Demircan et al, 2017; Turkes et al, 2020). Carbon cycle processes have changed in response to changes in aridity across different parts of Türkiye (Bilgili et al, 2020; Erşahin et al, 2016), further emphasizing the importance of understanding trends in GPP. Here, we analyze multi-decadal time series from six different GPP products across the seven different regions of Türkiye, employing various trend analysis methods, both standard and novel, to interpret patterns. In this study, alongside the conventional trend method (Mann-Kendall), the simultaneous interpretation of two graphical trend methods (ITA and EMD) aims to develop an original methodology in ecology and carbon cycle studies. This approach is designed to prevent an over-simplification of trend analyses, thereby identifying trends with higher temporal resolution. In addition, we aim to provide insight into how key components of the carbon cycle of Türkiye have been changing and to provide a framework for multi-method approaches for understanding these changes.

6.2 Materials

6.2.1 Study area

Türkiye is located at the crossroads of Europe and Asia and between 36° and 42° North and 26° and 45° East. The total land area of the country is approximately 780,000 km², and elevation above sea level (asl) generally increases from west to east (Aksu et al, 2023). Four main climate types, – Mediterranean, Continental, Marmara (Transition), and Black Sea– are observed across Türkiye. The Mediterranean climate is characterized by hot and dry summers and mild and rainy winters. The Continental climate, corresponding to the Dsa category in the Köppen climate classification, is found in the mountainous parts of the Eastern Region, Southeast Region and Inner Region of Türkiye (Demir and Citakoglu, 2023), characterized by hot summers and cold winters. The Marmara region serves as an ecological and climatic intermediary, bridging the Mediterranean climate found in the south with the Continental climate of

the north. The average annual temperatures are 14.5 °C in the eastern part of the Marmara Region and 13.6 °C in the north. While summer temperatures (in July) vary around 25 °C along the coasts, these values tend to decrease in the inland areas (Biltekin et al, 2023). The Black Sea region receives relatively high precipitation throughout the year and its eastern coast experiences the most rainfall, with more than 2200 mm annually, whereas the central Anatolian region only receives about 400 mm each year (Figure 6.1b). The warmest areas are along the Mediterranean and Aegean coasts, as well as in the southern and southeastern parts of Türkiye, where average temperatures are consistently higher than in the rest of the country (Figure 6.1c). The land cover map related to the study area is provided in Figure 6.1a. Combined, these differences in climate and vegetation create unique opportunities to study nation-wide trends in GPP.

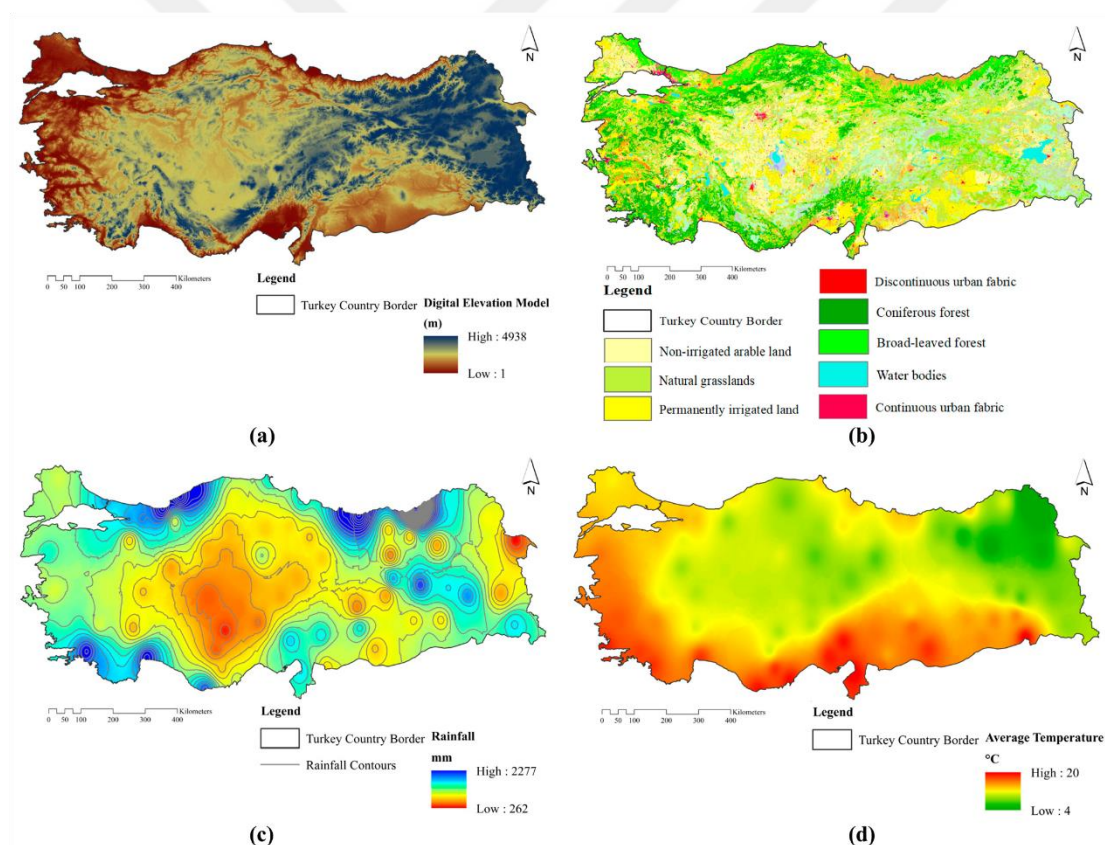


Figure 6.1 : Maps of elevation (a), annual average precipitation (1980–2020) (b), annual average temperature (1980–2020) (c), and land use/land cover (2021) (d) of Türkiye.

Türkiye is divided into seven different geographical regions – Aegean, Black Sea, Central Anatolia, Eastern Anatolia, Marmara, Mediterranean, and Southeastern Anatolia. These regions are defined based on similarities and differences in climatic

and topographic characteristics (Zeybekoğlu and Keskin, 2020). To provide a comprehensive understanding for the entirety of the country, we selected one location from each region to study trends in GPP using multiple remote sensing-based estimates. The criteria for the selection of grids from which GPP values are obtained encompass areas that are typically distant from residential areas in the regions with the most significant agricultural activity in their respective geographical locations.

6.2.2 GPP datasets

This section provides relevant information regarding the GPP datasets as well as the methods that they use to estimate GPP; additional details are provided in the Appendix A. To ascertain whether the similarity among GPP datasets can be analyzed using parametric methods, an analysis has also been conducted to determine whether these datasets are normally distributed using the Shapiro-Wilk test, with a null hypothesis that the data are normally distributed at the $\alpha < 0.05$ level. The map representing GPP values for 7 different points used as a study area is presented in Figure 6.2. In the study, due to the different temporal and spatial resolutions of the utilized datasets, monthly averages have been adopted for temporal resolution for comparison. The spatial resolution varies, with the highest being 0.05 degrees and the lowest 0.5 degrees. In the selection of spatial resolution, a lower resolution value has been chosen, and for datasets with higher resolutions, averages of 10×10 grids corresponding to 0.5×0.5 degrees in the selected area have been calculated. Descriptive information pertaining to the datasets used in the study is presented in Table 6.1.

Table 6.1 : Temporal and spatial information of GPP dataset.

Dataset	Date Range	Data Length (Months)	Units	Scale Factor	Spatial Resolution	Temporal Resolution
MODIS	02/2000–12/2020	253	kg C m ⁻² m ⁻¹	-	50 km	8 days
FLUXCOM	01/1979–12/2018	480	g C m ⁻² m ⁻¹	-	50 km	monthly
TL-LUE	01/1992–12/2020	348	g C m ⁻² m ⁻¹	0.1	5 km	monthly
PMLV2	01/1980–12/2014	420	umol m ⁻² s ⁻¹	-	5 km	monthly
GOSIF	03/2000–12/2022	274	g C m ⁻² m ⁻¹	0.01	5 km	monthly
MuSyQ	01/1981–12/2018	456	g C m ⁻² d ⁻¹	0.01	5 km	8 days

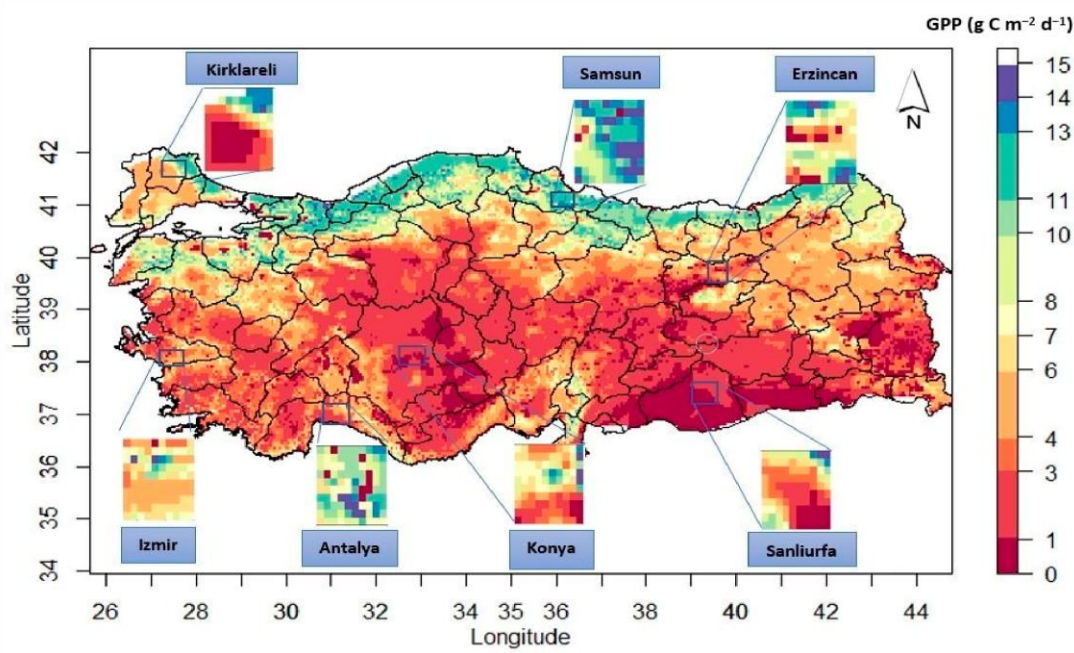


Figure 6.2 : Gross Primary Productivity across Türkiye as estimated by the TL-LUE dataset (Bi et al, 2022) for June 2020.

6.2.2.1 MODIS

The Moderate Resolution Imaging Spectroradiometer (MODIS), launched into space on the Terra satellite in 1999 and on the Aqua satellite in 2002, is designed to observe the Earth's surface, atmosphere, and oceans. One of the key functions of MODIS is to regularly measure the carbon dioxide uptake of the Earth's terrestrial vegetation through GPP products (Turner et al, 2006; Xie et al, 2023) that are generated by calculating the daily net photosynthesis values of plants and combining these values over 8-day observation intervals throughout the year at 1 km (MOD17). The MODIS GPP algorithm, developed by Running et al. (2004), is based on the Light Use Efficiency (LUE) concept originally proposed by Monteith (Monteith, 1972) and depends on the type of vegetation in a specific area and the amount of sunlight absorbed for photosynthesis. The MODIS GPP data used in the study were provided by (Kern, 2021).

6.2.2.2 MuSyQ

The Multisource Data Synergized Quantitative (MuSyQ) algorithm employs methods similar to the MODIS GPP algorithm, but it differentiates itself by using multiple remote sensing products and implementing certain improvements in the calculation of LUE by taking into account the finding that LUE for diffuse solar radiation is higher

than for direct solar radiation (Wang et al, 2021). Particularly in cloudy regions, in tropical, evergreen broadleaf forests, the accuracy of GPP predictions can be enhanced by incorporating the clearness index (CI) into the LUE estimation. The MuSyQ dataset provided by (Sun et al, 2020).

6.2.2.3 PMLV2

The Penman-Monteith-Leuning model (PML) model was initially developed by Leuning (Leuning et al, 2008) and introduced a biophysical model for surface conductance (G_s) based on the Penman-Monteith (PM) equation (Pei et al, 2020). Since then, there have been two improved versions of the PML model: PML-V1 and PML-V2. The PML-V2 employs a biophysical canopy conductance (G_c) model to combine Evapotranspiration (ET) with GPP, as described by Gan et al. (2018). To streamline the model, it retains only essential parameters that are of physiological significance for stomatal re-sponse and assimilation (Li et al, 2023). Unlike PML-V1, PML-V2 takes into account the effects of increasing Leaf Area Index (LAI), rising CO_2 levels, and climate forcing on ET and GPP (Naeem et al, 2023). PML-V2 products consist of five components: GPP, vegetation transpiration (E_c), soil evaporation (E_s), evaporation of intercepted rainfall (E_i), and the evaporation from ice, snow, and water bodies (ET_{water}) (see (Zhang et al, 2021)). The PMLV2 dataset used here was provided by Zhang et al, (2023).

6.2.2.4 TL-LUE

The two-leaf light use efficiency (TL-LUE) model calculates the GPP of sunlit and shaded leaves separately and sums them along the canopy (He et al, 2013). The revised (TL-LUE) adds an atmospheric CO_2 concentration regulation scalar and modifies the air temperature regulation scalar (Bi et al, 2022). TL-LUE GPP values were obtained from the study by (Bi and Zhou, 2022).

6.2.2.5 GOSIF

The GOSIF GPP product is generated from solar-induced chlorophyll fluorescence (SIF) data based on the relationship between SIF and GPP across various vegetation types, utilizing data from the Orbiting Carbon Observatory-2 (OCO-2), MODIS, and meteorological data from the Modern-Era Retrospective analysis for Research and Applications (MERRA-2). Developed by Li and Xiao (2019), this product has a spatial

resolution of 0.05° and offers temporal resolutions of 8 days, monthly, and annually. The GOSIF GPP dataset has been validated through comparisons with independent flux tower data, demonstrating high precision and the ability to effectively represent spatiotemporal variations in large-scale GPP (Gu et al, 2020). This validation, which considered each biome separately, confirmed the dataset's accuracy in representing photosynthetic activity, particularly in grassland ecosystems. The monthly GOSIF GPP values have been provided by Li and Xiao (2019).

6.2.2.6 FLUXCOM

FLUXCOM is an initiative formed through the collaboration of various experts, aiming to understand uncertainties in empirical scaling and to present global GPP values based on machine learning (Tramontana et al, 2016). In FLUXCOM, a variety of machine learning-based regression tools such as model tree ensembles, multiple adaptive regression splines, artificial neural networks, and kernel methods are used. The study aims to estimate CO_2 and energy fluxes using two different scaling strategies: (1) 8-day average fluxes based solely on remotely sensed data, and (2) daily average fluxes based on remotely sensed and meteorological data. Machine learning (ML) approaches have been applied and extensively cross-validated with flux towers for both setups, and it is important to note that FLUXCOM was designed to not have trends in the “RS + METEO” products used here (Jung et al, 2019). For the calibration of GPP values in the study, FLUXNET GPP values were used. The dataset has been obtained from <http://www.fluxdata.org>.

Due to the varying lengths of these time series across datasets, the start date of the latest commencing dataset, GOSIF, has been chosen as the baseline start date. Similarly, the end date of the earliest finishing dataset, PMLV2, was used as the overall end date to enable more accurate comparisons between the datasets.

6.2.3 Validation of GPP datasets

Validating a new global dataset is a challenging process, as ideal validation would involve measuring daily GPP across a diverse range of biomes and climates. An algorithm successful in a coniferous forest might be inadequate in a tropical savanna. Therefore, a global array of validation sites is necessary. More fundamentally, each available measurement for validating GPP represents a different spatial scale and presents unique limitations (Running et al, 2004). Three types of data appear useful

for validating global GPP data: direct measurements of biomass (Liao et al, 2023), tower flux measurements (Zhu et al, 2024), and measurements of atmospheric CO₂ concentrations (Joiner et al, 2018). However, each data type has limited precision or scope. Direct biomass measurements of GPP are typically measured on an annual basis (or longer), and the sample area is often less than 1 hectare (Running et al, 2004), with uncertain carbon losses due to respiration. Tower flux measurements are continuous over time and sample a footprint that may be of similar spatial extent as a 1 km pixel of satellite images, but towers only directly measure NEE such that deriving GPP requires additional modelling and subsequent assumptions. Measurements of atmospheric CO₂ concentrations integrate over a large, difficult-to-define land area and include anthropogenic CO₂ emissions and are often collected through a monthly repeat snapshot sampling process. In short, no measure matches the detail represented by GPP in terms of time, space, and ecosystem characteristics.

A network called FLUXNET, comprising hundreds of tower sites across all continents and biome types, provides accurate seasonal measurements about vegetation carbon uptake and local land CO₂ balances (Falge et al, 2002). These towers enable validation of GPP values derived from remote sensing products.

The MODIS GPP dataset has been validated using all station data in the FLUXNET network. The MuSyQ dataset was validated using measurements from stations in the BigFoot and FLUXNET networks (Wang et al, 2021). For the PMLV2 GPP dataset, 26 EC towers selected in China covering 9 PFTs were used and a Genetic Algorithm was employed during the calibration process (He et al, 2022). The TL-LUE model utilized 68 towers from the FLUXNET network for calibration and 25 for validation (Bi et al, 2022). The GOSIF dataset was developed using 91 FLUXNET sites and has been tested for 10 different biome types with satisfactory results (Li and Xiao 2019). The FLUXCOM dataset was developed using current machine learning techniques and information from 224 towers in the FLUXNET system to represent 15 different PFTs (Tramontana et al, 2016). Detailed technical information on how the GPP data sets used in the study were calculated by the developers is shared in the supplementary file.

6.3 Methods

6.3.1 Kruskal-Wallis and Mann-Whitney U

The Kruskal-Wallis test is used for comparing the median values of data obtained from different groups (Kruskal and Wallis, 1952). This non-parametric statistical method is preferred especially when the data is not normally distributed or when sample sizes are small (Pagliacci and Salpina, 2022). The Kruskal-Wallis test investigates the hypothesis that K groups of samples originate from a single statistical population or from comparable statistical populations, with respect to their means (Avand et al, 2022).

After the Kruskal-Wallis test identifies that at least one group is different, the Mann-Whitney U test can be applied for pairwise comparisons between groups. (This test is also referred to as Wilcoxon's rank sum (Shah et al, 2023). The Mann-Whitney U (MWU) test combines the data from two groups to form an overall ranking (Mann and Whitney, 1947) to jointly assess data from both groups and calculate rank sums. The test statistic is then determined based on the rank sums of both groups and the number of observations in each group and used to test whether there is a significant difference between the group medians.

6.3.2 Modified Mann-Kendall test

The Mann-Kendall test is a widely used method for detecting trends in time series data (Mann 1945). It does not account for autocorrelation (the correlation between consecutive observations in a time series), which can lead to misleading results. The Modified Mann-Kendall (MMK) test is designed to reduce the effects of autocorrelation (Hamed, 2008). This test constitutes the fundamental principles of the original Mann-Kendall test while considering the presence of autocorrelation. The specific equations of the MMK test are similar to those of the original Mann-Kendall test but include some additions and corrections to account for autocorrelation. Detailed information about MMK can be accessed at Berhanu et al. (2024).

6.3.3 Innovative trend analysis

Innovative trend analysis (ITA) is a non-parametric method used for identifying trends in time series data (Şen 2012) regardless of data distribution and length and allows for a quick graphical understanding of trends. In this method, observational data are

divided into two equal lengths starting from the beginning date and sorted from smallest to largest; one of the key advantages of the ITA method is its capability to interpret the trend behaviors of low, medium, and high-valued data (Güçlü, 2020). The first sub-series is placed on the x-axis and the second sub-series on the y-axis on the Cartesian coordinate plane. If the data fall on or very close to the 1:1 (45°) line, it is practically understood that there is no significant trend in the series being examined. If the data fall above (or below) the 1:1 (45°) line and continue to deviate from the straight line in an increasing (or decreasing) manner, the trend is monotonically increasing (or decreasing). Series that do not show a continuous increase (or decrease) at smaller values but exhibit an increasing (or decreasing) character as the values increase are referred to as series with a non-monotonic increasing (or decreasing) trend.

The steps of obtaining the confidence interval limits used to determine the significance of the trend values produced by the ITA method are described below.

The trend slope and indicator of the ITA are calculated using equations (6.1) and (6.2).

$$s = \frac{2(\bar{x} - \bar{y})}{n} \quad (6.1)$$

where s is the trend slope, n is the number of data points available, $\bar{x}$ is the mean of the first part of the series, and $\bar{y}$ is the mean of the second part of the series.

$$D = \frac{1}{n} \sum_{i=1}^n \frac{10(y_i - x_i)}{\bar{x}} \quad (6.2)$$

D is a trend indicator (a positive value indicates an increasing trend and a negative value indicates a decreasing trend), n is the number of data points in each subseries, $\bar{x}$ is the mean value of the first subseries, and x_i and y_i are the observed data values in the first and second subseries, respectively.

The significance of the ITA is estimated using a probability distribution function. The confidence interval of a standard normal distribution function with zero mean and standard deviation of the ITA trend slope (σ_s) at the significance level α is calculated using the confidence limit of the trend slope (CL):

$$CL_{(1-\alpha)} = 0 \mp s_{\sigma_s} \quad (6.3)$$

When the estimated ITA trend slope (i.e., s) is greater than the critical value, the null hypothesis of no slope is rejected. Therefore, if the data points fall outside the specified confidence limit, s calculated from the data series is considered statistically significant. In this study, a 95% confidence level is considered. An example of a graphical representation of the ITA method is presented in Figure 3.3.

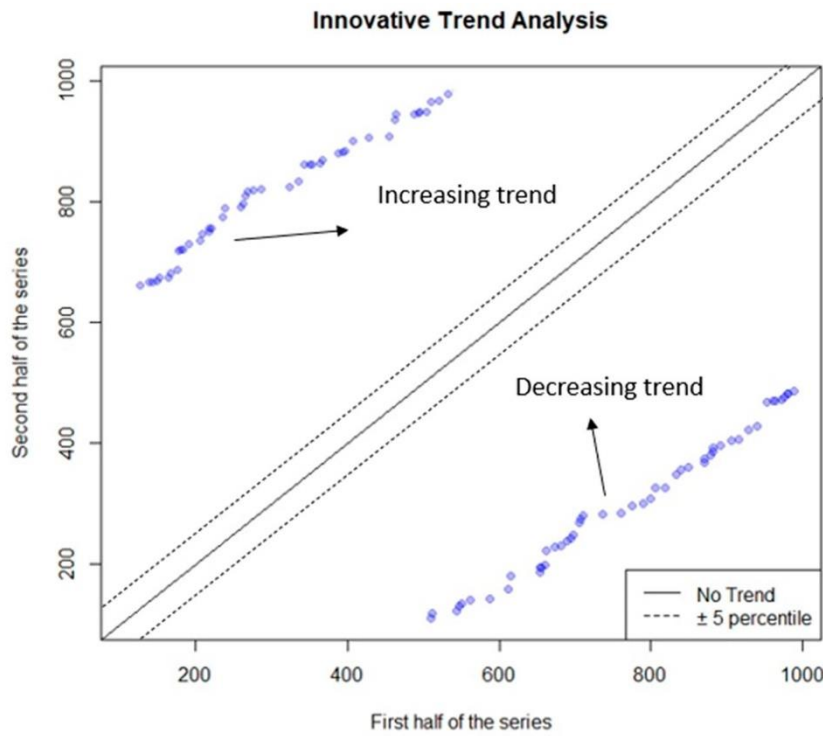


Figure 6.3 : ITA example.

Among the areas where the ITA method is most frequently used is the examination of hydrological and meteorological variables. In this context, considering the strong link between carbon cycles and water cycles (Yu et al, 2024), we seek to investigate the utility of ITA GPP trend analysis. For the ITA analysis, the trendchange package (Patakamuri and Das, 2022) in R was utilized (R Core Team 2023).

6.3.4 Empirical mode decomposition

Empirical Mode Decomposition (EMD) is a signal processing method that aims to uncover the internal structure of data by decomposing a signal into its sub-components at different frequencies (Huang et al, 1998). EMD has been used for various

applications, such as analyzing time series data, identifying periodic or continuous structures in data, and decomposing signals in the frequency domain. These components are called Intrinsic Mode Functions (IMFs). Each IMF represents the local frequency modulation of the data and reflects the time–variance relationship in the signal. IMFs preserve the intrinsic characteristics of the data while determining their fundamental frequencies. EMD decomposes a signal into IMFs through a series of progressive steps. For a signal component to be considered as an IMF, it must satisfy two conditions: (i) the number of local maxima and minima and the number of zero crossings must be the same or differ at most by one and (ii) the arithmetic mean of the envelope defined by the local maxima and the envelope defined by the local minima must be zero. The sum of the IMF and residual mode components, determined through the decomposition and elimination processes of a signal by EMD, reconstructs the original signal. EMD calculation steps are defined as follows.

Step 1: All extrema in the time series are identified. The local maxima and minima are then used for interpolation and the upper $U_x(t)$ and lower $L_x(t)$ envelopes are derived using cubic spline.

Step 2: The mean envelope value $m(t)$ and the detailed component $d(t)$ are calculated.

$$m(t) = \frac{U_x(t) + L_x(t)}{2} \quad (6.4)$$

$$d(t) = x(t) - m(t) \quad (6.5)$$

Step 3: If the resulting $d(t)$ meets the stopping criteria, it becomes the first IMF. Otherwise, the procedure is repeated on $d(t)$ until the stopping criteria are met, where the stopping criteria are defined as follows:

$$\sum_{t=1}^T \frac{|d_j(t) - d_{j-1}(t)|^2}{d_{j-1}^2(t)} \leq SD \quad (6.6)$$

where T represents the length of the time series and j denotes the number of iterative calculations. The default value of SD is usually set between 0.2 and 0.3. Thus, in this study, the default SD value was set to 0.2.

Step 4: Continue iterating steps two to three until all IMFs and the detailed signal are obtained. Finally, the original time series $f(t)$ can be decomposed as given below:

$$f(t) = \sum_i^N IMF_i(t) + R_N(t) \quad (6.7)$$

Where $R_N(t)$ is the final residual and can be considered as the global trend signal.

The residual band obtained through the EMD method is commonly referred to as the long-term trend component (Souza et al, 2022), which identifies changes in the nonlinear trend present in the time series over time. In the literature, its application is noted in various fields such as precipitation (Guo et al, 2016), air temperature (Liu et al, 2016), extreme sea levels (Lee, 2013), and regional drought (Adarsh and Janga Reddy, 2019).

We decomposed the GPP time series into sub-series with different frequencies using the EMD method. This decomposition process utilized the EMD library (Kim and Oh, 2009, 2021) in R. While the EMD algorithm performs a data-based decomposition, it still contains certain hyperparameters including the maximum number of siftings and the maximum number of IMFs. The maximum number of siftings refers to the maximum iterations the algorithm will perform during the extraction of each IMF, which influences how detailed the algorithm will process each IMF and how long it will run. A larger maximum sifting number can lead to longer extraction processes for each IMF. This can create a significant computational cost, especially for large datasets. Excessive sifting may sometimes result in over-sifting of the signal, causing the IMFs to be overly detailed and fail to reflect the main signal's characteristics (Jia et al, 2023). Depending on the specific application of the algorithm and the characteristics of the dataset, the optimal value of this hyperparameter can vary. In practical applications, it is typically adjusted through trial and error, and we did so by choosing values from 10 to 30. In the EMD algorithm, the “maximum number IMFs” setting determines the maximum number of IMFs the algorithm will extract from a signal. This setting is a crucial hyperparameter that directly affects how EMD works and its outcomes. The maximum number of IMFs was also set using a trial-and-error method. In this study, the maximum number of IMFs varied between 4 and 8. After extracting the IMFs and the trend component, the variance contribution rate (VCR) values of all the sub-series were calculated. This value measures the extent to which

each sub-series explains the variance of the original time series. Although simple and practical, the EMD method also has some disadvantages. For example, EMD-based techniques are sensitive to the noise contained in the signal. A small change in the signal-to-noise ratio of the signal can lead to significantly different signal decomposition results (Yang et al, 2017). EMD has also mixing problem (Santhosh et al, 2018), which is caused by the mixing of frequency components between various IMFs. These issues should be taken into account when using the EMD algorithm.

6.4 Study Design

The flowchart of the study is illustrated in Figure 6.4. First, we obtained global Gross Primary Production (GPP) datasets from various sources. These datasets were carefully selected to ensure a comprehensive representation of GPP values across different geographical locations. These datasets were then subjected to a sub-setting process in which data corresponding to specific points in seven different regions in Türkiye were extracted for a more localized analysis of GPP. Given the concern that GPP values examined at the highest spatial resolution could potentially be misleading and may not accurately reflect larger-scale variability, 50×50 km areas were defined within each region. For each of these defined areas, average GPP values were calculated. Once the time series data for each region and the corresponding dataset were obtained, a series of statistical analyses were performed. First, correlation analyses were conducted to determine the degree of similarity and strong relationships between data within the same region. Non-parametric statistical tests, specifically the Kruskal–Wallis (KW) test and the Mann–Whitney U (MWU) test, were used to evaluate differences between the datasets. Following the identification of similarities and differences through these statistical tests, temporal trends in GPP values were analyzed using three different methods. These methods were chosen to comprehensively capture and analyze temporal dynamics and trends in GPP values over time. Each method provided unique insights into the temporal evolution of GPP.

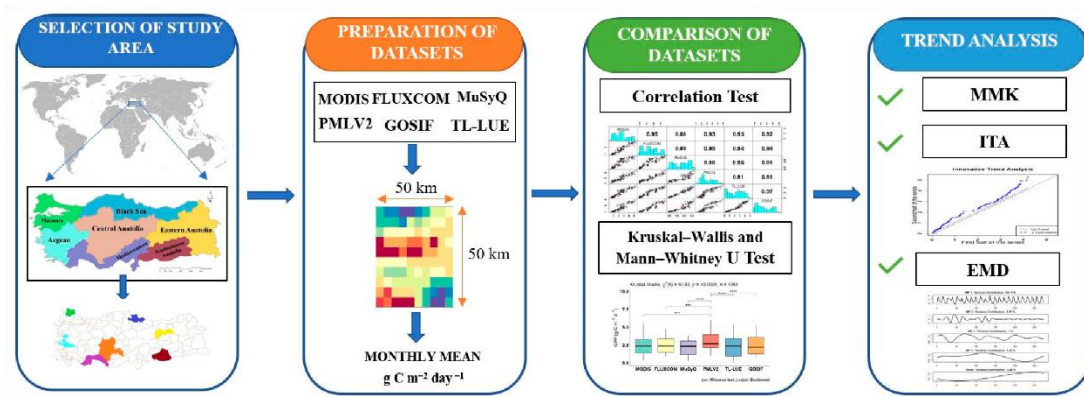


Figure 6.4 : Flowchart of the study.

6.5 Results

6.5.1 Statistical evaluation of the GPP datasets

None of the GPP product data follow a normal distribution across the study regions. The time series of MODIS GPP for all regions, and histograms related to these time series, are shared in Figure A.3. The average GPP values for Antalya, Erzincan, Izmir, Kırklareli, Konya, Samsun, and Sanliurfa are 2.53, 1.29, 2.51, 2.52, 0.97, 3.27, and 1 $\text{g C m}^{-2} \text{ d}^{-1}$, respectively. The lowest GPP value, 0 $\text{g C m}^{-2} \text{ d}^{-1}$, was found in the Erzincan region, while the highest GPP value, 8.72 $\text{g C m}^{-2} \text{ d}^{-1}$, was observed in the Samsun region. The time series of MuSyQ GPP for all regions, and histograms associated with these time series, are presented in Figure A.4. The average GPP values from MuSyQ were as follows: Antalya 2.04, Erzincan 1.13, Izmir 3.43, Kırklareli 2.51, Konya 0.68, Samsun 3.00, and Sanliurfa 1.05 $\text{g C m}^{-2} \text{ d}^{-1}$. The lowest GPP value was 0.03, observed in the Erzincan region, while the highest GPP value was 8.23 $\text{g C m}^{-2} \text{ d}^{-1}$, which was observed in the Samsun region. Distribution graphs and time series graphs for the study areas are presented in Figure A.5. The lowest GPP estimates from the PMLV2 were 0.11 $\text{g C m}^{-2} \text{ d}^{-1}$ and 0.33 $\text{g C m}^{-2} \text{ d}^{-1}$, recorded in the regions of Erzincan and Konya, respectively, while the maximum GPP values recorded were between 7.27 $\text{g C m}^{-2} \text{ d}^{-1}$ and 9.04 $\text{g C m}^{-2} \text{ d}^{-1}$, observed in the Antalya and Samsun regions, respectively (Figure A.5). Recent years have seen notably high values in the Antalya, Erzincan, Izmir, and Konya regions; the increase in the frequency of high values in recent years can be observed from the time series graphs (Figure A.6). In the TL-LUE GPP datasets, the highest values were detected at 11.93 $\text{g C m}^{-2} \text{ d}^{-1}$ in Samsun and 8.51 $\text{g C m}^{-2} \text{ d}^{-1}$ in Kırklareli. The lowest GPP from GOSIF values were identified in Erzincan, while the highest were observed in Samsun (Figure A.7).

Similar to MuSyQ GPP values, almost all of the GOSIF GPP values also exhibited a distribution that was heavily right-skewed (e.g., Figure 6.5). The lowest values from FLUXCOM occurred in Erzincan, while the highest values were identified in Samsun. The FLUXCOM GPP average values for Antalya, Erzincan, Izmir, Kirklareli, Konya, Samsun, and Sanliurfa were 2.49, 1.34, 3.39, 2.07, 0.85, 2.79, and 1.31 g C m⁻² d⁻¹, respectively (Figure A.8).

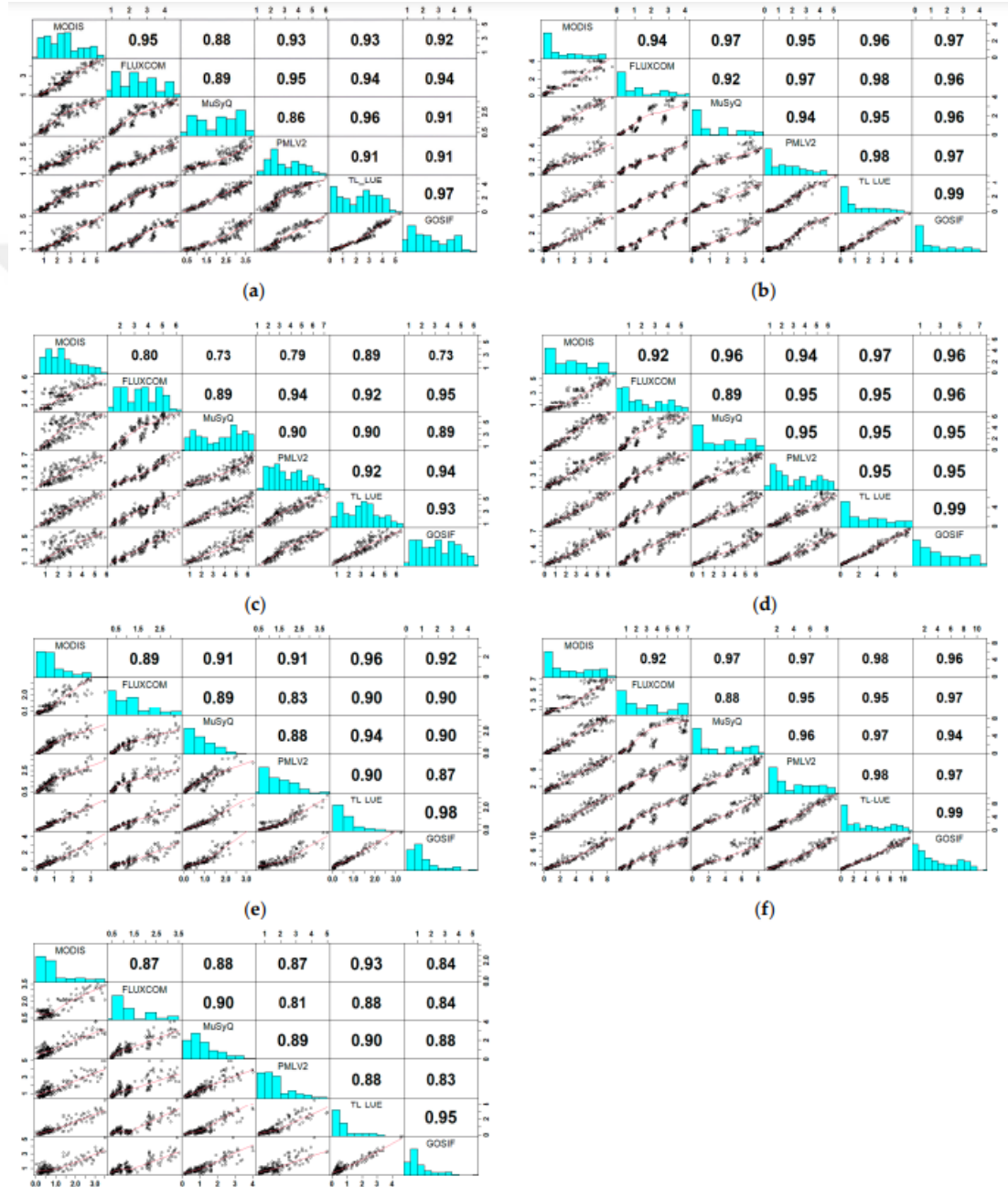


Figure 6.5 : Correlation matrices of GPP datasets from the (a) Antalya, (b) Erzincan, (c) Izmir, (d) Kirklareli, (e) Konya, (f) Samsun, and (g) Sanliurfa regions.

First, correlation coefficient values were calculated to determine relationships between different GPP products from each data point from different regions (Figure 6.5). All

relationships were statistically significant, with important differences in the strength of the agreement. For instance, in the Antalya region, the highest correlation of 0.97 was observed between GOSIF and TL-LUE, while the lowest was 0.88 between MODIS and MuSyQ datasets. The scatter plots illustrate that variance tended to be lower for lower GPP values and vice versa for higher GPP values. A noticeable increase in variance can be observed between FLUXCOM and MuSyQ, and MuSyQ and GOSIF for high GPP values. In the case of the Erzincan region, the highest correlation was 0.99 between GOSIF and TL-LUE, and the lowest was 0.92 between FLUXCOM and MuSyQ datasets. Scatter diagrams for the Erzincan region indicate a relatively lower variance at low values compared to medium and high values. Regarding the Kırklareli region, the highest correlation coefficient of 0.99 was measured between GOSSIF and TL-LUE, while the lowest correlation coefficient of 0.89 was observed between FLUXCOM and MuSyQ. For the İzmir region, the higher variance can be observed for medium and high values compared to lower values. When examining scatter plots for the Konya region, the variance between low and medium values is significantly lower compared to very high values. The highest correlation coefficient in Konya was 0.98 between GOSIF and TL-LUE, while the lowest was 0.83 between PMLV2 and FLUXCOM. In the Samsun region, an almost perfect scatter was observed between.

TL-LUE and GOSIF, resulting in a correlation coefficient of 0.99. The lowest correlation, at 0.88, occurred between MuSyQ and FLUXCOM. When examining scatter plots between FLUXCOM and other datasets, a noticeable difference is apparent. Finally, for the Sanliurfa region, the highest correlation coefficient of 0.95 was between TL-LUE and GOSIF, while the lowest correlation was 0.84, both between GOSIF and MODIS and GOSIF and FLUXCOM. Low variance can be observed among low values, but variance noticeably increased for medium and high values.

6.5.2 Kruskal-Wallis and Mann-Whitney U test results

Upon examining the KW test results, the KW test statistic values for all regions were calculated as $\chi^2 = 50.8$, $\chi^2 = 21.9$, $\chi^2 = 76.9$, $\chi^2 = 42.6$, $\chi^2 = 149$, $\chi^2 = 34.5$, and $\chi^2 = 168$ ($p < 0.05$), respectively: at least one group among the GPP estimation methods

differed significantly from the others. Following the KW test, the MWU test was applied to test which groups had significant differences.

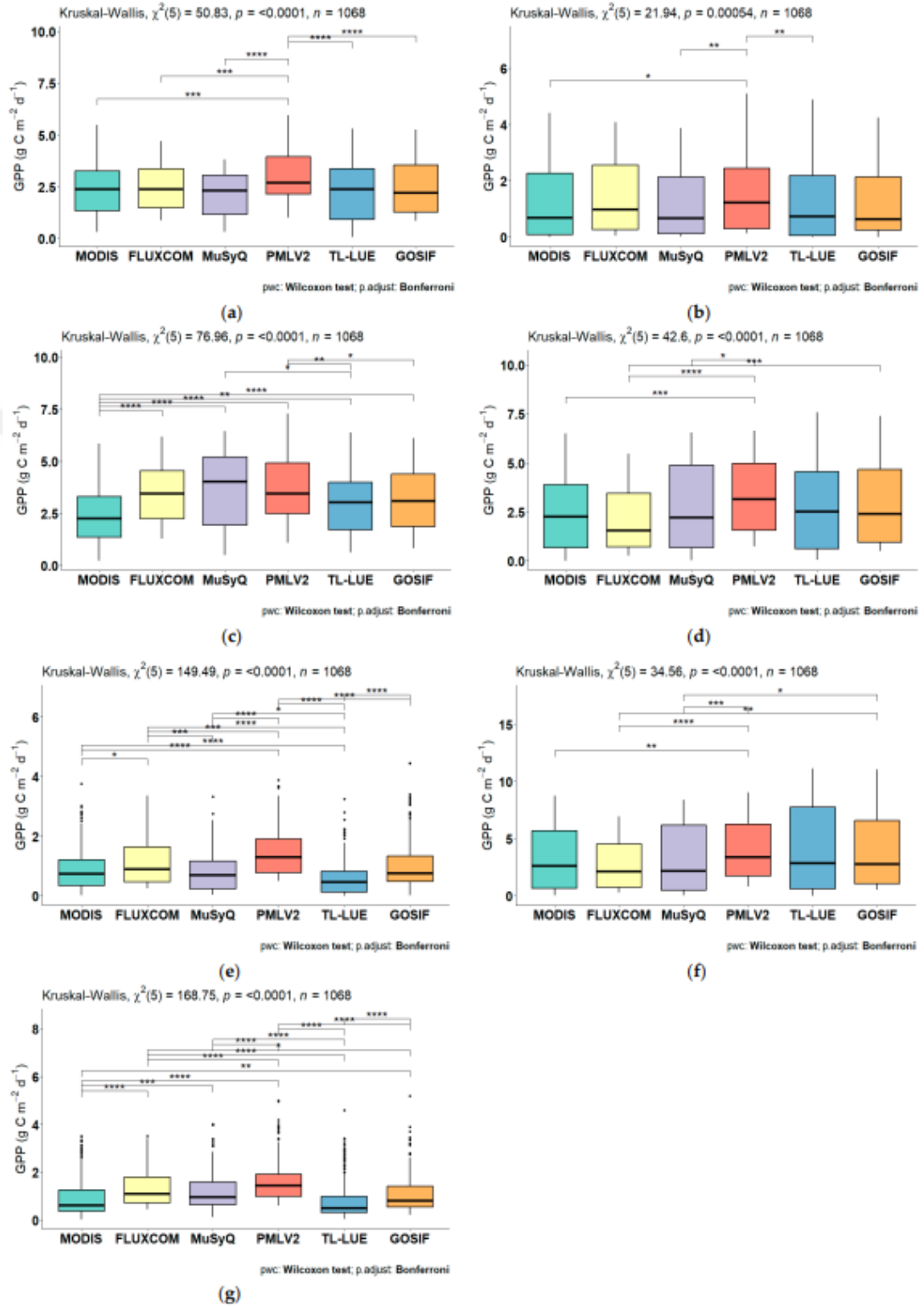


Figure 6.6 : GPP dataset similarity test results for the (a) Antalya, (b) Erzincan, (c) Izmir, (d) Kırklareli, (e) Konya, (f) Samsun, and (g) Sanliurfa regions of Türkiye (**** $p < 0.0001$, *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$).

This test enabled pairwise comparisons and groups differing from each other were tested at different levels of significance. Due to the number of groups evaluated simultaneously being more than two, the Bonferroni correction was applied to prevent an increase in Type I errors during the MWU test. The KW and MWU test results for all regions are presented, comparatively, using box plot graphs in Figure 6.6. The quartile information of the distributions of the GPP datasets can be read on the box plot diagrams, and the data groups that are significantly different on the graph are marked with an asterisk. A significant difference was found between the PMLV2 and all other datasets in Antalya. For the Erzincan region, PMLV2 showed significant differences compared to MODIS, MuSyQ, and TL-LUE. In the Izmir region, the MODIS dataset was significantly different from all other datasets, and significant differences were also found between MuSyQ-TL-LUE, PMLV2-TL-LUE, and PMLV2-GOSIF. In the Kırklareli region, the pairs identified with significant differences, in order, were MODIS-PMLV2, FLUXCOM-PMLV2, MuSyQ-PMLV2, and FLUXCOM-GOSIF. In the Konya region, nearly all datasets showed significant differences, with only MODIS-MuSyQ, MODIS-GOSIF, FLUXCOM-GOSIF, and MuSyQ-GOSIF appearing similar. In the Samsun region, MODIS-PMLV2, FLUXCOM-PMLV2, FLUXCOM-GOSIF, and MuSyQ-GOSIF were significantly different. In the Sanliurfa region, similar to Konya, almost all datasets were significantly different, except for MODIS-TL-LUE, FLUXCOM-MuSyQ, and MuSyQ-GOSIF.

6.5.3 Modified Mann-Kendall results

The present study involves the temporal analysis of GPP values, employing Mann–Kendall (MK) tests. These non-parametric tests provide insights into the direction of trends in time series data. The purpose of using the Modified Mann–Kendall (MMK) test in this study is to address the issue of autocorrelation, which can lead to inaccurate results in datasets with high autocorrelation when using the MK test alone. The MMK test attempts to eliminate internal dependency, aiming for statistically more reliable results. The monthly trend results are presented in Table 6.2. The values pertaining to the Antalya region have been interpreted in detail as an example.

In the analyses conducted on a monthly basis, no significant trend was observed in any dataset for January in the Antalya region. In February, MODIS indicated a decrease in

GPP, while other datasets (except PMLV2) showed an increase. In March, no trend was observed in the GOSIF dataset, while MODIS showed a decrease, and other datasets indicated an increase. In April, MODIS indicated a decrease in GPP, FLUXCOM showed no trend, and other datasets showed an increase. In May, MODIS again showed a decrease, FLUXCOM had no trend, and other datasets exhibited a significant increase. In June, there was no trend in FLUXCOM and MuSyQ, while other datasets showed an increase. The same was observed in July and August. In September, no significant trend was detected in the FLUXCOM, MODIS, and PMLV2, while other datasets showed an increasing trend. In October, FLUXCOM and PMLV2 showed no significant trend, while other datasets continued to show an increasing trend. In November, all datasets except FLUXCOM showed an increasing trend and, in December, all datasets except MuSyQ showed an increasing trend.

The TL-LUE dataset shows no trend during the winter months, while it displays an increasing trend in all regions during the spring and summer months. The GOSIF dataset generally exhibits an increasing trend and MODIS, distinct from other datasets, shows a tendency to decrease in many regions during the months of March, April, and May. MuSyQ shows no trend in January and December, while exhibiting an increasing trend in all other months. PMLV2 shows no trend during the winter months but an increasing trend during the spring and summer months. Monthly trend values for all regions are shared in the Appendix A (Tables A.1–A.7).

After conducting an analysis of the monthly values, the datasets were examined using the MMK test. According to the results presented in Table 6.3, significant increasing trends were identified in TL-LUE, MuSyQ, and PMLV2 for the Antalya region. In the Erzincan region, a notable increasing trend was detected exclusively in MuSyQ. In Izmir, significant trends were observed in TL-LUE, GOSIF, and PMLV2. For the Kırklareli region, statistically significant increasing trends were found only in TL-LUE and MuSyQ. In Konya, an upward tendency was identified in TL-LUE, GOSIF, MuSyQ, and PMLV2. In Samsun, increasing trends were observed in MuSyQ and PMLV2. Lastly, in the Sanliurfa region, significant increases were detected in all datasets except for FLUXCOM and MODIS. MuSyQ showed a strong and statistically significant upward trend in the Kırklareli, Konya, Samsun, and Sanliurfa regions with a 99% confidence interval.

Table 6.2 : Monthly MMK trend analysis results.

Region	Data sets	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Antalya	TL-LUE	↔	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑
	FLUXCOM	↔	↑	↑	↔	↔	↔	↔	↔	↔	↔	↔	↑
	GOFIS	↔	↑	↔	↑	↑	↑	↑	↑	↑	↑	↑	↑
	MODIS	↔	↓	↓	↓	↓	↑	↑	↑	↔	↑	↑	↑
	MuSyQ	↔	↑	↑	↑	↑	↑	↔	↔	↔	↑	↑	↔
	PMLV2	↔	↔	↑	↑	↑	↑	↑	↑	↑	↔	↑	↑
Erzincan	TL-LUE	↔	↔	↑	↑	↑	↑	↑	↔	↔	↔	↔	↔
	FLUXCOM	↔	↔	↔	↔	↔	↔	↔	↔	↔	↔	↔	↔
	GOFIS	↔	↔	↔	↔	↑	↑	↔	↔	↑	↑	↑	↔
	MODIS	↔	↓	↓	↓	↓	↓	↑	↑	↑	↑	↑	↑
	MuSyQ	↔	↑	↑	↔	↑	↑	↑	↑	↑	↑	↑	↔
	PMLV2	↔	↔	↑	↑	↑	↑	↑	↑	↑	↑	↑	↔
Izmir	TL-LUE	↔	↑	↑	↑	↑	↑	↑	↔	↔	↔	↑	↑
	FLUXCOM	↔	↔	↔	↔	↔	↔	↔	↓	↔	↑	↑	↑
	GOFIS	↔	↑	↑	↑	↔	↑	↑	↑	↔	↔	↔	↔
	MODIS	↔	↓	↓	↓	↔	↑	↑	↔	↓	↔	↑	↑
	MuSyQ	↑	↑	↑	↑	↑	↑	↔	↔	↔	↑	↑	↑
	PMLV2	↔	↔	↑	↑	↑	↑	↑	↔	↑	↑	↑	↑
Kirklareli	TL-LUE	↔	↔	↑	↑	↑	↑	↑	↔	↔	↔	↑	↑
	FLUXCOM	↔	↑	↑	↔	↔	↔	↔	↔	↔	↑	↔	↔
	GOFIS	↔	↑	↔	↔	↑	↑	↑	↑	↔	↔	↔	↔
	MODIS	↔	↔	↓	↓	↓	↔	↑	↑	↑	↑	↑	↑
	MuSyQ	↔	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↔
	PMLV2	↔	↔	↔	↑	↑	↑	↑	↑	↑	↑	↑	↔
Konya	TL-LUE	↔	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↔
	FLUXCOM	↔	↔	↔	↔	↔	↔	↔	↔	↔	↔	↔	↔
	GOFIS	↔	↔	↔	↔	↔	↔	↑	↑	↑	↑	↔	↔
	MODIS	↔	↔	↔	↓	↔	↑	↑	↑	↑	↑	↑	↑
	MuSyQ	↔	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↔
	PMLV2	↔	↔	↑	↑	↑	↑	↑	↑	↑	↔	↑	↔
Samsun	TL-LUE	↔	↑	↑	↑	↑	↑	↑	↔	↔	↔	↑	↔
	FLUXCOM	↔	↔	↔	↔	↔	↔	↔	↔	↔	↔	↔	↔
	GOFIS	↔	↑	↔	↔	↑	↑	↑	↑	↔	↔	↔	↑
	MODIS	↔	↓	↓	↓	↓	↓	↑	↑	↑	↑	↑	↑
	MuSyQ	↔	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↔
	PMLV2	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑
Sanliurfa	TL-LUE	↔	↑	↑	↑	↑	↔	↑	↑	↑	↑	↑	↔
	FLUXCOM	↑	↑	↑	↓	↔	↔	↓	↔	↓	↔	↔	↔
	GOFIS	↔	↔	↔	↔	↔	↑	↑	↑	↑	↑	↔	↔
	MODIS	↔	↔	↓	↓	↑	↑	↑	↔	↔	↔	↑	↑
	MuSyQ	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑
	PMLV2	↔	↔	↑	↑	↔	↓	↑	↑	↑	↔	↑	↔

6.5.4 Innovative trend analysis results

The ITA results for the GPP values for all regions are shared in Figure 6.7. Examining the results for the Antalya region, a monotonic increasing trend can be observed throughout the TL-LUE dataset, with distinct increasing trends in both low and high GPP values. The FLUXCOM dataset is scattered along the trendless 1:1 line, indicating an absence of trend. In the GOSIF dataset, while no clear trend is evident, there is a slight increasing trend in medium GPP values. The MODIS dataset nearly shows a trendless pattern; however, an increasing trend is noticeable at higher GPPs values. In the MuSyQ dataset, while low and medium values show an increasing trend, high values are distributed near the trendless region. In the PMLV2 dataset, there is no trend in low GPP values and an increasing trend as GPP becomes higher. In the Erzincan region, except for the FLUXCOM dataset, all datasets exhibit a notable monotonically increasing GPP trend that is more pronounced at high values. Regarding the ITA results for the Izmir region, no trend is observed in the low and medium GPP values except for the MuSyQ, which increase towards medium values with a smaller increase in high values and some higher values showing a decreasing trend. TL-LUE, GOSIF, and PMLV2 also exhibit an increasing trend at high GPP values, while MODIS shows slight decreasing trends in higher values. In Izmir, a monotonic increasing trend is observed in TL-LUE, GOSIF, and PMLV2, while MODIS and MuSyQ display a non-monotonic, increasing trend at medium values but trendless behavior at low and high values. The ITA results for the Konya region display a markedly different trend compared to other regions. Except for the FLUXCOM and GOSIF datasets, all other datasets show a monotonically increasing trend. In the GOSIF dataset, while the medium values show a clear increasing trend, its high values are scattered in the trendless region. In the PMLV2 dataset, very pronounced increases are observed at high GPP values.

In examining the GPP values in the Samsun region, a distinct monotonic increasing trend can be observed in MuSyQ and PMLV2. The TL-LUE dataset shows an increasing trend at high values, and an increasing trend at medium values is apparent in MODIS and GOSIF. In Sanliurfa, a notably different trend distribution is present compared to other regions. In the GOSIF dataset, an increasing trend in low values slightly diminishes in medium values but increases again in high values. The common point in the trend analyses across all regions is the tendency for increases at high

values. While there are clear differences between the trends in the Sanliurfa and Konya regions, the trend movements in other regions are similar. To provide a more detailed analysis, the trend results obtained with the ITA method are shared in the appendix for each time series (Figure A.9).

Table 6.3 : Modified Mann–Kendall trend analysis results for GPP estimates from different remote sensing data products from selected areas in different regions of Türkiye.

Data set		Antalya	Erzincan	Izmir	Kirklareli	Konya	Samsun	Sanliurfa
TL-LUE	Z	3.16*	1.17	2.61*	2*	3.03*	1.58	3.68**
	p	.001	.23	.008	.044	.002	.11	.0002
	Slope	0.0021	0.00005	0.002	0.0015	0.0005	0.001	0.0007
FLUXCOM	Z	0.61	0.24	0.36	0.21	-1.7	-0.98	-0.48
	p	.539	.8	.71	.82	.86	0.92	.62
	Slope	0.0001	0.00002	0.00004	0.00003	-	-	-0.0005
GOFIS	Z	1.9	1.05	2.13*	1.41	2.6*	1.37	4.69**
	p	.056	.29	.032	.15	.009	.16	.000
	Slope	0.0013	0.0004	0.002	0.0014	0.0012	0.0013	0.0021
MODIS	Z	1.38	0.95	0.8	1.07	1.73	0.9	1.73
	p	.166	.34	.42	.28	.08	.36	.08
	Slope	0.0015	0.0003	0.0008	0.0013	0.0008	0.0013	0.0006
MuSyQ	Z	2.07*	2.09*	0.29	3.57**	5.28**	2.27*	6.66**
	p	.038	.035	.29	.0003	.000	.022	.000
	Slope	0.0006	0.0002	0.001	0.0019	0.0008	0.001	0.0014
PMLV2	Z	3.08*	1.75	2.74*	1.93	3.44**	2.66*	2.87*
	p	.002	.08	.006	.052	.000	.007	.004
	Slope	0.0012	0.0003	0.0015	0.001	0.0007	0.0016	0.0006

*P<0.05, **p<0.001

6.5.5 Empirical mode decomposition results

Using monthly data, it was generally observed that the second IMF had a high VCR value. This measure essentially evaluates the explanatory power of the trend component on the original series and provides insights into the significance of the trend. One of the key features distinguishing the trend component obtained by EMD from other trend analysis methods is its nonlinear structure and its ability to provide information about changes over time. Figure 6.8 presents an example of the IMFs and trend component obtained after EMD analysis. This graph shares insights from the analysis of the MuSyQ dataset values from the Konya region. The related time series was decomposed into a total of six sub-series, with the last sub-series representing the trend. The sub-series that explained the most variance of the original data was IMF1 with 57.1%, followed by the trend component with 16.5%. Examining the trend component overlaid on the original time series reveals a decreasing trend during the first 100 months, followed by an increasing trend up to the 400th month, and then a

slight decrease thereafter. However, the overall trend from the beginning to the end points is increasing.

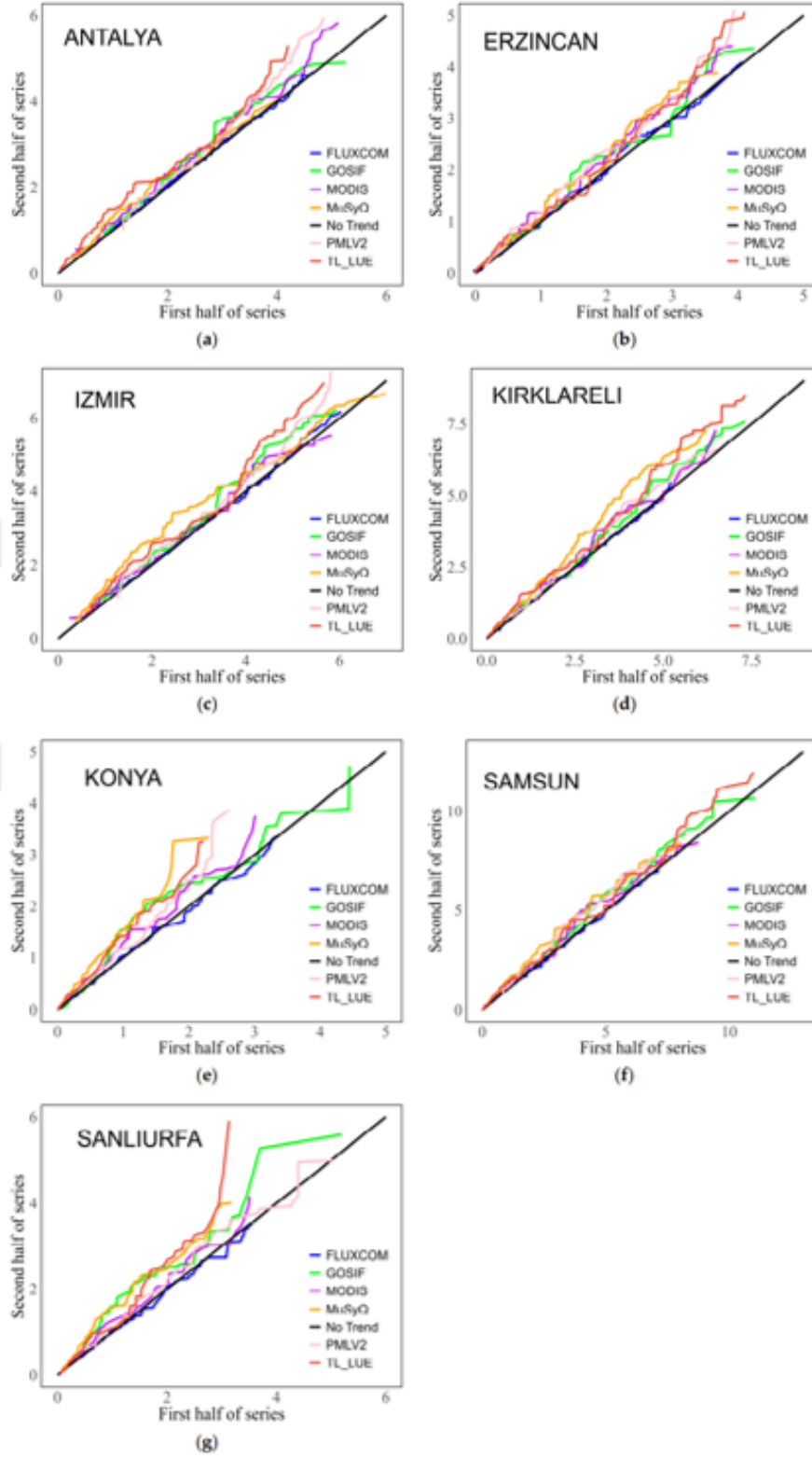


Figure 6.7 : ITA results for (a) Antalya, (b) Erzincan, (c) Izmir, (d) Kırklareli, (e) Konya, (f) Samsun, and (g) Sanliurfa.

All the GPP values for each region were analyzed in this manner, and the sub-series obtained through EMD along with their VCR values are shared in the Appendix A (Table A.8). Upon examining Figure 6.9, for the Antalya region, an increasing trend can be observed in all datasets except for FLUXCOM. In TL-LUE and PMLV2, a nearly linear increase is observed, whereas in GOSIF, MODIS, and MuSyQ, a trend of initial decrease followed by an increase is detected. The lowest VCR, at 0.13%, is calculated for FLUXCOM, while the highest, at 5.5%, belongs to GOSIF; the GOSIF and MODIS trend bands exhibit similar behavior. In Erzincan, an increasing trend is apparent in most datasets with the highest VCR (4.1%) in TL-LUE. In the Izmir region, a slight increasing trend is observed in the FLUXCOM dataset with a VCR of 1.73%, while TL-LUE, MuSyQ, GOSIF, and PMLV2 show an initial decrease followed by an increase. In Konya,

TL-LUE and MuSyQ exhibit similar behavior, as do MODIS and GOSIF. The most pronounced VCR, at 16.45%, is calculated for the trend in the MuSyQ data series. In Samsun, a near-linear increase is seen in TL-LUE and a decrease followed by an increase is observed in GOSIF, MODIS, and MuSyQ; this pattern is even more pronounced in PMLV2. In Sanliurfa, the most distinct trend behavior belongs to the MuSyQ dataset with a VCR of 11.61%. The TL-LUE trend series follows a near-linear course; the trend in the PMLV2 dataset can be interpreted as showing an overall trendless movement, decreasing until the midpoint and then increasing. All the time series and trend components obtained with EMD are shared in the Supplementary Materials (Figure A.10).

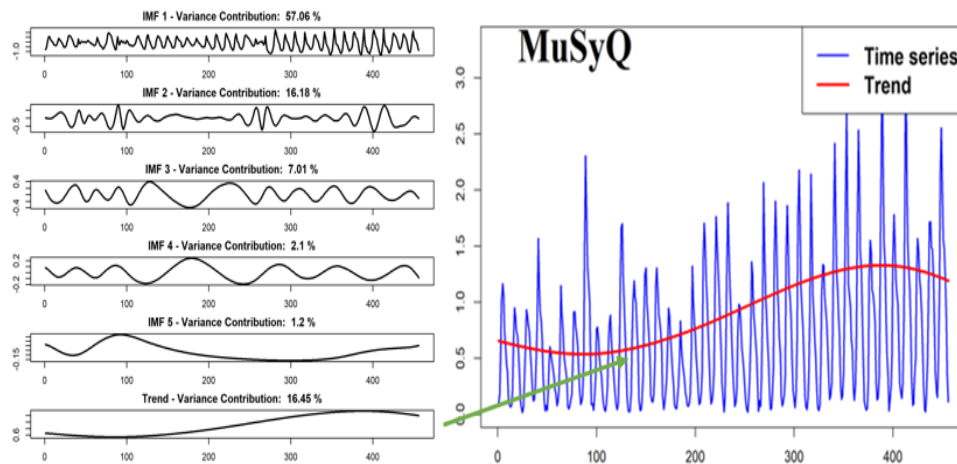


Figure 6.8 : Example of EMD subseries for MuSyQ GPP estimates from Konya, Türkiye.

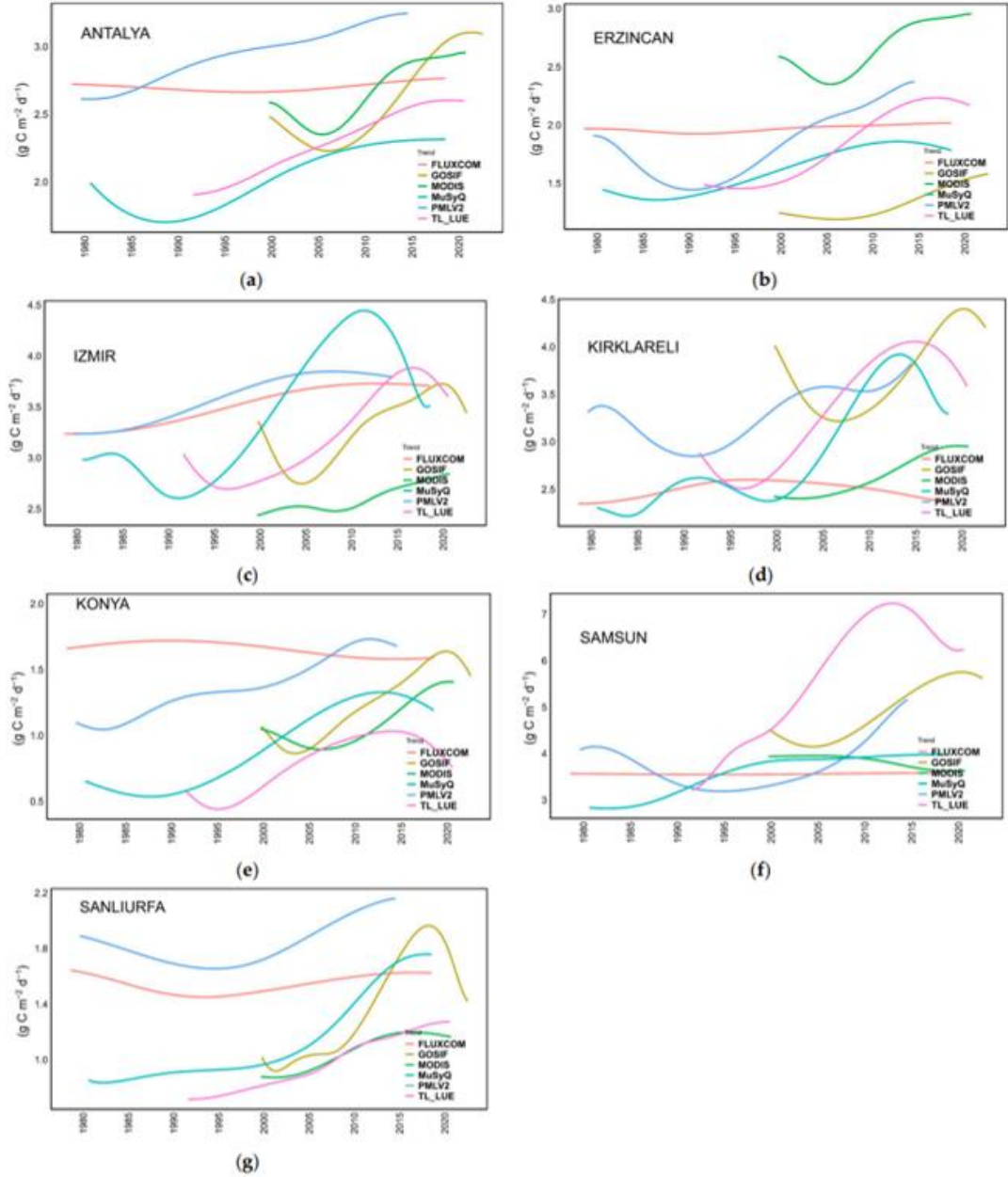


Figure 6.9 : Trends in GPP extracted from the last mode of an Empirical Mode Decomposition (EMD) on GPP estimates from different remote sensing data products from different regions of Türkiye, (a) Antalya, (b) Erzincan, (c) Izmir, (d) Kırklareli, (e) Konya, (f) Samsun and (g) Sanliurfa.

6.6 Discussion

6.6.1 Overview

This study's regional focus on Türkiye, segmented into seven distinct areas, provided a unique opportunity to explore GPP trends in varied ecological and climatic contexts. The use of six different GPP datasets – that were unable to be validated due to the limited presence of eddy covariance systems – aligns with existing literature showing

an increasing trend in GPP across many global regions (Piao et al, 2019; Zhu et al, 2016). Our methodology, which incorporated multiple GPP estimates and trend analysis methods, lends confidence in the overall finding that GPP increased at the study locations but also leads to important nuances when interpreting results that we describe in more detail below.

The datasets utilized, derived from various remote sensing methodologies, underscore the potential and challenges of using satellite data for GPP estimation. The inclusion of different calculation methods, namely Light Use Efficiency models (e.g. MODIS, MuSyQ, TL-LUE), machine learning (FLUXCOM), GPP proxies (GOSIF), and biophysical models (PMLV2) reinforces our results; for example, that there are numerous outliers in the MODIS, GOSIF, and TL-LUE GPP datasets in the Konya region, whereas in the Sanliurfa region, similar findings are observed in the MODIS, PMLV2, GOSIF, and TL-LUE GPP datasets. Moreover, ITA analyses have demonstrated that there is a pronounced increasing trend in TL-LUE across all regions at high GPP values, meaning that large carbon uptake events are becoming larger. Significant increases at high values of the PMLV2 dataset have been observed in the regions of Antalya, Erzincan, Izmir, and Konya. At the same time, the TL-LUE data exhibits a declining trend in recent years across all areas and GPP values from the GOSIF dataset show a slight decreasing trend in recent years across all regions, except for Erzincan, and PMLV2 data has maintained a continuous increasing trend, which demonstrates the importance of using multiple datasets for a conservative interpretation of GPP trends.

Although most trend analysis methods agree on the presence of an increasing trend across most regions, they diverge in certain details. The MMK method follows a classical ranking-based approach, allowing for the interpretation of the presence of a trend based on a certain Z value, but lacks the graphical extrapolation capabilities of ITA and EMD. ITA provides detailed information about the trend status at low, medium, and high values, whereas EMD is more successful in explaining how the trend changes over time. The multi-method analysis lends confidence that results are less likely to be an artefact of the particular method used (Vernay et al, 2024).

6.6.2 Spatial analysis of GPP datasets

We noticed nuanced insights into the factors influencing GPP in various regions of Türkiye. Samsun, positioned along the Black Sea coast, emerged as the region with the highest average GPP, consistent with wetter conditions and coniferous forests in the area including Anatolian black pine (*Pinus nigra*) with relatively high photosynthetic capacity (Deligöz et al, 2019) that is relatively insensitive to ecoclimatic setting (Fkiri et al, 2023). Additionally, Samsun's climate, particularly its relative humidity, is more favorable to plant growth than in other regions resulting in a relatively lower vapor pressure deficit (VPD). GPP is strongly constrained by VPD (Fu et al, 2022) and the combination of favorable climate and dense forest cover create an optimal environment for high GPP. The Kırklareli region, known for its extensive agricultural activities, displayed the next highest average GPP. This region, situated in the north of Thrace, benefits from the western Black Sea's precipitation patterns, receiving rainfall levels above the national average. The İzmir and Antalya regions, both influenced significantly by their coastal locations, exhibit distinct GPP dynamics. İzmir's annual average precipitation tends to be lower than that of the coastal regions of Samsun and Antalya. However, its average temperature is higher than Samsun's and similar to Antalya's. Warmer temperatures in İzmir favor agricultural practices that depend on warm and humid conditions, and this intensive agricultural management impacts GPP, such that the selected region of İzmir has higher GPP than Antalya albeit with similar trends.

The Konya region is one of the least rainy areas in Türkiye and has a predominantly arid climate. The majority of agricultural activities in this region rely on dryland farming due to the limited availability of water. Konya has little forest cover, and the vegetation is largely dependent on the diversity of seasonally cultivated agricultural crops. According to data from the Turkish Statistical Institute (TUIK), the cultivated area in the region has increased from 1,326,220 hectares in 2004 to 1,507,451 hectares in 2022, which is consistent with the observed trend of increasing GPP values in the study region. Similarly, the Sanliurfa region is characterized by low rainfall but has greater irrigation intensity than Konya, which supports large-scale agricultural operations (Ozdogan et al, 2006). After Konya, Sanliurfa has some of the lowest average GPP values among the regions studied but likewise with an increasing trend. These findings highlight the complex interplay between climatic factors, vegetation

types, and topography in determining regional GPP and point to the importance of understanding ongoing GPP trends for ecological management and forecasting the impacts of climate change on regional productivity.

6.6.3 Temporal analysis of GPP datasets

In the analysis of the temporal trends of GPP values, the study employed various trend analysis methods, all of which were non-parametric, thus avoiding the strict pre-conditions associated with parametric methods like linear regression. Among these, the MK method is one of the most commonly used trend analysis methods. Although not a parametric method, it is generally preferred for data without high autocorrelation (Mirabbasi et al, 2020). To enhance the reliability of the results, a modified MK method (MMK) was used in this study. Despite efforts to mitigate false trends arising from autocorrelation through pre-whitening processes, some studies have reported that MMK can still be prone to errors and may produce misleading results (Bürger, 2023). Notably, it interpreted significant trends far less often than ITA and EMD for the study datasets such that the interpretation of GPP trends will often differ if MK (or MMK) or the graphical trend analysis methods, ITA or EMD, were used. ITA is adept at clearly delineating trends in both low and high values, while EMD provides information on how trends in GPP have increased or decreased over specific years. When assessing all regions collectively. The trend analysis results from ITA and EMD were similar.

Analysis with the MMK method using monthly GPP data across all regions revealed that the FLUXCOM dataset, consistent with literature (Jung et al, 2020), showed a trendless pattern. In contrast, the MODIS dataset indicated a decreasing GPP trend during February, March, and April – although this was not observed in other datasets – which would lead one to search for the mechanisms underlying GPP changes during late winter and early spring. The most pronounced increases in GPP values were calculated during the summer months, which would lead one to study how changes in climate, leaf area index, and perhaps agricultural intensity have altered GPP since the 1980s.

Izmir and Antalya, displayed similar trends in GPP values, it is believed that this similarity is influenced by the fact that both regions share the same Köppen climate classification (Csa), indicating a similarity in their climatic conditions.

In examining the results of the EMD method, the FLUXCOM dataset consistently showed no trend across all regions, as expected (Jung et al, 2020). A key distinction of the EMD method was its indication of initially decreasing trends in GPP values, followed by increasing trends, due likely in part to its search for the oscillating signals that underlie a dataset (Guo et al, 2016). The GOSIF dataset showed a slight decreasing trend in recent years in all regions except Erzincan. Overall, these findings highlight the complexity of analyzing GPP trends and the importance of selecting appropriate trend analysis methods that can accurately reflect the temporal dynamics of GPP in various ecological settings. As mentioned by Du et al, (2024) a potential reason is that the GOSIF dataset was reconstructed using MODIS surface reflectance data, which may have led to some bi-ases in its algorithms.

6.6.4 Differences and similarities between GPP datasets

In the study, correlation analysis and the Kruskal-Wallis test were applied to analyze the similarities and differences among the datasets. Initially, scatter plots were used to determine at which levels the datasets were more similar or different. It was generally observed that while the similarity was high at lower values, it decreased with higher values, indicating the presence of increasing variance. Correlation values confirmed high positive relationships across all regions, but since these values were linear and not statistically powerful enough to compare the entire group together, the KW test was additionally applied. The results of the KW test and subsequent pairwise comparisons indicated that in each region, at least one dataset possessed a distinct structure. The most significant difference among the GPP datasets was observed in the PMLV2 dataset. Its emphasis on following physical processes could be the reason for this divergence and, given its explicit coupling of the carbon and water cycles, one may argue that it has advantages for the temperate, semi-arid, and arid climate zones of Türkiye. A notable finding was that in arid regions like Konya and Sanliurfa, almost all datasets exhibited differences, further emphasizing the importance of water limitation on GPP, GPP products, and their trends. Zhang and Aizhong Ye. (2021) reported similar results. Basically, the differences in various remote sensing-based GPP products can be attributed to different algorithms using different input data and having different parameter values when describing environmental mechanisms over climate zones (Chen et al, 2019). In addition, stomatal closure when drought occurs in terrestrial ecosystems has been shown to affect photosynthesis due to limited leaf and

canopy water content or high VPD in the atmosphere. Different GPP models using different model parameters, such as maximum LUE and optimum temperature for different ecosystems, are the second major contributor to the possible differences in arid regions. Moreover, different GPP datasets in the literature have used different flux tower data to calibrate LUE parameters, which may also contribute to differences in simulated GPP (Lv et al, 2023). The use of different leaf area index data and land cover data (Bi et al, 2022) can also cause significant differences. Additionally, the number of outliers in all datasets for these regions was higher than in other regions. These analyses suggest that variations in the algorithms of the datasets significantly impact the calculation of values in arid regions. Therefore, it is recommended that new GPP prediction methods being developed should consider these variations and devise algorithms that account for the distinct characteristics of arid regions.

6.6.5 The importance of the findings

When making a general commentary on the results obtained from this study, a trend of increasing greenness in Türkiye has been identified, consistent with many global regions (Piao et al, 2019). The observed low GPP values in the semi-arid regions of Türkiye (Konya, Sanliurfa, and Erzincan) are attributed to the high sensitivity of semi-arid ecosystems to water availability. Adverse trends in rainfall can lead to abrupt declines in vegetation productivity or the loss of ecosystem resilience (Lee et al, 2023). Additionally, vegetation in arid areas is more sensitive to changes in environmental factors, which can significantly underestimate maximum LUE values in arid ecosystem models (Wang et al, 2019). Consequently, the lack of appropriate LUE parameters in arid areas and the inability to accurately simulate the complex relationship between water and LUE contribute to most LUE models yielding low GPP estimates (Zhang et al, 2023).

GPP values vary across different geographical regions, and it is strongly recommended that these values be examined on a region-specific basis. Identifying the parameters that most significantly affect GPP and, if necessary, using local adjustment coefficients to enhance the representational accuracy of GPP datasets is crucial. Given Türkiye's geographical position, surrounded by seas on three sides and hosting a variety of micro-climates, it is essential to increase the number of ground observation stations to enable more accurate carbon cycle analyses. Doing so across Türkiye's

geographical regions would be beneficial given the differences in GPP magnitudes and trends. These measurements should then be used to calibrate remote sensing datasets, ensuring that they accurately reflect the distinct environmental and ecological characteristics of each region in Türkiye. This approach will not only improve the accuracy of current GPP estimations but also enhance our understanding of regional ecological dynamics, contributing to better environmental management and policymaking.

6.6.6 Limitations of study

Our study has some limitations. The most important limitation is that the GPP values produced by remote sensing systems cannot be verified with site measurement data in the study area. This is due to the very limited number of eddy covariance systems in Türkiye. The spatial resolution of the datasets used in the study is limited to 0.05 degrees, and more accurate analyses could result from higher-resolution data in light of recent developments in remote sensing systems.

6.7 Conclusions

In this study, spatiotemporal analysis of GPP values for seven major regions in Türkiye was conducted. We evaluated different remote sensing products for GPP, since there are inadequate measurement systems, i.e., flux towers and tower networks, in the study area. While many studies in the literature have focused on the global-scale analysis of GPP values, local-scale analyses are rarely conducted. Moreover, many studies in the literature have focused on annual GPP sums, whereas this study presents monthly analyses of GPP. The key conclusions drawn from this study can be listed as follows:

- GPP values show high similarity among datasets at low values, but this similarity decreases at high GPP values. The inputs used by the algorithms differ and more meticulous regional calibration and validation are necessary. When conducting a monthly trend analysis of GPP values, only the MODIS dataset exhibited a decreasing trend at certain months. The factors causing this trend should be further investigated. Additionally, significant increasing trends are detected in GPP values during summer months. There are many potential mechanisms that underlie these trends, which should be studied further to understand the changing carbon cycle of Türkiye.

- The ITA and EMD methods could be promising alternatives to MMK as they provide additional insight into how time series change. This comparison in the field of GPP is a first, and future studies may benefit from these methods for trend analysis. The ITA and EMD methods have already made substantial contributions to the literature in terms of visualizing trends. While ITA can easily detect trends at low and high values, EMD stands out for answering how the trend follows a nonlinear path over time. Optimizing the hyperparameters of the EMD method could lead to the more rational extraction of trend component information.

We hope that the present analysis helps motivate more studies on carbon cycling and its trends, in Türkiye and other regions that are under-represented in flux networks, to improve our mechanistic understanding of the carbon cycle across larger regions of the globe. Also, researchers are encouraged to use data with higher spatial resolution in future studies, to work with data classified according to plant functional type, and to analyze variables that may affect the trends in GPP values.

7. COMBINED DROUGHT INDEX USING HIGH-RESOLUTION HYDROLOGICAL MODELS AND EXPLAINABLE ARTIFICIAL INTELLIGENCE TECHNIQUES IN TÜRKİYE⁶

7.1 Introduction

Drought is a natural disaster characterized by insufficient water resources in a given region to meet demand, typically associated with periods of low precipitation (Başakın et al, 2021; Mishra and Singh, 2010). In addition to the widely accepted simple definition, drought can also be described as a complex process involving trends of soil moisture depletion, reduced streamflow, and declining groundwater levels, often accompanied by increasing temperatures (Spinoni et al, 2017). This situation leads to negative impacts on ecosystems, agriculture, water resources and human life (Bhaga et al, 2020). Drought can be classified in several ways: meteorological drought results from periods of below-average precipitation (Kumari et al, 2023), hydrological drought occurs due to reduced surface water flows (Hasan et al, 2019), agricultural drought is caused by insufficient soil moisture to support normal crop development (Wu et al, 2024), and socio-economic drought affects human lives and leads to economic disruption (Pathak and Dodamani, 2021). These are conventional classifications, but today the terms flash drought, characterized by rapid onset and assessed on a 5-day basis (Otkin et al, 2018), and ecological drought, characterized by reduced water availability leading to reduced vegetation activity (with consequent disruption of ecosystem services) (Raposo et al, 2023; Wang et al, 2023), are also commonly used. Each of these types (except socio-economic) is defined and assessed based on specific hydroclimatic variables (Raza et al, 2021), while socio-economic drought is related to local water supply and monitors water demand through socio-economic systems (Tu et al, 2018). As a natural disaster with increasing costs each

⁶ This chapter is based on the paper “Başakın, E. E., Stoy, P. C., Demirel, M. C., Ozdogan, M., and Otkin, J. A. (2024). Combined Drought Index Using High-Resolution Hydrological Models and Explainable Artificial Intelligence Techniques in Türkiye, Remote Sensing, 16(20), 3799. doi:10.3390/rs16203799.”

year, drought has a significant and widespread impact on many economic activities and people at any given time (Li et al, 2024).

Due to its complex and nonlinear nature, the mathematical characterization of drought events remains a substantial challenge. Dynamic drought processes are often simplified by defining drought indices, which are simplified formulae derived by analyzing long-term trends of variables like precipitation, evaporation, and temperature using statistical methods (Zhang et al, 2024). Drought indices, which can characterize drought in terms of duration, severity and intensity (Cavus et al, 2023), are indispensable in drought assessment as they facilitate drought risk analysis and monitoring by comparing current and future drought characteristics with past events (Han and Singh, 2023). Some of the drought indices commonly used in the literature include the Standardized Precipitation Index (SPI) (McKee et al, 1993), Standardized Precipitation Evapotranspiration Index (SPEI) (Vicente-Serrano et al, 2010), Palmer Drought Severity Index (PDSI) (W. C. Palmer, 1965), Soil Moisture Anomaly (SMA) (Bergman et al, 1988), Evapotranspiration Deficit Index (ETDI) (Narasimhan and Srinivasan, 2005), Streamflow Drought Index (SDI) (Nalbantis and Tsakiris, 2009), Enhanced Vegetation Index (EVI) (Huete et al, 2002), Normalized Difference Vegetation Index (NDVI), Drought Reconnaissance Index (DRI) (Tsakiris & Vangelis, 2005), Evaporative Demand Drought Index (EDDI) (Hobbins et al, 2016) and Effective Drought Index (EDI) (Byun and Wilhite, 1999), all of which were designed with different drought impacts in mind. These indices can represent a single class of drought, while some can represent multiple classes of drought. For instance, SPI is commonly used to define meteorological drought, SDI for hydrological drought, SMA for agricultural drought, and PDSI for both meteorological and agricultural drought (Gunda et al, 2016). These indices may be effective for describing particular drought features, but were not necessarily designed to incorporate all drought impacts.

Drought indices can be classified as single, multivariate or combined based on the number of parameters they use. While single drought indices are easier to compute, their success in reflecting drought conditions may be less than that of multivariate and combined indices (Li et al, 2024). Multivariate indices are derived by combining different variables. For example, the SPEI uses both precipitation and evapotranspiration values, while the PDSI includes precipitation, temperature, and

water-holding capacity of soil. Combined drought indices (CDI) are formed by merging several single drought indices to highlight the multidimensional characteristics and multifaceted effects of drought (AghaKouchak et al, 2015), and therefore may be of particular use for characterizing key drought impacts that single and multivariate models may miss.

With the advancement of technology, there has been a significant increase in the number of parameters measured at the earth's surface. These measurements can be made by both ground-based stations and remote sensing systems. In contrast to ground-based measurements, remote sensing products (attempt to) provide continuous measurements in both time and space, despite varying spatial, temporal, and spectral resolutions. The data obtained from remote sensing systems, such as land surface temperature, precipitation, soil moisture, and vegetation reflectance characteristics, have introduced numerous drought indices into the literature. Some of the most commonly used indices include the Temperature Condition Index (TCI) (Kogan, 1995), the Vegetation Condition Index (VCI) (Kogan, 1995), the Normalized Difference Water Index (NDWI) (Gao, 1996), and the Soil Moisture Condition Index (SMCI) (Zhang and Jia, 2013), all of which measure different physical phenomena. CDIs, derived from the combination of these and/or other indices, play a critical role in the detection of agricultural drought events (Hanadé Houmma et al, 2023). One of the earliest examples of such combined indices is the Vegetation Health Index (VHI), which combines the TCI and VCI (Kogan, 2001). Danodia et al. (2021) conducted a study in India using TCI, VCI, PCI, and SMCI values to calculate CDI values and found significant positive relationships between CDI values and crop yield. Karimi et al. (2022) used TCI, PCI, Soil Water Index, and VCI inputs to calculate CDI for their drought analysis in Iran and reported that the resulting CDI values better represented the 2008 and 2014 droughts than SPI. In a study conducted in Niger, Houmma et al. (2023) constructed a CDI index using Normalized Rainfall Efficiency Index, PCI, TCI, and smoothed vegetation index values. They reported that the index had a high correlation with VHI and PDSI, thus providing important information for agricultural drought monitoring. Li et al. (2024) proposed a new CDI index using PCI, TCI, and Evapotranspiration Condition Index values as inputs. They found that the new index had a higher correlation with SPI-1, a one-month SPI, than the individual indices. In

addition, the proposed index showed a negative correlation with Drought Affected/Disaster Area and Net Primary Productivity.

In reviewing the literature on CDI studies, there are a limited number of studies using gross primary productivity (GPP) and NDWI despite opportunities for integrating carbon and other water-based metrics into drought indices. In Wei et al. (2023), GPP was not used as an input, but rather to validate the derived CDI. NDWI is sensitive to the water content of the vegetation canopy (Shashikant et al, 2021), and numerous studies have shown that NDWI is effective in monitoring agricultural drought (Patil et al, 2024; Zhang et al, 2023); for example Delbart et al. (2005) found that NDWI was more responsive than NDVI to crop responses to drought. Similarly, in another study, Olmos-Trujillo et al. (2020) found that NDWI was more sensitive to agricultural drought than NDVI in semiarid regions. The GPP, as a critical component of terrestrial ecosystem productivity and surface carbon cycling, represents photosynthesis at the ecosystem level (Başakın et al, 2024). Recent studies have highlighted the importance of the relationship between drought and GPP (Wei et al, 2023; Yang et al, 2023; Zhang et al, 2024). The study by Pendergrass et al. (2020) found that changes in GPP should also be considered when examining the causes of regional meteorological and agricultural droughts.

Most methods for creating a CDI derive a linear equation, with the optimal calculation of the equation's coefficients being one of the main challenges. To address this issue, Kogan (2001) calculated the Vegetation Health Index by taking the arithmetic mean of TCI and VCI, giving equal importance to both inputs. In their study, Hao et al. (2015) optimized the CDI equation weights through trial and error by assigning different weights to the TCI, SMCI, and PCI indices. This trial-and-error method requires evaluating the CDI values against a target. Zhang et al. (2017) used seven different weighting scenarios in a trial-and-error approach, with the PDSI, Z-index, and SPI as target variables to determine the optimal scenario. In studies conducted without a target variable, researchers often utilize methods such as principal component analysis (PCA) (Arab et al, 2022; Parthsarathi Pandya et al, 2022) entropy (Tang et al, 2023; Waseem et al, 2015), and analytic hierarchy process (AHP) (Cai et al, 2023; Shi et al, 2022). In CDI calculations, PCA captures the common variance of the single indices and creates principal components for analysis, assigning weights to

each single index. These weights are determined by each input's contribution to the principal components, with higher variance components receiving more weight. CDIs based on PCA assume linearity among random variables, which can lead to information loss post-analysis (Suo et al, 2024). This linearity assumption can be problematic in drought studies involving complex hydrological and meteorological systems with nonlinear characteristics (Kanthavel et al, 2022). AHP methods, on the other hand, rely on expert opinions to weigh criteria. Some disadvantages of this method include high computational requirements even for small problems, a subjective nature that can depend on emotional judgments, and increased time and effort required for additional pairwise comparisons (Salvia et al, 2019).

A review of the CDI literature reveals that very few studies have used machine learning (ML) methods to calculate the weights of the equation. Of those that have employed ML, Houmma et al. (2022) used the Random Forest (RF) method to determine the weights of CDI inputs. However, metrics used in the RF model, such as Gini impurity and permutation importance, can be biased, especially for features with high cardinality (many categories). This can inflate the importance of certain features, thereby leading to an inaccurate assessment of their true contribution to the model (Wallace et al, 2023).

Recently, methods have been developed to explain nonlinear relationships and work alongside ML techniques to evaluate the power of inputs in predicting the output, taking into account their intricate relationships. One of the most frequently used methods that can adapt to almost any type of ML model is the SHapley Additive exPlanations (SHAP) method. SHAP uses Shapley values from game theory, which are used to fairly determine each player's (in this case, the features in the model) contribution to a game (the model's prediction). A feature's contribution to the model is calculated as the average of its marginal contribution across all possible combinations. While Random Forest's feature importance ranks the general contributions across the entire model, it does not provide local (individual prediction level) explanations. In contrast, SHAP provides individual explanations for each prediction, allowing for a more detailed and fair understanding of the model's decisions (He et al, 2023).

The study aims to answer the following questions:

1. Does adding NDWI and GPP values improve CDI calculations?
2. How successful is the developed CDI in detecting agricultural drought?
3. What are the relationships between CDI, formed by different combinations of single drought index indicators, and crop yield?
4. What is the nonlinear behavior of each input on the output?

To find answers to the questions listed above, this study is to define a new CDI, that will reflect winter wheat yield values. In addition to the commonly used TCI, PCI, VCI, and SMCI, the NDWI and GPP have been included to calculate CDI values. To demonstrate the effectiveness of including the CDI, correlation coefficient between CDI and winter wheat yield were calculated. This is intended to more accurately detect agricultural drought events. Furthermore, high-resolution hydrological model outputs have been used to obtain SMCI, thus employing soil moisture values derived with the Mesoscale Hydrological Model (MhM) instead of directly using satellite-derived soil moisture values, which were not available across the study period. The SHAP method has been used to determine the equation weights for the CDI, and its effectiveness has been evaluated by comparing it with PCA and empirical method results. By using SHAP analysis to assess the nonlinear behavior of each input variable (TCI, PCI, VCI, SMCI, NDWI, and GPP) on the CDI, we aim will provide a deeper understanding of how each individual index contributes to the overall CDI and its impact on crop yield, revealing complex interactions and the relative importance of each input variable. Furthermore, the developed index is compared against the widely used SPEI and scPDSI to quantify differences and lend further insight into drought indices for agricultural yield estimation.

7.2 Materials

7.2.1 Study area

This study focuses on four of the most critical areas for Türkiye's food security, namely Aksaray, Karaman, Konya, and Niğde, which are characterized by fertile soils and cover an area of 54,908 km² (Ergün and Demirel, 2023). These four regions are part of Türkiye's driest interior zone, with an average annual rainfall of 380 mm. The elevation of the study area ranges from 284 to 3,583 meters, with agricultural activities primarily conducted in the lower northern sections (Koycegiz et al, 2023). These

provinces mainly practice rainfed agriculture for wheat. According to the annual reports of the Turkish Statistical Institute (TUIK), approximately 70% of the region is engaged in rainfed agriculture. There are 49 administrative districts in these provinces; we excluded districts with minimal agricultural land as indicated by land cover maps in the present study. The list of districts included in the study and their statistical information is presented in Table B1 and Figure 7.1b. The districts with limited or no cropland land areas are excluded from the study area and labelled as “Unstudied” in Figure 7.1b. From elevation map it is obvious that arable land is very limited due to topographical conditions.

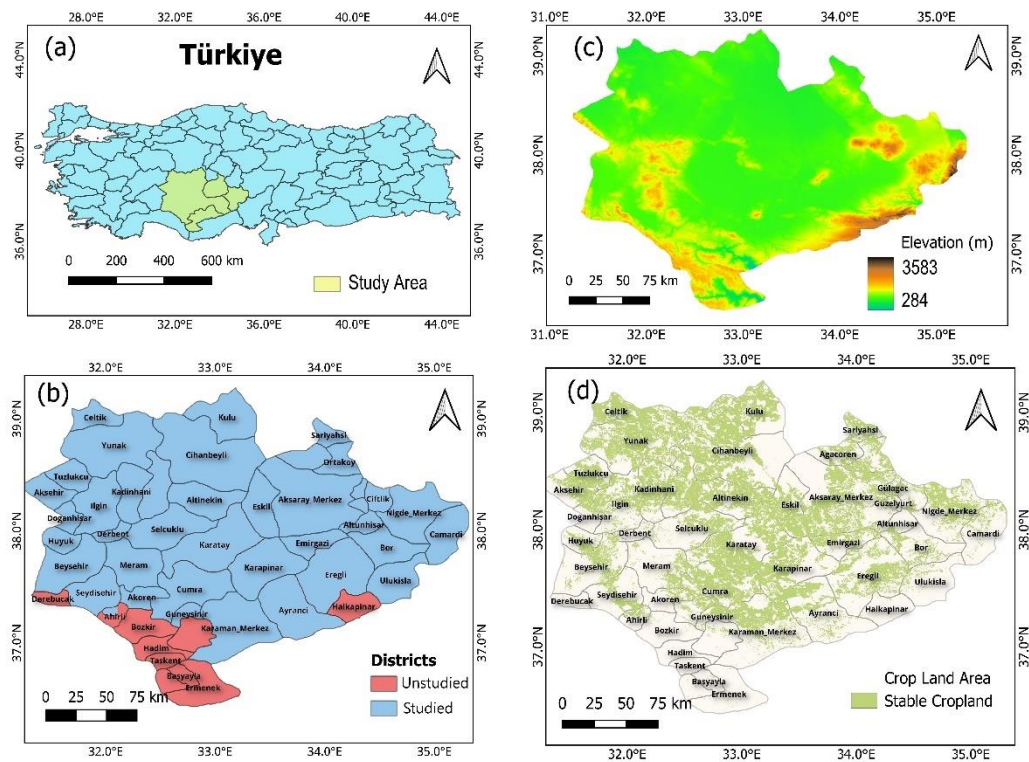


Figure 7.1 : (a) The study area in Türkiye, (b) districts in the study area, (c) a digital elevation map, and (d) cropland areas.

7.2.2 Data description

The precipitation values used to calculate PCI were obtained from the Climate Hazards Group InfraRed Precipitation with Station (CHIRPS) database. CHIRPS is a widely used precipitation dataset for climate data analysis and research. It contains global precipitation estimates from 1981 to the present, available on daily, 10-day, and monthly bases. The dataset is produced by combining two main data sources: satellite-based estimates and ground station data. CHIRPS uses precipitation data provided by

the U.S. National Oceanic and Atmospheric Administration Climate Prediction Center and enhances satellite estimates by incorporating precipitation data from thousands of ground stations worldwide (Funk et al, 2015). Due to the discrete nature of precipitation data, there can be challenges in obtaining precipitation from remote sensing systems, which necessitates research on their correlation with ground observations. Studies conducted specifically for Türkiye have shown that monthly CHIRPS data have a high correlation with ground observations (Aksu and Akgül, 2020). In addition, CHIRPS data used in drought studies specific to Türkiye have been verified to produce satisfactory results by Aksu et al. (2022) and Orieschnig and Cavus (2024).

Land surface temperature (LST) and NDVI were obtained from the Moderate Resolution Imaging Spectroradiometer (MODIS) database for the calculation of TCI and VCI. In addition, NDWI was also similarly obtained from MODIS. Launched in 1999 on the Terra satellite and in 2002 on the Aqua satellite, both satellites scan the entire Earth's surface every 1 to 2 days, collecting data in 36 spectral bands. The instrument's detectors collect data at three different spatial resolutions: 250 m, 500 m, and 1,000 m. MODIS bands 31 (centered at 11,030 μm) and 32 (12,020 μm), are thermal infrared bands used in LST calculations (Kim et al, 2024). NDVI, an indicator widely used to monitor vegetation health and density, is obtained using MODIS Band 1 (red: 645 nm) and Band 2 (NIR: 858 nm). Finally, NDWI values are obtained using MODIS Band 2 (NIR: 858 nm) and Band 5 (SWIR: 1240 nm).

In previous studies, researchers have used soil moisture (SM) data from various databases, including the Global Land Data Assimilation System (GLDAS) (Yin and Zhang, 2023), Advanced Microwave Scanning Radiometer for EOS (Jiao et al, 2019), Famine Early Warning Systems Network Land Data Assimilation System (Baig et al, 2020), ERA5-Land (Rahman et al, 2024), and Soil Moisture and Ocean Salinity (Wang et al, 2022). In this study, SM values were obtained using the mHM, a physically-based high-resolution distributed hydrologic model used to simulate hydrological processes in watersheds at different scales. The mHM divides watersheds into small cells and performs calculations for each cell separately, allowing for more realistic results by accounting for spatial heterogeneity. The model is based on physical principles and to simulates hydrological processes including evaporation,

transpiration, runoff, and groundwater flow (Demirci and Demirel, 2023). While the model is designed to predict surface runoff using inputs such as precipitation, potential ET, and mean air temperature, it also outputs SM at various depths, including 0-50 cm. In this study, these SM values were used to calculate the SMCI.

Gross Primary Productivity (GPP) values were obtained from the Penman-Monteith-Leuning Evapotranspiration V2 (PMLV2) dataset (Zhang et al, 2023). The Penman-Monteith-Leuning (PML) model was first developed by Leuning et al. (2008), who introduced a biophysical model for surface conductance based on the Penman-Monteith (PM) equation (Pei et al, 2020). Since then, two improved versions of the PML model have been developed: PML-V1 and PML-V2. PML-V2, as described by Gan et al. (2018), uses a biophysical canopy conductance model to integrate evapotranspiration (ET) with GPP.

Temporal and spatial resolutions, dates, and data sources are given in Table 7.1. In addition, the wheat yields used in the study were obtained from the TUIK database. The yields are in t/ha and were obtained at the district level. The land cover/land use map of the Global Land Analysis and Discovery (GLAD) team with a resolution of 30 m was used to identify the cultivated areas (Potapov et al, 2022).

Table 7.1 : Descriptive information of the data.

Data	Spatial Resolution	Temporal Resolution	Time Range	Sources	Product Name
Rainfall	0.25 degree	Monthly	1 January 2000–30 December 2021	CHIRPS	CHIRPS-2.0
LST	1000 m	Daily	1 February 2000–30 December 2021	MODIS	MOD11A1
NDVI	1000 m	Monthly	1 February 2000–30 December 2021	MODIS	MOD13A3
NDWI	0.05 degree	Daily	24 February 2000–30 December 2021	MODIS	MOD09GA v061
GPP	0.05 degree	Monthly	1 March 2000–30 December 2021	MODIS and PML_V2 0.1.7 GLDAS	
Land Cover	30 m	-	2020	GLAD	
Crop Yield	District Based	Annual	2001–2020	TUIK	
Soil Moisture	500 m	Monthly	1 January 2000–30 December 2021	mHM	

7.3 Methods

7.3.1 Calculations of drought indices

7.3.1.1 The precipitation condition index

The PCI was calculated by comparing current precipitation to long-term maximum and minimum precipitation. The precipitation used to calculate the PCI was obtained from the CHIRPS data set. The equation used to calculate the PCI is shown below.

$$PCI_i = \frac{CHIRPS_{i,j} - CHIRPS_{i,j,min}}{CHIRPS_{i,j,max} - CHIRPS_{i,j,min}} \quad (7.1)$$

In this context, i represents the pixel, and j represents the month. $CHIRPS_{i,j,max}$ denotes the maximum precipitation for the i -th pixel in the j -th month, while $CHIRPS_{i,j,min}$ denotes the minimum precipitation for the i -th pixel in the j -th month. This index ranges from 0 to 1, with values close to 0 (1) indicating dry (wet) periods.

7.3.1.2 The temperature condition index

The TCI is calculated using LST values obtained from the MOD11A1 dataset. Similar to the calculation method of the PCI, the TCI ranges from 0 to 1, but higher TCI values are associated with drought conditions, while lower values indicate non-drought periods. The TCI, like the PCI, ranges between 0 and 1.

$$TCI_i = \frac{LST_{i,j,max} - LST_{i,j}}{LST_{i,j,max} - LST_{i,j,min}} \quad (7.2)$$

7.3.1.3 The vegetation condition index

The VCI was proposed by Kogan (1995). This index indicates how close the current month's NDVI value is to the minimum NDVI calculated from long-term records. It is developed by scaling each location's and pixel's VCI values from 0 to 1:

$$VCI_i = \frac{NDVI_{i,j} - NDVI_{i,j,min}}{NDVI_{i,j,max} - NDVI_{i,j,min}} \quad (7.3)$$

High VCI is associated with wet periods, whereas low VCI is associated with drought periods.

7.3.1.4 The soil moisture condition index

The calculation steps of the SMCI are the same as those of the PCI. SMCI values also range from 0 to 1, where low (high) SMCI indicates drought (wet) periods. The steps for calculating SMCI are provided below.

$$SMCI_i = \frac{SM_{i,j} - SM_{i,j,min}}{SM_{i,j,max} - SM_{i,j,min}} \quad (7.4)$$

The SM values produced by the mHM model are used to calculate the SMCI. Technical information about the mHM model is provided in Section 4.3.2.

7.3.1.5 The normalized difference water index

NDWI is strongly associated with plant water content. Shortwave infrared (SWIR) reflectance captures changes in both spongy mesophyll composition and plant water content. In contrast, NIR (Near-Infrared) reflectance is determined by the internal structure of the leaf and its dry matter content rather than its water content. The combination of NIR and SWIR is designed to improve the accuracy of SM determination by eliminating variations due to internal leaf composition and leaf dry matter quality.

$$NDWI_i = \frac{NIR_i - SWIR_i}{NIR_i + SWIR_i} \quad (7.5)$$

NDWI values range between +1 and -1. An NDWI value close to +1 indicates higher plant water content, while a value close to -1 indicates the absence of plant water content. In this study, the NDWI was normalized using the same min-max normalization applied to the other indices, ensuring they fall within the same range (0-1). NDWI values close to 1 are associated with wet periods, whereas values close to 0 are associated with dry periods.

7.3.1.6 The gross primary productivity

The GPP data used in this study consist of data obtained using the PMLv2 method (Zhang et al, 2023). The PMLv2 model is a biophysical model designed to estimate evapotranspiration (ET) and GPP by coupling water and carbon processes. It improves on earlier versions by incorporating stomatal conductance and the effects of carbon dioxide on both water loss and photosynthesis, allowing for a more accurate simulation

of land–atmosphere interactions. PMLv2 uses remote sensing data, such as MODIS-derived LAI and albedo, and is driven by meteorological inputs, including air temperature, humidity, wind speed, and solar radiation. Detailed information regarding PMLv2 is provided in the Appendix B.

7.3.2 Mesoscale hydrological model

The Mesoscale Hydrological Model (mHM) is a physically-based, distributed hydrological model designed to understand and model the components of the water cycle, particularly for assessing water resources over large areas and extended periods (Samaniego et al, 2010). The mHM utilizes a distributed approach to model spatially variable hydrological processes using meteorological (e.g., precipitation, temperature, evaporation) and soil data to compute hydrological processes for each grid cell (Kumar et al, 2013). The importance of SM modeling is the rainfall–runoff module, which simulates processes such as precipitation falling on the soil surface, infiltration, surface runoff, and river flow (Samaniego et al, 2010). The soil moisture module calculates the distribution and changes in moisture within the soil profile. In this study, SMCI was calculated using the SM values generated by the mHM. Precipitation, temperature, and PET were obtained from ERA-5 for use in the mHM. SM values were produced at a spatial resolution of 1 km for each grid cell within the study area. Figure B1 provides an insight into the working principles of the mHM. During the development of the mHM, the calibration was carried out using measurements from 10 streamflow stations within the study area, which includes four provinces: Aksaray, Karaman, Konya, and Nigde. The calibration set achieved a R of 0.78 and an RMSE of 3.2 m³/s between the simulated and observed values. For the validation set, an R of 0.77 and an RMSE of 3.3 m³/s were obtained.

7.3.3 Crop yield detrending

In this study, the weights of the CDI values were optimized using crop yield as the target variable. When comparing drought indices with yield, it is essential to remove the trend component from the original series. This trend is often associated with technological advancements, soil improvements, and fertilization effects in agriculture (Araneda-Cabrera et al, 2021). Common methods for detrending include polynomial fitting (linear or cubic) (Zipper et al, 2016), moving average (Chen et al, 2018), and locally weighted regression model (LOWESS) (Lischeid et al, 2022; Liu et al, 2022).

Due to certain assumptions associated with linear trends (Başakın, 2022), the difficulties in using cubic splines with small datasets (<60) and their tendency to remove components other than the trend (Lu et al, 2017), this study opted for the nonparametric LOWESS method (Lu et al, 2017). A standardization process was carried out to compare the detrended series with standardized series such as SPEI and scPDSI, and yield anomalies were obtained. The standardization process is given in equation (7.6).

$$detrend\ yield_{std} = \frac{detrend_i - \overline{detrend}}{\sigma_{detrend}} \quad (7.6)$$

Where $detrend_i$ represents the detrended yield in the i-th month, $\overline{detrend}$ represents the temporal average of the detrended yield in a district, and $\sigma_{detrend}$ represents the standard deviation of the detrended yield in a district.

7.3.4 Benchmarking methods

7.3.4.1 Standardized precipitation evapotranspiration index

The SPEI was developed by Vicente-Serrano et al. (2010) to assess drought by considering the climatic balance between precipitation and potential evapotranspiration (PET). Unlike the SPI, the SPEI considers temperature in addition to precipitation. Similar to SPI, SPEI can be calculated on different time scales, with negative values indicating drought periods. Due to the sensitivity of vegetation to evapotranspiration, the SPEI is considered to have a relative advantage in being closely associated with drought events. After calculating the climatic water balance over different accumulation periods, the data are fitted to a parametric distribution and the probabilities are converted to a per-centile of the standard normal distribution to obtain the SPEI values (Laimighofer and Laaha, 2022). Typically, the three-parameter log-logistic distribution is used as the reference distribution. The parameters are estimated using L-moments, following the approach proposed by Beguería et al. (2014). The SPEI used in this study was obtained from the global dataset produced by Beguería (2022), which has a spatial resolution of 0.5 degrees and a monthly temporal resolution. This dataset uses precipitation and temperature data produced by the Climate Research Unit to perform SPEI calculations. To assess the effectiveness of

CDI in this study, 1-, 3-, 6-, and 9-month SPEI values were obtained, represented by the abbreviations SPEI_1, SPEI_3, SPEI_6, and SPEI_9.

7.3.4.2 Self-calibrated palmer drought severity index

The Palmer Drought Severity Index (PDSI) (Palmer, 1965) characterizes the balance of moisture deficit and surplus based on the difference between actual precipitation and climatically appropriate precipitation. Although this index is primarily a meteorological drought index, it is often used to monitor agricultural drought events (Shen et al, 2019). The parameters of the original PDSI were obtained based on observations from the Midwestern United States, which limits its applicability to other regions. To address this issue, Wells et al. (2004) proposed the self-calibrating PDSI (scPDSI) by incorporating weighting and persistence factors to improve spatial comparability. The core idea of scPDSI is to determine the weighting and persistence factors based on locally observed hydrometeorological variables. The empirical parameters in the analysis procedure can be self-calibrated, thereby improving the spatiotemporal comparability of drought monitoring results using scPDSI. In this study, scPDSI values with a spatial resolution of 0.5 degrees were obtained from Van Der Schrier et al. (2013).

7.3.5 eXtreme gradient boosting

EXtreme Gradient Boosting (XGBoost) is a powerful ML algorithm used for classification and regression problems, created by sequentially boosting decision trees (Zhang et al, 2024). Gradient boosting attempts to minimize errors by updating model parameters along the gradient of a loss function. XGBoost improves on this by using a second-order Taylor series expansion to use the second derivatives of the loss function for more precise optimization. Thus, the model is optimized faster and more accurately compared to traditional gradient boosting methods that rely solely on first derivative information. In particular, it uses L1 (lasso) and L2 (ridge) regularization terms to control model overfitting (Gupta et al, 2024). Regularization helps maintain control over tree depth and number of trees, leading to more general results on complex datasets. XGBoost also has a builtin algorithm that handles missing data and automatically processes missing values. While most gradient boosting algorithms work sequentially, XGBoost processes different parts of the dataset in parallel,

speeding up the tree-building process (Bushenkova et al, 2024). This parallel processing capability allows XGBoost to work effectively on large datasets.

7.3.6 Shapley additive explanations

SHAP determines feature importances by calculating each feature's contribution to the prediction. This becomes particularly valuable when trying to explain the internal working mechanism of the model. SHAP is based on a concept called Shapley values (Shapley 1953) from cooperative game theory to fairly distribute the contributions of players (features) to the game (prediction model) (Ekmekcioğlu et al, 2022). This method calculates the marginal contribution of each feature to the model's prediction and does so for all possible feature combinations (Koc, 2023). Unlike traditional feature importance methods (permutation importance and Gini importance etc.), SHAP accounts for interactions between features when measuring each feature's contribution. This enhances the accuracy and reliability of SHAP, as features in most real-world data often interact with one another. Calculating SHAP values typically has high computational costs, but these costs can be minimized through optimization techniques like the TreeSHAP algorithm proposed by Lundberg and Lee (2020). In this study, the TreeSHAP method, which is particularly effective for decision trees and tree-based ensemble models like XGBoost, was also used.

7.3.7 Principal component analysis

PCA is a statistical technique used to better understand and analyze multivariate data. The primary goal of PCA is to explain the largest variance in a dataset by creating new variables, called principal components, which are linear combinations of the original variables (Ma et al, 2024). PCA begins by calculating the covariance matrix of the variables in the dataset, which shows the relationships and shared variance between them. Using this covariance matrix, eigenvalues and eigenvectors are computed. Eigenvalues represent the variance of the principal components, while eigenvectors determine the direction of these components. The eigenvector with the largest eigenvalue defines the first principal component and explains the most variance in the dataset. The eigenvector with the second largest eigenvalue defines the second principal component, explaining less variance and being orthogonal to the first component. This process continues until as many principal components as the number of variables are created. Since the principal components are linear combinations of the

original variables, understanding the relationship between these components can help reveal structural features and hidden patterns in the dataset. However, PCA has some limitations. For instance, it only considers linear relationships and may require data normalization (Jolliffe and Cadima, 2016). Additionally, interpreting the principal components can be challenging because they are complex combinations of the original variables.

7.3.8 Model evaluation metrics

In this study, statistical metrics were used to evaluate the performance of ML models and to analyze the relationship between drought indices and crop yield. The metrics employed include Spearman's correlation coefficient (R), root mean square error (RMSE), and Akaike information criterion (AIC). Due to the generally non-normal distribution of the data and the limited number of yield values available at the district level (20 samples, excluding ML models), Spearman's correlation coefficient was preferred over the Pearson correlation coefficient. Spearman's correlation coefficient is nonparametric, robust to outliers, and capable of revealing nonlinear relationships (Zhao et al, 2024). Additionally, RMSE and AIC were used alongside the correlation coefficient to assess the performance of the ML models. The equations for RMSE, AIC, and Spearman's correlation coefficient are provided below.

$$RMSE = \sqrt{\sum_{i=1}^n \frac{(y_i - \hat{y}_i)^2}{n}} \quad (7.7)$$

$$R = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \quad (7.8)$$

$$AIC = n \times \ln\left(\frac{SS_e}{n}\right) + 2K \quad (7.9)$$

where d_i is the difference between the two ranks of each observation, n is the number of observations, y_i is the actual yield, $\hat{y}_i$ is estimated yield, SS_e is sum square of errors, and K is number of parameters. Its value lies within the range $-1 \leq r \leq 1$, where $R < 0$ indicates a negative correlation, $R > 0$ indicates a positive correlation, and $R = 0$ indicates no correlation. The larger the absolute value of r is, the stronger the correlation between two variables. The RMSE ranges from 0 to infinity, with models

that achieve RMSE closer to 0 being considered superior. Lower AIC values indicate that the model is more useful.

7.3.9 Drought classification

Threshold values are commonly employed to categorize the severity of climatic events, such as droughts and excessive moisture. These classifications are crucial in assessing the extent to which regional climate conditions deviate from expected patterns. Such insights are instrumental in shaping long-term water management strategies and informing climate policies. The use of these classifications is essential for comparing climatic anomalies across different temporal scales and geographic regions. Given that the drought severity levels of CDI and SPEI are derived from similar statistical processes and have comparable lower and upper bounds, a widely used classification in the literature was adopted (Nigatu et al, 2021). In view of the considerable quantitative range encompassed by the severity values of scPDSI, a representation comprising a greater number of classes was selected to facilitate a more comprehensive analysis (Gavrilov et al, 2020; Raphael et al, 2017). The classifications for CDI, SPEI, and scPDSI are provided in Table 7.2.

Table 7.2 : Classification Thresholds for CDI, SPEI, and scPDSI.

CDI and SPEI	Class	scPDSI	Class
≤ -2	Extreme Drought	≤ -4	Extreme Drought
-1.5 to -1.99	Severe Drought	-3 to -3.99	Severe Drought
-1 to -1.49	Moderate Drought	-2 to -2.99	Moderate Drought
-0.5 to -0.99	Mild Drought	-1 to -1.99	Mild Drought
-0.49 to -0.49	Near Normal	-0.5 to -0.99	Incipient Drought
0.5 - 0.99	Slightly Wet	0.49 to -0.49	Near Normal
1 - 1.49	Moderately Wet	0.5 - 0.99	Incipient Wet
1.5 - 1.99	Severely Wet	1 - 1.99	Slightly Wet
≥ 2	Extremely Wet	2 - 2.99	Moderately Wet
		3 - 3.99	Very Wet
		≥ 4	Extremely Wet

7.4 Model Development

In this study, machine learning methods and the capabilities of the SHAP method were utilized to develop a new drought index by combining individual indices into a CDI. The data used for calculating single indices within the study area were obtained from remote sensing systems and subsequently underwent normalization processes for the calculation of each index. The normalized images were cropped to cover the study

area. The input data utilized to obtain the CDI were rescaled to have a temporal resolution of monthly and a spatial resolution of 1×1 km. This was achieved by averaging the daily data on a monthly basis. The bilinear algorithm was applied for spatial rescaling, transforming all cropped images to a target resolution of 1×1 km, and they were subsequently masked to include crop areas based on crop land information obtained from GLAD. After the masking process, spatial averages were taken for each month from October to June, considering the district boundaries within the study area. Thus, a single value representing each month was obtained for each district. Using these values, the weights of the CDI equation were calculated using three different methods. Of these, PCA and empirical methods do not require a target variable, but for the SHAP method, crop yield values measured in each district were used as the target variable. To provide inputs for the SHAP algorithm, prediction models were created using the XGBoost method. The inputs of the XGBoost model were TCI, VCI, SMCI, PCI, GPP, and NDWI values, while the target variable was the detrended crop yield values. Since the GPP input has been available since 2000, the period between 2001 and 2020 was selected for modeling. During the ML model creation phase, all possible scenarios for the six inputs (63 different combinations) were tested. During the modeling phase, a random splitting strategy was embraced, with 75% of the data were used for training and 25% for testing. The districts selected for the test set were Aksehir, Altinekin, Altunhisar, Beysehir, Bor, Camardi, Cihanbeyli, Ilgin, Karapinar, and Selcuklu. In the next step, the calculation of mean SHAP values was carried out using the results of the model with high prediction accuracy. Using the SHAP feature importance values obtained, the weights of the CDI equation were determined, thereby obtaining accurate weights using a method that calculates the marginal contributions of all variables in a nonlinear manner. The correlation between the CDI equations obtained by three different methods and crop yield was examined to measure the success of the new index in representing agricultural drought. Additionally, a comparative analysis with SPEI and scPDSI values, which are also used in agricultural drought studies in the literature, was conducted. Furthermore, the similarities and differences between CDI, SPEI, and scPDSI were examined. The analysis steps are shared in Figure 7.2.

The hyperparameters of the XGBoost algorithm used in the study were optimized using the grid search method. Additionally, early stopping was applied to overcome

the overfitting problem. The search space information and optimum values required for parameter optimization of the XGBoost model are provided in Table B2.

In this study, the empirical method, used to represent conventional approaches, is the simplest method. The CDI equation obtained by assigning equal importance to all inputs is as follows. Thus, the weights of all inputs for this method were set at 0.1666 (i.e., 1/6th).

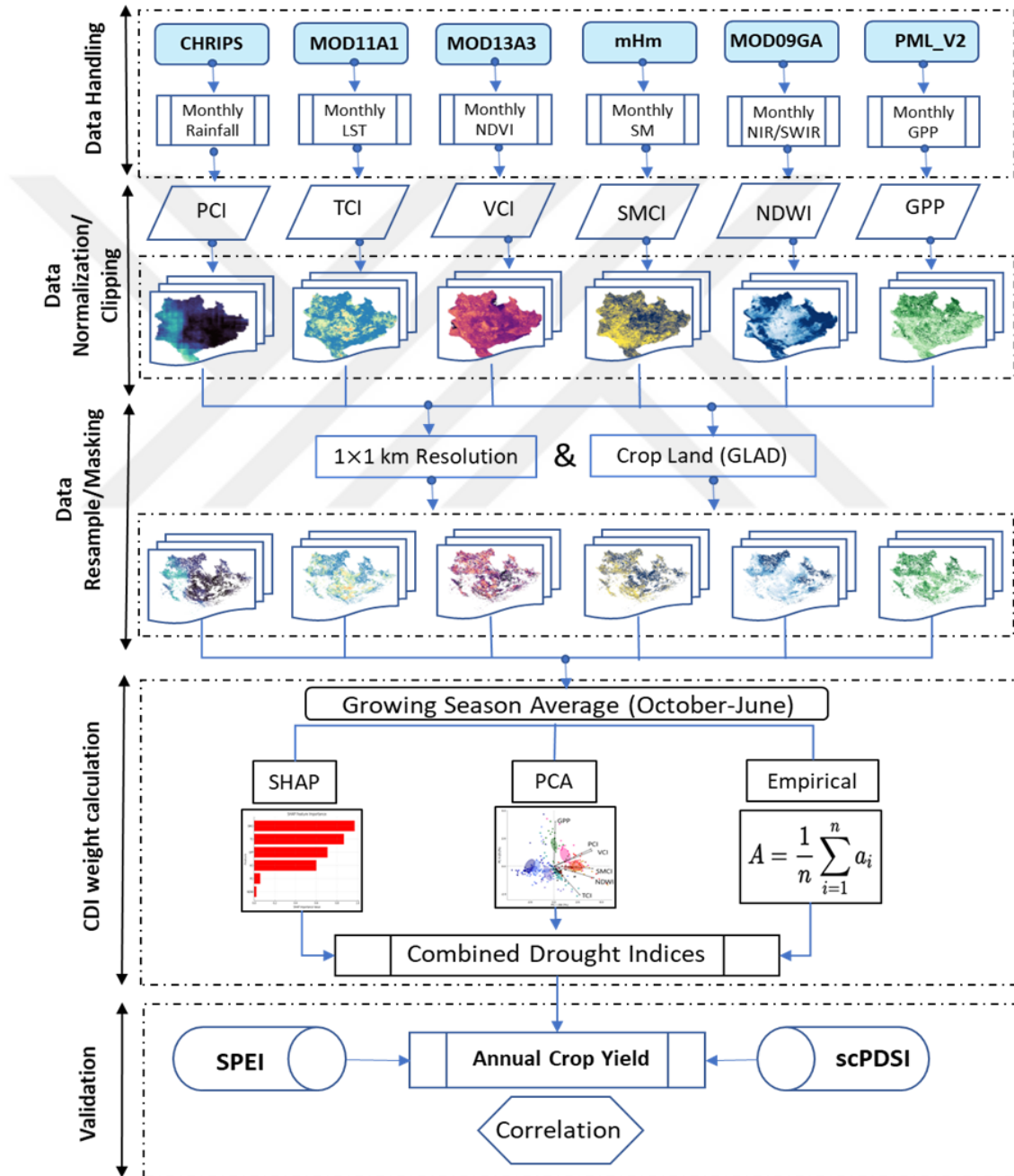


Figure 7.2 : Study flowchart.

7.5 Result And Discussion

7.5.1 Machine learning results

For the model where all inputs were used, the training R value was found to be 0.82, and the test R value was 0.68 (the prediction model results for all combinations are provided in Table B3). The training R values obtained by reducing the inputs one by one were all found to be lower than the initial model. Since the SHAP values' feature importance reflects the results obtained from the training set, the weights were calculated using the results of the model with the highest R value. In this study, to compare with PCA and empirical methods, the model using all inputs was selected, and PCA and empirical method calculations were performed on this model. Thus, a comprehensive analysis was achieved by using TCI, PCI, SMCI, VCI, NDWI, and GPP values, resulting in an extensive evaluation of factors potentially affecting drought for CDI calculations.

7.5.2 Weight obtained by SHAP

A feature importance plot was generated using the SHAP method after obtaining the prediction results from the XGBoost model (Figure B2). In generating these plots, the original feature importance values were used to recalculate relative values on a scale of 100 to determine the impact of each input on the output. NDWI was the most influential variable for yield, contributing 41%. It was followed by TCI (19%), PCI (15%), GPP (11%), VCI (8%), and SMCI (6%). GPP, SMCI, and VCI had lower importance. The CDI equation calculated using the weights obtained from the SHAP method are as follows:

$$CDI = 0.41 \times NDWI + 0.19 \times TCI + 0.15 \times PCI + 0.08 \times VCI + 0.11 \times GPP + 0.06 \times SMCI \quad (7.10)$$

7.5.3 Weight obtained by PCA

For a fair comparison with SHAP, PCA weights were obtained using the data in the training set, and the weights derived from this dataset were validated using the test set (Table B4). When examining CDI studies, information from the first component or the first two components is commonly used (Chang et al, 2022). The contribution ratios of the principal components following the PCA analysis for the districts are shown in Table B5. The first two components accounted for more than 60% of the variance.

When examining the weights obtained from the PCA analysis, the input with the highest weight is VCI at 0.23, followed by TCI at 0.19. The third highest weight is GPP at 0.18, while the fourth is NDWI at 0.17. The lowest weight is attributed to PCI, with an average of 0.10. Compared to SHAP, only the TCI input shares the same ranking, while the weight rankings for all other inputs are significantly different.

$$CDI = 0.16 \times NDWI + 0.19 \times TCI + 0.10 \times PCI + 0.23 \times VCI + 0.18 \times GPP + 0.10 \times SMCI \quad (7.11)$$

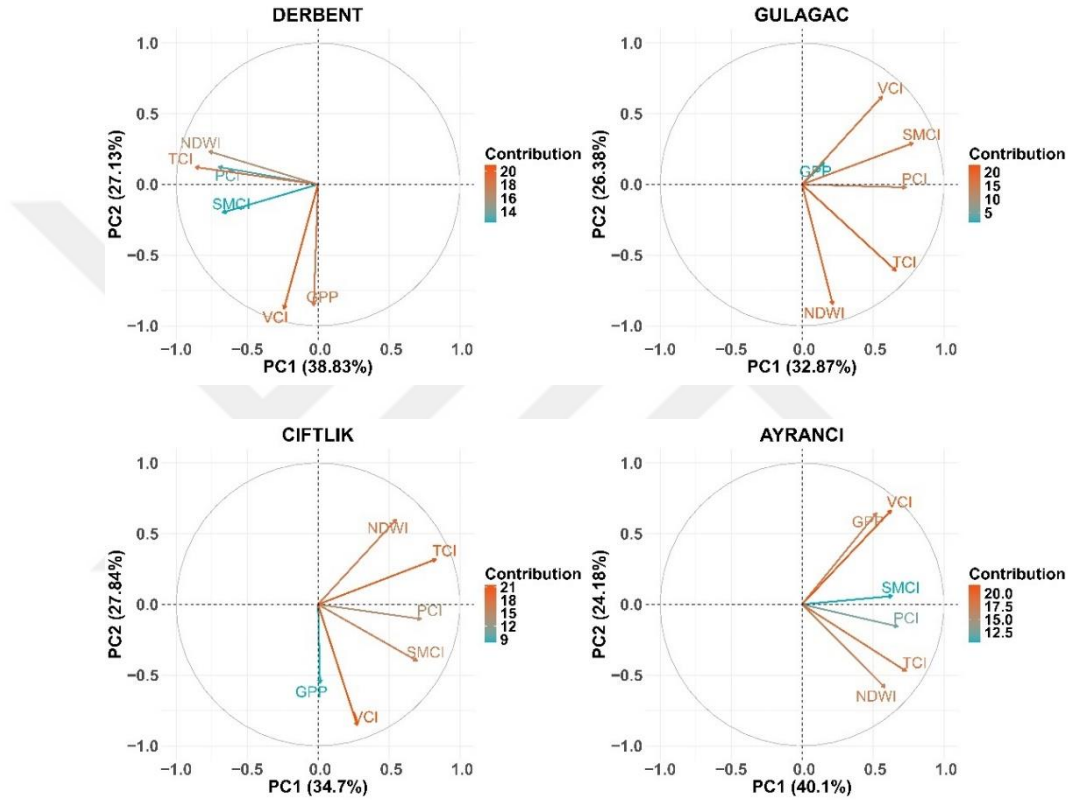


Figure 7.3 : Loading plots of PCA analysis results represent the four selected provinces within the study area, with the selected districts for Aksaray, Karaman, Konya, and Niğde being Gülagaç, Ayrançı, Derbent, and Çiftlik, respectively.

To analyze the relationship between the inputs after the PCA analysis, PCA loadings plots were created. These plots allow us to see which inputs are closely related to each other in each district and their contribution to explaining the variance. The axes represent the components that explain the variance in the data, with the direction of the vectors indicating the relationship between the variables and the components, and their length reflecting their contribution. The angle between vectors signifies the correlation between variables; small angles indicate a positive relationship, while large angles suggest a weak association. Vectors pointing in the same direction imply that

the variables exhibit similar behavior. The districts with the highest explanation ratio in the first two principal components were selected to create Figure 7.3. Among the values of the four districts in the graph, VCI provides the highest contribution to both the first (PC1) and second (PC2) components and TCI significantly contribute to PC1 and PC2. PCI and SMCI values have relatively low contributions. Additionally, a strong positive relationship can be observed between VCI and GPP in all regions, while a negative correlation can be observed between VCI and TCI and between NDWI and GPP, as these inputs intersect each other perpendicularly.

7.5.4 Crop yield correlation results obtained by three different methods

The relationship between crop yield and CDI equations calculated using three different methods was examined. The results of the correlation analysis for all districts are shared in Table 7.3. In 7 out of the 40 districts where the study was conducted, correlations were not statistically significant for all methods. Among all the significant models, the SHAP model provided the best results with an average correlation of 0.56.

Table 7.3 : Correlation Results for Different CDI Weight Calculation Methods.

District	SHAP	PCA	Empirical	District	SHAP	PCA	Empirical
Agacoren	0.008	-0.039	-0.011	Guneysinir	0.182	0.120	0.080
Akoren	0.439*	0.417*	0.438*	Guzelyurt	0.340	-0.056	0.036
Aksaray_Merkez	0.681*	0.636*	0.644*	Huyuk	0.689*	0.723*	0.708*
Aksehir	0.583*	0.319	0.389*	Ilgm	0.794*	0.759*	0.705*
Altunekin	0.547*	0.353	0.347	Kadinhani	0.829*	0.752*	0.740*
Altunhisar	0.738*	0.719*	0.716*	Karaman_Merkez	0.729*	0.549*	0.556*
Ayranci	0.269	0.008	0.008	Karapinar	0.635*	0.453*	0.454*
Beysehir	0.623*	0.591*	0.614*	Karatay	0.571*	0.445*	0.468*
Bor	0.644*	0.574*	0.565*	Kulu	0.331	0.402	0.430*
Camardi	0.504*	0.320	0.305	Meram	0.456*	0.365	0.385
Celtik	0.701*	0.487*	0.493*	Nigde_Merkez	0.547*	0.438*	0.441*
Ciftlik	0.408	0.475*	0.483*	Ortakoy	0.565*	0.444*	0.516*
Cihanbeyli	0.762*	0.689*	0.692*	Sarayonu	0.377	0.311	0.287
Cumra	0.558*	0.498*	0.480*	Sariyahsi	0.600*	0.580*	0.538*
Derbent	0.573*	0.304	0.367	Selcuklu	0.665*	0.603*	0.576*
Doganhisar	0.520*	0.477*	0.417*	Seydisehir	0.002	-0.110	-0.114
Emirgazi	0.800*	0.701*	0.713*	Tuzlukcu	0.558*	0.505*	0.550*
Eregli	0.254	0.155	0.167	Ulukisla	0.672*	0.592*	0.618*
Eskil	0.505*	0.268	0.332	Yalihuyuk	0.128	0.075	0.180
Gulagac	0.735*	0.600*	0.662*	Yunak	0.641*	0.639*	0.627*

*In bold are test set districts; * $p < 0.05$.*

Following the SHAP model, the highest average correlation was 0.47, obtained using the empirical method. Another significant finding is that the correlation results between the PCA and empirical methods are quite similar.

The correlation values between the drought indices used in this study and crop yield are presented in Table B6. When evaluating all 40 districts, the highest correlation was obtained using SHAP in 33 districts, PCA in 1 district, the empirical method in 3 districts, SPEI in 3 districts, and scPDSI in 0 districts. In the comparison among the SPEI indices themselves, SPEI_3 exhibited the highest average correlation at 0.39, while SPEI_9 had the lowest average correlation (0.26). Among all the methods, scPDSI showed the lowest average correlation (0.22). A box plot illustrating the correlation values for the districts in the training and test sets is presented in Figure 7.4. This figure, which serves as a visual representation of Table B6, allows for the easy identification of the distribution, minimum and maximum values, and median values of the correlation results for each method. The highest median values in the training and test sets belong to CDI(SHAP), while CDI(Empirical) and CDI(PCA) show a very similar distribution. It is observed that SPEI_1, SPEI_3, and SPEI_6 produce similar results in the training set, but the differences become apparent in the districts within the test set. The spatial distribution of the correlation across the 40 districts in the study area is shown in Figure 7.5. SHAP demonstrates high correlations in a large number of districts, followed closely by PCA. Higher correlations between the drought indices and crop yield are observed in the northern regions of the study area, while a decrease in correlations is observed in the southern regions, which are at a higher elevation. Topographic variations in mountainous regions have the potential to influence precipitation distribution and the conditions under which wheat can be cultivated. Mountainous regions may give rise to orographic effects, which can result in greater (or lower) precipitation accumulation. Conversely, lower elevations may exhibit a more homogeneous distribution of rainfall. This may result in a weakening of the correlation between the drought index and yield in the southern and mountainous regions.

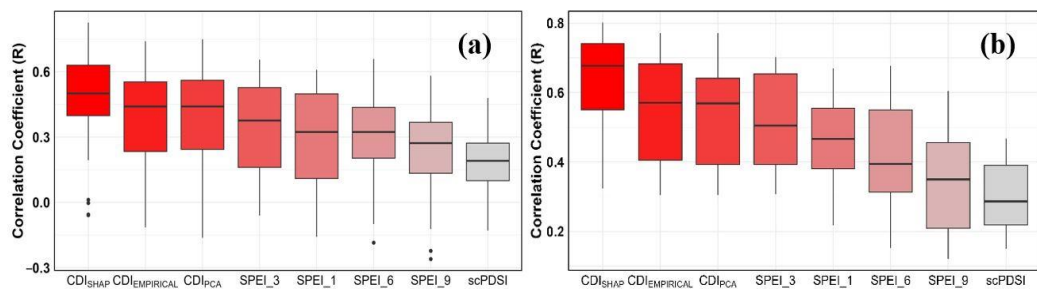


Figure 7.4 : Box plots of correlation coefficients between drought indices and crop yield (a) train set, (b) test set.

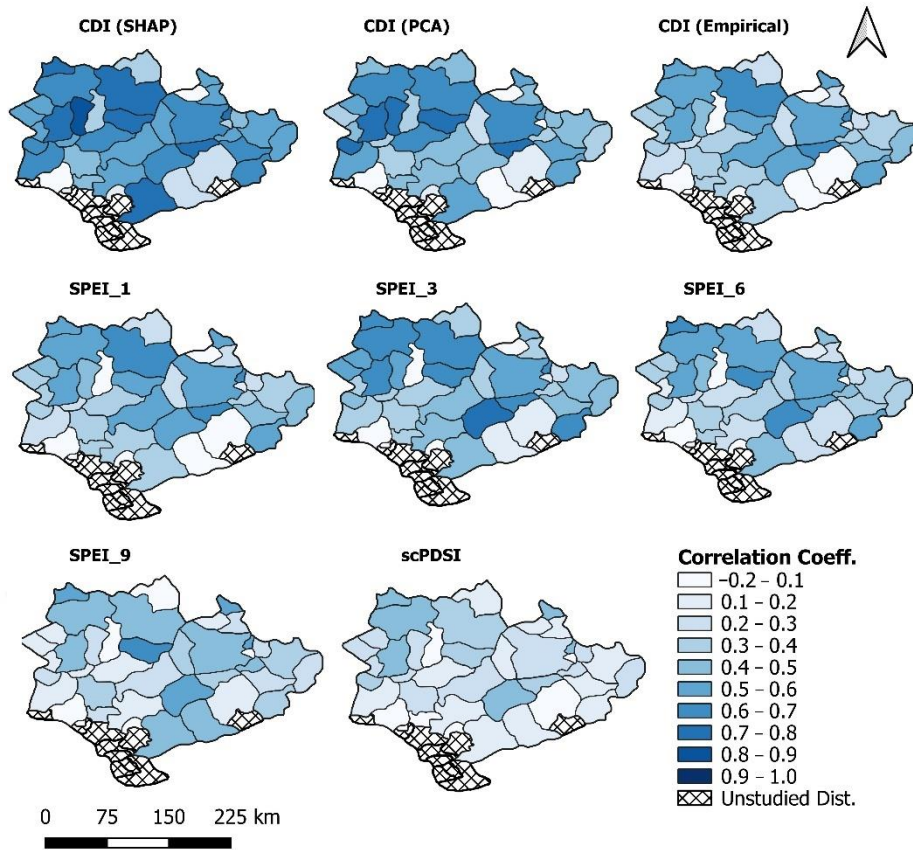


Figure 7.5 : Spatial distribution of correlation results between drought indices and crop yield.

To assess the temporal relationship between the drought indices and crop yields for the districts in the test set, time series of the indices and crop yield are presented in Figure 7.6. The standardized crop yields, along with CDI and SPEI indices which belong to the same classification, are placed on the left axis, while the scPDSI values are placed on the right axis for visual comparison. To avoid clutter, only the SPEI value with the highest correlation at the district level was used, rather than all SPEI indices. Overall, a good alignment between CDI and SPEI with crop yield was observed. In contrast, scPDSI failed to capture crop yield trends in certain years, a conclusion that can be quickly drawn from the visual inspection. In all districts, the relationship between all indices and crop yield reached its peak during the wet year of 2015. Between 2004 and 2009, scPDSI deviated from the crop yield trends, showing a smooth movement without significant fluctuations, whereas CDI and SPEI responded promptly to crop yield anomalies.

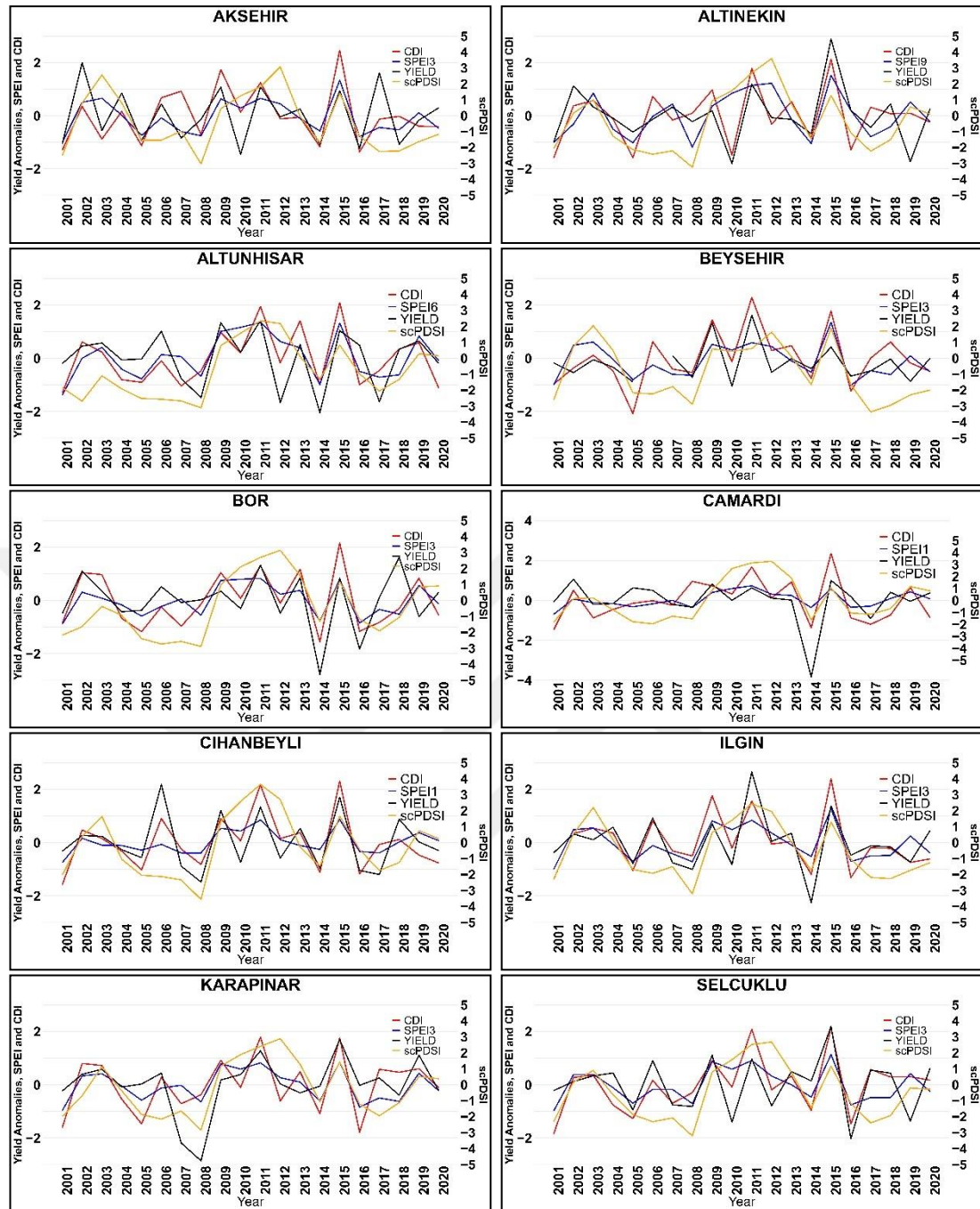


Figure 7.6 : Annual time series of drought indices and crop yield of the test districts.

7.5.5 Spatiotemporal distribution of drought indices

In this study, several maps were prepared to evaluate the spatial distribution of drought indices. Drought maps for the crop development periods in 2016, which was consistently represented as a drought year across all indices, are presented in Figure 7.7.

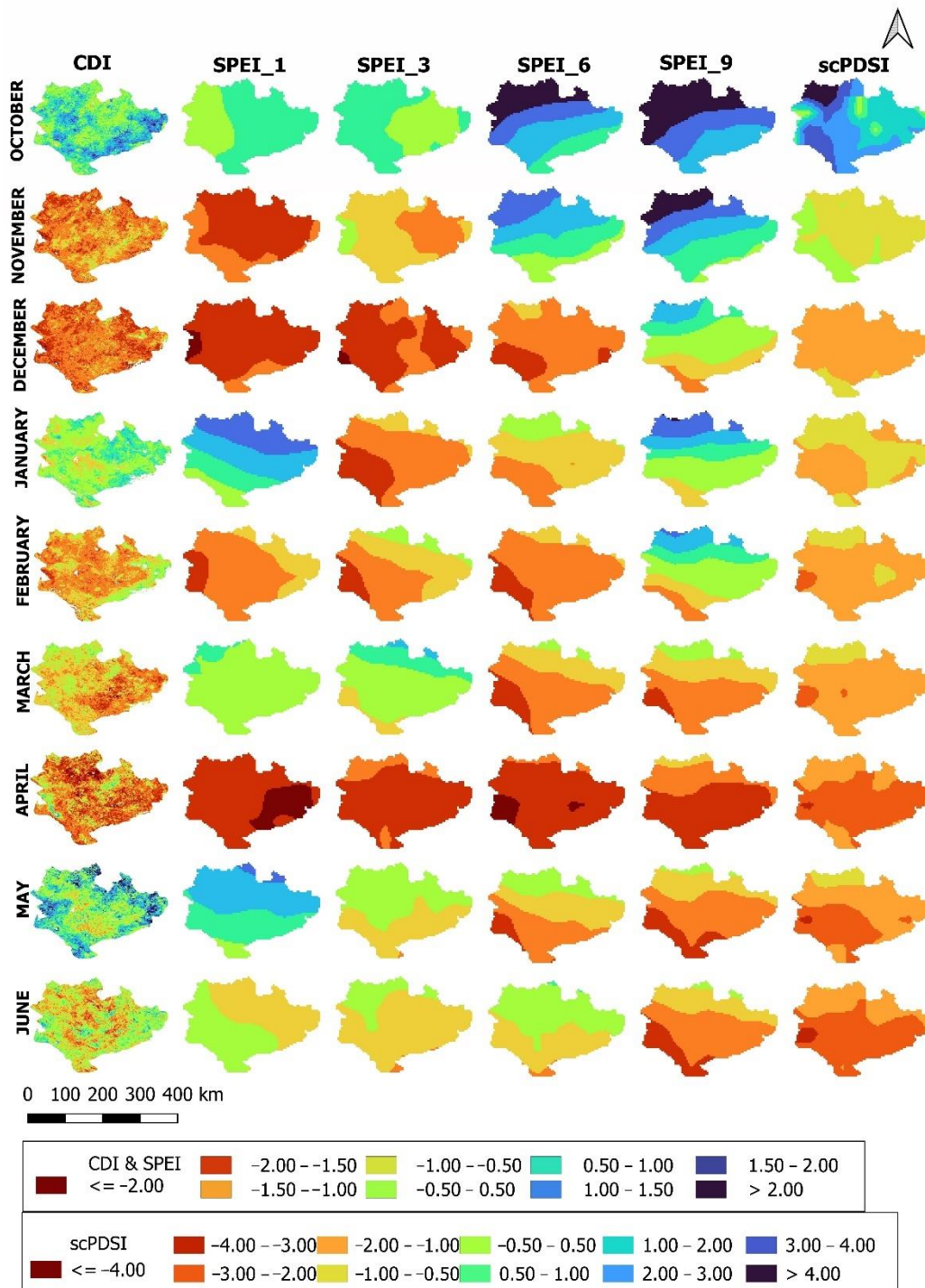


Figure 7.7 : CDI, SPEI and scPDSI maps for all months of the winter wheat growing season during the 2016 dry year.

In the first month of the development period, all indices indicate a wet month; however, according to the CDI, this is predominantly observed in the southeastern region, while SPEI and scPDSI show wetter conditions in the northern regions. Additionally, it is observed that as the SPEI month index increases, there is an increase in wet areas. In November 2016, CDI indicates widespread drought, while SPEI₆ and

SPEI_9 show increasing wetness in the north. December also shows high drought according to CDI, SPEI_1, SPEI_3, SPEI_6, and scPDSI. In January, CDI suggests near-normal to slightly wet conditions, but SPEI indices are inconsistent, with some showing drought and others indicating wet areas in the north. February sees CDI and scPDSI aligning, showing drought in the west. By March, CDI notes drought in the east, with mixed conditions across SPEI indices. April is dominated by drought across all indices. In May, CDI shows wet conditions near the borders and drought in the central area of the study region, while SPEI transitions from wet to drought. June ends with normal to mild drought, with scPDSI highlighting widespread drought for most of the period.

The variability of the CDI over the study period is illustrated in Figure 7.8. The maps were generated using the average values for the nine months of the growing season, from October to June. The years 2001, 2005, 2008, 2014, 2016, and 2020 were mild to moderate drought years, while 2011 and 2015 were moderately to severely wet years. Drought years generally show variability along the north–south axis. In the maps produced with the same resolution as CDI (1 km), a resolution of 5 km was chosen for mapping SPEI and scPDSI values to avoid issues related to scaling. Additionally, in Figure 7.9, which analyzes changes over the years, SPEI_3 values were used because they generally align better with crop yield. Similar to CDI, SPEI_3 was classified as mild drought across most of the study area in 2001, 2005, 2008, and 2016. The years 2011 and 2015 were classified as slightly wet and moderately wet, respectively. Unlike CDI values, which were generally classified as near-normal and mild drought in 2004, 2007, and 2020, SPEI_3 remained in the near-normal class across the entire region during these years. However, in 2014 and 2019, CDI and SPEI_3 were observed to belong to similar drought classes in similar areas. When examining Figure 7.10, which illustrates the distribution of scPDSI values across the study area, 2001 is mildly to moderately drought-affected, similar to CDI and SPEI_3. However, when analyzing the spatial distribution, drought is more pronounced in the western and southern regions according to scPDSI and predominantly observed in the northern regions in CDI and SPEI_3. In 2002, a drought occurred in the northeastern region, which was not captured by CDI and SPEI_3.

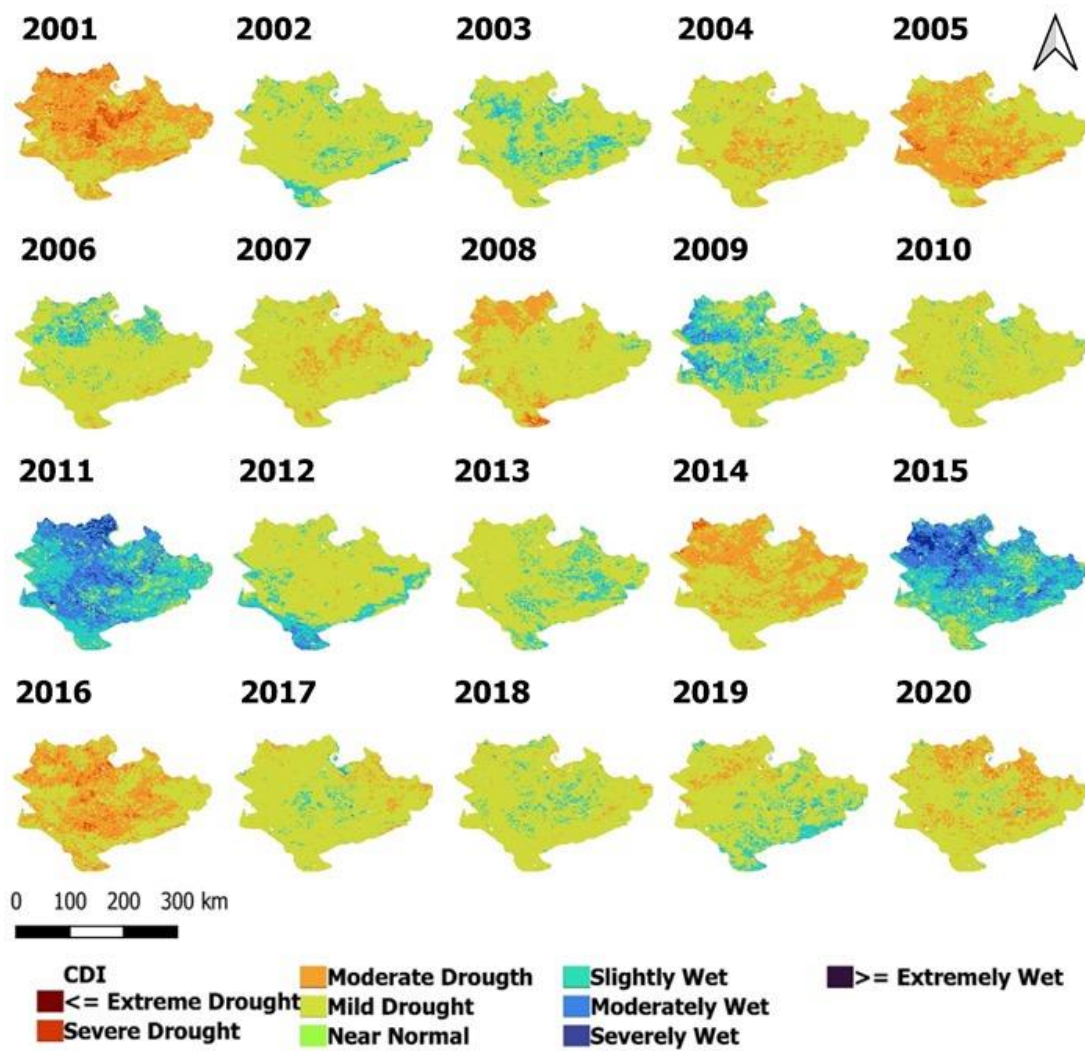


Figure 7.8 : The CDI drought map for winter wheat growing seasons in the study region.

One noteworthy feature of the scPDSI values is the variation of drought along the northeast–southwest axis during consecutive years, particularly in 2007, 2008, and 2009. While 2012 could be classified as wet according to CDI and SPEI_3, scPDSI placed it in the very wet category. In contrast to CDI and SPEI_3, 2017 is generally depicted as drier in scPDSI.

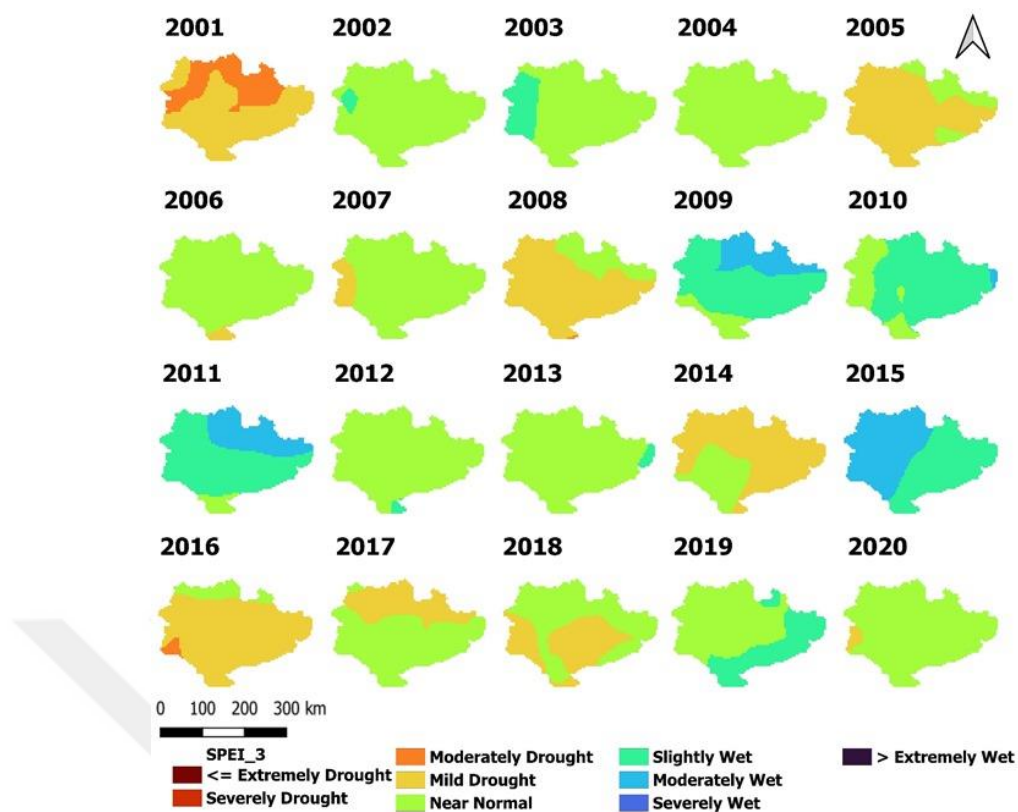


Figure 7.9 : SPEI's drought map for the winter wheat growing season across the study.

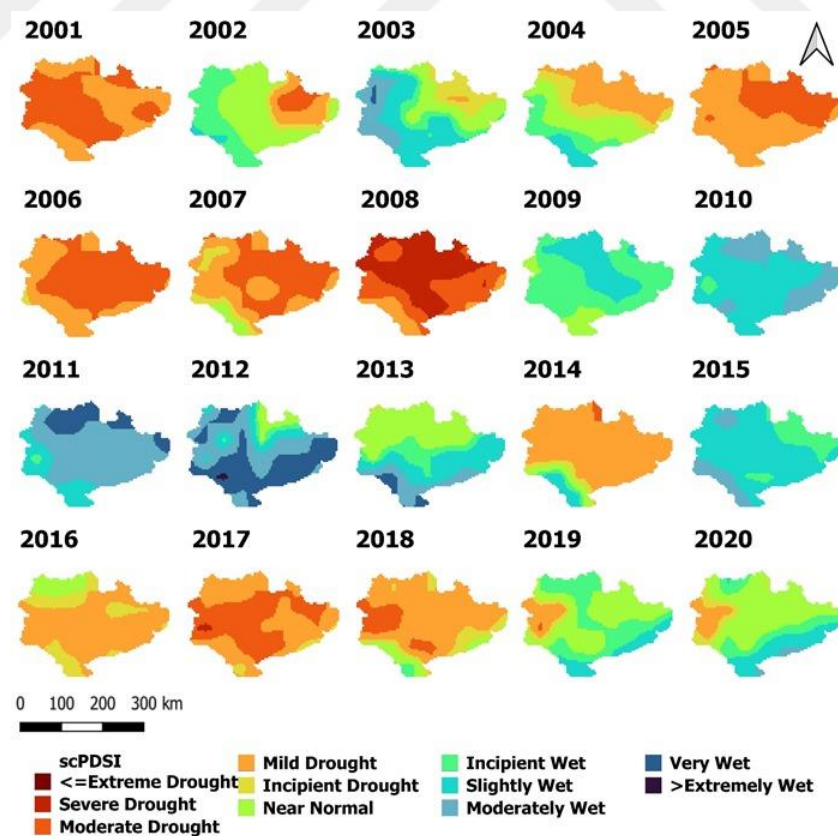


Figure 7.10 : The same as Figure 9 but for scPDSI.

7.5.6 Long term evolution of drought indices

In this study, time series were generated to analyze the long-term drought values obtained through the CDI. As shown in Figure 7.11, the CDI and SPEI time series are presented on the left y-axis, while the scPDSI time series are depicted on the right y-axis. The reason for using two different vertical axes is that the scPDSI varies between -5 and $+5$, which might pose challenges in the multi evaluation of it with CDI and SPEI, which fluctuate between -2 and $+2$. One of the most significant findings is the similar behavior exhibited by the CDI and SPEI. The fluctuations in scPDSI values are less pronounced compared to the other two indices. SPEI, in terms of its composition, utilizes precipitation and ET calculations, which are largely based on temperature. This similarity in inputs results in SPEI displaying a similar trend to CDI. However, SPEI focuses only on PET, which is insufficient to represent surface water processes fully (Abrar Faiz et al, 2022), and lacks input that provides information about vegetation conditions, leading to significant divergences in some cases. For instance, in the Altinekin district, SPEI values indicate a mild drought and fewer fluctuations compared to CDI values (Figure 7.11).

The scPDSI generally exhibits a similar pattern to CDI and SPEI, but significant differences are observed in certain years. For example, scPDSI indicates the presence of a prolonged drought from 2003 to 2009, with drought values remaining constant over several months. A similar pattern is observed between 2016 and 2020. In the districts of Aksehir, Beysehir, and Ilgin, a long-term drought with consistent values is present. One of the reasons for this could be that scPDSI values are calculated using the drought values from the previous month, which contributes to the persistence of similar values over extended periods.

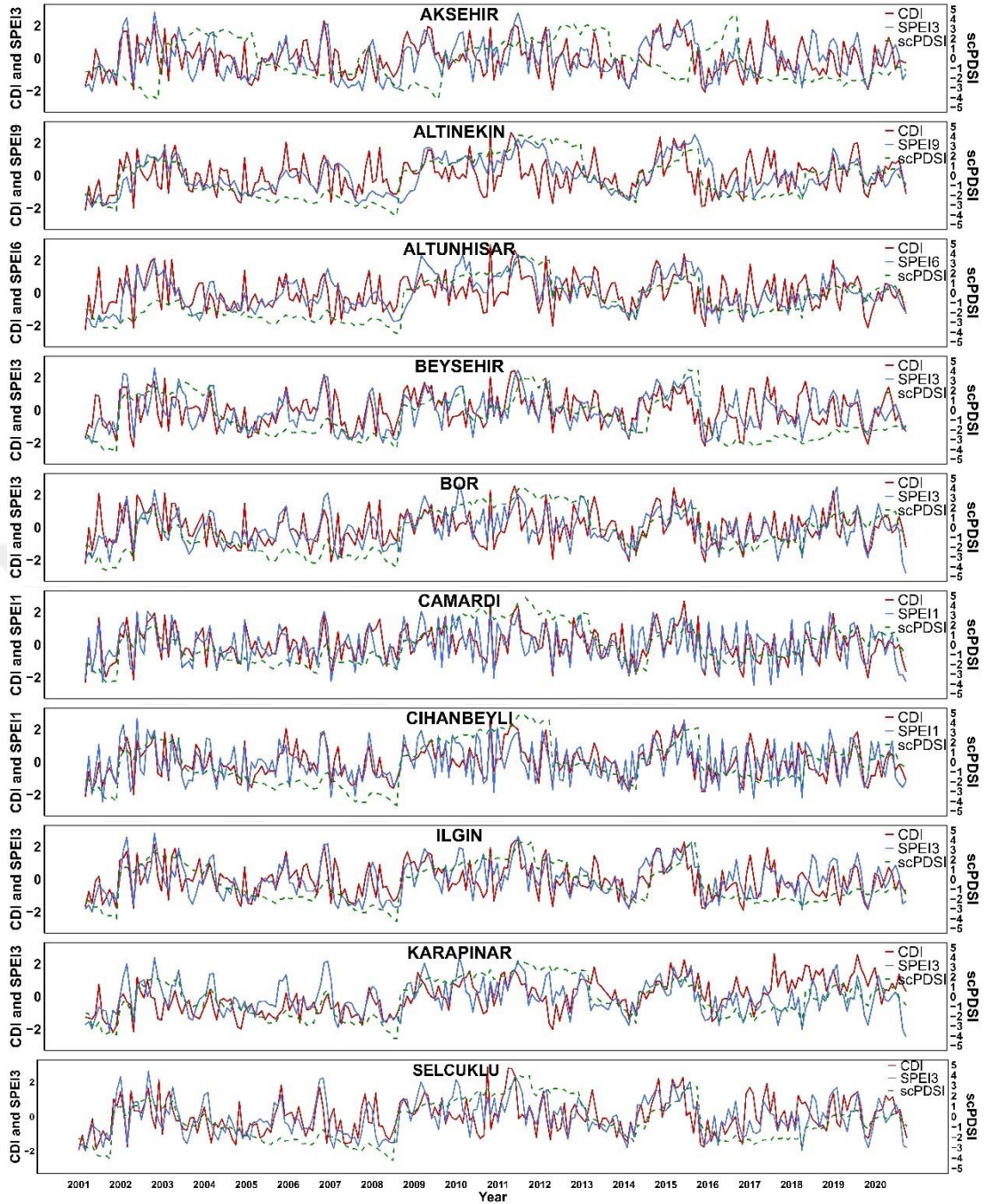


Figure 7.11 : Long-term time series of CDI, SPEI and scPDSI for the districts in the test set.

7.5.7 In-depth analysis of feature importance and dependencies

In this study, the SHAP algorithm's feature importance values were used in the process of optimizing the weights for the CDI equation. This method not only calculates the importance of features but also visually represents their impact on the output through graphs. The SHAP summary plot in Figure 7.12 describes how individual indices affect crop yield and demonstrates their nonlinear responses.

Extremely low NDWI has a very strong negative impact, while higher values have a relatively lower positive impact. The second most important index, TCI, also shows that higher values lead to increased crop yield, while medium and low values have a decreasing effect on crop yield. The PCI slightly increases crop yield at average values, while very low values decrease it. GPP and VCI have a similar effect; higher values lead to an increase in crop yield, while lower values have a minimal negative effect. As for SMCI, very high values cause a slightly positive effect, while medium values have a slightly negative effect. However, SMCI turned out to be the individual index with both the smallest impact size and the narrowest range of effects.

Research examining the relationship between NDWI values and crop yield has also confirmed that NDWI values significantly impact crop yield (Tuvdendorj et al, 2019). There is no consensus on which input is the most important in these studies, as individual indices can vary greatly regionally and may have different levels of influence depending on the climatic classification of the area.

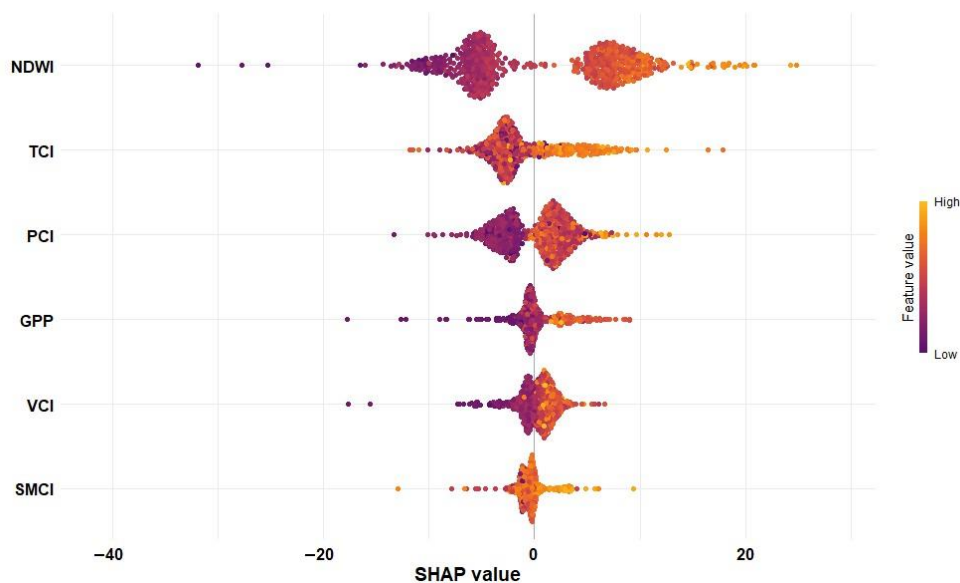


Figure 7.12 : SHAP summary plot calculated from predictions obtained using the XGBoost method.

One of the SHAP method's outputs is the dependence plots, which analyze the non-linear effects of each input on the output. Figure 7.13 presents dependence plots for the six indices used in predicting crop yield. A detailed examination reveals that normalized GPP values between 0.2 and 0.3 significantly reduce crop yield, while values above 0.6 have a positive or stable effect. The NDWI plot shows a nearly

monotonic positive relationship, with significant yield increases above 0.55. PCI values exhibit a decreasing effect between 0.2 and 0.35, followed by a fluctuating increase above 0.6. SMCI shows a neutral effect up to 0.75, with positive effects emerging afterward. TCI values demonstrate a negative effect up to 0.5, followed by an increase, while VCI shows a negative effect up to 0.4, turning slightly positive above 0.45. These analyses provide a comprehensive view of both the weights and nonlinear effects of the indices on crop yield.

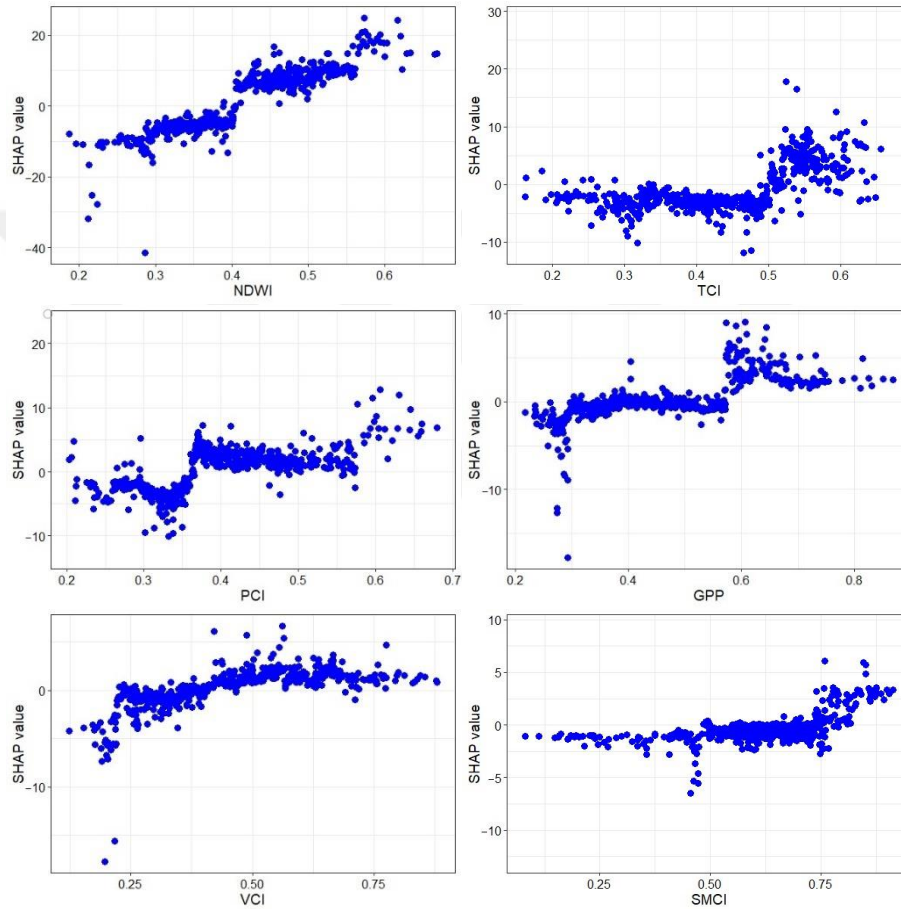


Figure 7.13 : Dependences plots obtained after SHAP analysis of single indices used to calculate CDI.

7.5.8 Future directions, recommendations and practical applications

Although the CDI developed in this study was created using innovative algorithms, it has certain limitations. The winter wheat yield values used in the study are limited to 20 years. Developing models with more extensive data is encouraged, noting that remote sensing time series are often limited in duration. Additionally, it has not been investigated whether the remote sensing data used in the study area require any additional corrections; the observations used are derived from established global

algorithms. Researchers are strongly advised to explore this aspect and incorporate the need for corrections between the remote sensing data and ground observations in their study areas. In addition to improving temporal resolution, enhancements in spatial resolution could help with yield modeling in individual fields. The present study's XGBoost hyperparameters optimization was challenging, and grid search algorithm addressed this challenge. Nevertheless, future research should investigate the potential of advanced heuristic methods. Given that XGBoost was employed for the construction of predictive models, TreeSHAP was utilized. In the context of larger datasets and broader regions, it is possible that deep learning models and the KernelSHAP method may yield superior results.

The model inputs include remote sensing datasets containing information on temperature, precipitation, soil moisture, vegetation condition, plant water status, and primary productivity. However, the study did not use evapotranspiration (ET) values. Future studies could benefit from incorporating the Evaporative Stress Index (Anderson et al, 2007), which represents the ratio of potential ET to actual ET. Moreover, variables representing global ocean–atmosphere interactions, such as El Niño and La Niña, could further improve CDI calculations if they include independent information from the indices explored here.

Our study focused on one of the largest grain-producing regions in Türkiye. The primary motivation for this work was the fact that the area predominantly practices rain-fed agriculture, as crops grown in rainfed areas tend to respond more sensitively to meteorological variables. In regions where irrigation data and the genetic characteristics of the crops used are available, it is recommended that these factors be included in the study to develop even more advanced indices.

Just as drought affects all sectors, the sector where its direct impacts are often most clearly seen is agriculture (Sharma, 2024). Among the most important stakeholders in drought studies are companies operating in the agricultural sector, local producers, and traders involved in the sale and marketing of agricultural products (Orimoloye, 2022). We hope that this study will contribute to the development of agricultural drought monitoring systems at the local level to enable decision-makers to address complex drought monitoring processes in greater detail and accelerate effective decision making.

In the USA, a country known for its advanced drought monitoring systems, CDIs are being developed using similar applications (Svoboda et al, 2002). Various CDIs can be generated using different combinations of inputs. Countries like Türkiye, where agricultural products constitute a significant share of the gross national product, could improve decision-making processes by utilizing CDI-like indices. Since all the inputs used in creating CDI indices are obtained from free sources and are continuously updated, this study is valuable for the development of online drought monitoring systems that provide local-scale information to end-users.

7.6 Conclusion

In this study, the newly developed combined drought index, designed to more effectively monitor agricultural drought events, may offer more effective usage when used in conjunction with existing indices. The use of SHAP algorithm determined weights and a mesoscale hydrologic model enhanced the reliability and accuracy of this new index. Notably, the index exhibited a higher correlation with wheat yields than widely used indices like SPEI and scPDSI, marking a significant contribution to agricultural drought monitoring. Furthermore, it is anticipated that the integration of increasing volumes of data from remote sensing systems with technological advancements will further enhance the efficiency of agricultural drought monitoring systems in the future. We anticipate that CDIs enhanced using ML techniques will continue to improve agricultural drought monitoring.



8. CONCLUSIONS AND RECOMMENDATIONS

This thesis has made significant strides in advancing the understanding of drought dynamics and improving the methodologies for drought monitoring, particularly in regions with semi-arid climates such as Türkiye. Through a comprehensive and multifaceted approach, the study has successfully developed a Combined Drought Index (CDI) that integrates various environmental factors, including agricultural, meteorological, and hydrological data. The novel contribution of this research lies in its integration of Gross Primary Productivity (GPP) data—a key indicator of ecosystem health—into drought monitoring, leveraging machine learning techniques to enhance both the accuracy and responsiveness of drought indices.

One of the primary conclusions of this study is that the inclusion of GPP values, derived from remote sensing data, provides a more dynamic and ecologically relevant metric for assessing drought impacts on agricultural productivity. Unlike traditional drought indices that rely heavily on precipitation and evaporation data, the use of GPP allows for a more comprehensive evaluation of drought, as it directly reflects the photosynthetic activity and health of vegetation. This is particularly important in agricultural drought monitoring, where the direct impact on crops needs to be accurately assessed. The spatio-temporal analysis of GPP trends across Türkiye revealed significant regional variations, underscoring the need for region-specific drought indices that account for local environmental conditions and crop types.

Another critical outcome of this research is the development of a new algorithm for imputing missing solar radiation data, which significantly enhances the completeness and reliability of the datasets used in drought models. This, combined with the forward-looking predictive model for evaporation, contributes to more accurate and robust drought monitoring tools. These innovations address a longstanding challenge in environmental and hydrological studies, where incomplete or unreliable data can hinder effective drought prediction and management.

The systematic review of missing data imputation methods and the development of a novel approach for solar radiation data imputation also contribute to the broader field

of environmental sciences. By providing a detailed comparison of existing methods and introducing a new, more accurate technique, this study sets a foundation for future research in data processing for environmental models. The forward-looking evaporation model presented in this thesis further enhances the predictive capability of drought indices, allowing for better anticipation of drought events and their potential impacts.

Furthermore, the thesis's bibliometric analysis of GPP research highlights the growing significance of GPP as a key metric in both ecological and drought studies. This analysis not only sheds light on global research trends but also identifies gaps in the current literature, providing a roadmap for future research efforts in this field.

While the contributions of this thesis are substantial, they also pave the way for further research and development. There are several promising avenues for future studies that could build upon the findings of this work:

- **Expansion of Drought Index Models:** Future research could explore the integration of additional environmental variables into the Combined Drought Index, such as soil moisture data from more advanced hydrological models or evapotranspiration rates. By incorporating a wider range of indicators, the CDI could be further refined to capture the complex interactions between climate, soil, and vegetation during drought events.
- **Application of CDI to Other Regions:** While this study focused on Türkiye, the methodologies developed here could be applied to other regions with similar climatic challenges, such as parts of the Mediterranean, Middle East, or Central Asia. Comparative studies between different semi-arid regions could yield insights into the universal applicability of the CDI and identify region-specific adaptations that may be necessary.
- **Long-Term Drought Impact Analysis:** Future studies could focus on the long-term impacts of drought on ecosystem services and agricultural systems. By using the CDI over extended time periods, it would be possible to assess how recurring drought events affect crop productivity, biodiversity, and water resources over time. This would provide valuable information for developing sustainable agricultural practices and water management strategies.

- **Real-Time Drought Monitoring and Early Warning Systems:** Leveraging advances in machine learning and remote sensing technologies, future research could focus on developing real-time drought monitoring systems based on the CDI. These systems could provide early warnings of impending droughts, allowing for more timely and effective responses from policymakers and agricultural managers. The integration of cloud-based platforms for data processing and dissemination could further enhance the utility of such systems.
- **Incorporation of Climate Change Projections:** As climate change continues to alter weather patterns, future studies could incorporate climate models to predict how drought frequency and intensity might evolve in the coming decades. By integrating these projections into the CDI framework, it would be possible to assess how agricultural productivity and water resources will be affected under different climate scenarios. This would provide valuable information for long-term planning and risk management.
- **Impact of Drought on Socio-Economic Factors:** Another area of potential research is the examination of how drought affects socio-economic conditions in rural areas. By linking the CDI with socio-economic data, future studies could assess how drought-induced losses in agricultural productivity contribute to rising food prices, rural poverty, and migration. This would offer a more holistic understanding of drought impacts, beyond its environmental consequences.
- **Enhanced Machine Learning Techniques:** While the current study employed machine learning for the development of the CDI, future research could explore more advanced techniques, such as deep learning or ensemble models, to improve the accuracy of drought predictions. These approaches could also be used to identify non-linear relationships between different environmental variables, potentially leading to even more accurate drought indices.
- **Stakeholder Involvement and Policy Integration:** Finally, future studies could focus on integrating the CDI into decision-making processes by involving key stakeholders, such as farmers, water resource managers, and policymakers. By developing user-friendly tools and interfaces for accessing CDI data,

researchers could ensure that the CDI becomes a practical tool for managing agricultural drought in real-world settings.

In conclusion, this thesis has laid the groundwork for a more nuanced and accurate understanding of drought dynamics through the development of a novel Combined Drought Index. By integrating machine learning techniques and remote sensing data, the CDI represents a significant advancement in drought monitoring, particularly for agricultural applications. The findings and methodologies developed in this study offer a wealth of opportunities for future research, both in Türkiye and globally, as the need for effective drought management becomes increasingly urgent in the face of climate change.



REFERENCES

- Abbaszadeh, M., Soltani-Mohammadi, S., & Ahmed, A. N.** (2022). Optimization of support vector machine parameters in modeling of Iju deposit mineralization and alteration zones using particle swarm optimization algorithm and grid search method. *Computers & Geosciences*, 165, 105140. doi:10.1016/j.cageo.2022.105140.
- Abrar Faiz, M., Zhang, Y., Tian, X., Tian, J., Zhang, X., Ma, N., & Aryal, S.** (2022). Drought index revisited to assess its response to vegetation in different agro-climatic zones. *Journal of Hydrology*, 614(PA), 128543. doi:10.1016/j.jhydrol.2022.128543.
- Abudu, S., Bawazir, A. S., & King, J. P.** (2010). Infilling Missing Daily Evapotranspiration Data Using Neural Networks. *Journal of Irrigation and Drainage Engineering*, 136(5), 317–325. doi:10.1061/(ASCE)IR.1943-4774.0000197.
- Adarsh, S., & Janga Reddy, M.** (2019). Evaluation of trends and predictability of short-term droughts in three meteorological subdivisions of India using multivariate EMD-based hybrid modelling. *Hydrological Processes*, 33(1), 130–143. doi:10.1002/hyp.13316.
- Adarsh, S., Sanah, S., Murshida, K. K., & Nooramol, P.** (2018). Scale dependent prediction of reference evapotranspiration based on Multi-Variate Empirical mode decomposition. *Ain Shams Engineering Journal*, 9(4), 1839–1848. doi:10.1016/j.asej.2016.10.014.
- Afrifa-Yamoah, E., Mueller, U. A., Taylor, S. M., & Fisher, A. J.** (2020). Missing data imputation of high-resolution temporal climate time series data. *Meteorological Applications*, 27(1), 1–18. doi:10.1002/met.1873.
- AghaKouchak, A., Farahmand, A., Melton, F. S., Teixeira, J., Anderson, M. C., Wardlow, B. D., & Hain, C. R.** (2015). Remote sensing of drought: Progress, challenges and opportunities. *Reviews of Geophysics*, 53(2), 452–480. doi:10.1002/2014RG000456.
- Aghili, M., Tabarestani, S., & Adjouadi, M.** (2022). Addressing the missing data challenge in multi-modal datasets for the diagnosis of Alzheimer's disease. *Journal of Neuroscience Methods*, 375, 109582. doi:10.1016/j.jneumeth.2022.109582.
- Akın, M., Eydurán, S. P., Rakszegi, M., Yıldırım, K., & Rocha, J. M.** (2023). Statistical modeling applications to mitigate the effects of climate change on quality traits of cereals: A bibliometric approach. In *Developing Sustainable and Health Promoting Cereals and Pseudocereals* (pp. 381–396). Retrieved from <https://www.sciencedirect.com/science/article/abs/pii/B9780323905664000096>

- Aksu, H., & Akgül, M. A.** (2020). Performance evaluation of CHIRPS satellite precipitation estimates over Turkey. *Theoretical and Applied Climatology*, 142(1–2), 71–84. doi:10.1007/s00704-020-03301-5.
- Aksu, H., Cavus, Y., Aksoy, H., Akgul, M. A., Turker, S., & Eris, E.** (2022). Spatiotemporal analysis of drought by CHIRPS precipitation estimates. *Theoretical and Applied Climatology*, 148(1–2), 517–529. doi:10.1007/s00704-022-03960-6.
- Aksu, H., Taflan, G. Y., Yaldiz, S. G., & Akgül, M. A.** (2023). Evaluation of IMERG for GPM satellite-based precipitation products for extreme precipitation indices over Türkiye. *Atmospheric Research*, 291, 106826. doi:10.1016/j.atmosres.2023.106826.
- Alavi, N., Warland, J. S., & Berg, A. A.** (2006). Filling gaps in evapotranspiration measurements for water budget studies: Evaluation of a Kalman filtering approach. *Agricultural and Forest Meteorology*, 141(1), 57–66. doi:10.1016/j.agrformet.2006.09.011.
- Ali, A., Abu-Elkheir, M., Atwan, A., & Elmogy, M.** (2023). Missing values imputation using Fuzzy K-Top Matching Value. *Journal of King Saud University - Computer and Information Sciences*, 35(1), 426–437. doi:10.1016/j.jksuci.2022.12.011.
- Altan, N. T., & Ustundag, B. U.** (2012). Reconstruction of missing meteorological data using wavelet transform. *2012 1st International Conference on Agro-Geoinformatics, (Agro-Geoinformatics)*, 1-7. doi:10.1109/Agro-Geoinformatics.2012.6311644.
- Anav, A., Friedlingstein, P., Beer, C., Ciais, P., Harper, A., Jones, C., ... Zhao, M.** (2015). Spatiotemporal patterns of terrestrial gross primary production: A review. *Reviews of Geophysics*, 53(3), 785–818. doi:10.1002/2015RG000483.
- Anda, A., Soos, G., Teixeira, J. A., & Kozma-bognar, V.** (2015). Agricultural and Forest Meteorology Regional evapotranspiration from a wetland in Central Europe, in a 16-year period without human intervention. *Agricultural and Forest Meteorology*, 205, 60–72. doi:10.1016/j.agrformet.2015.02.010.
- Anderson, M. C., Norman, J. M., Mecikalski, J. R., Otkin, J. A., & Kustas, W. P.** (2007). A climatological study of evapotranspiration and moisture stress across the continental United States based on thermal remote sensing: 1. Model formulation. *Journal of Geophysical Research: Atmospheres*, 112(D10). doi:10.1029/2006JD007506.
- Andridge, R. R., & Little, R. J. A.** (2010). A review of hot deck imputation for survey non-response. *International Statistical Review*, 78(1), 40–64. doi:10.1111/j.1751-5823.2010.00103.x.
- Ang, T. zhang, Salem, M., Kamarol, M., Das, H. S., Nazari, M. A., & Prabakaran, N.** (2022). A comprehensive study of renewable energy sources: Classifications, challenges and suggestions. *Energy Strategy Reviews*, 43(August), 100939. doi:10.1016/j.esr.2022.100939.

- Anjomshoaa, A., & Salmanzadeh, M.** (2019). Filling missing meteorological data in heating and cooling seasons separately. *International Journal of Climatology*, 39(2), 701–710. doi:10.1002/joc.5836.
- Arab, S. T., Noguchi, R., & Ahamed, T.** (2022). Yield loss assessment of grapes using composite drought index derived from landsat OLI and TIRS datasets. *Remote Sensing Applications: Society and Environment*, 26(September 2021), 100727. doi:10.1016/j.rsase.2022.100727.
- Araghi, A., Adamowski, J., & Martinez, C. J.** (2020). Comparison of wavelet-based hybrid models for the estimation of daily reference evapotranspiration in different climates. *Journal of Water and Climate Change*, 11(1), 39–53. doi:10.2166/wcc.2018.113.
- Araneda-Cabrera, R. J., Bermúdez, M., & Puertas, J.** (2021). Assessment of the performance of drought indices for explaining crop yield variability at the national scale: Methodological framework and application to Mozambique. *Agricultural Water Management*, 246, 106692. doi:10.1016/j.agwat.2020.106692.
- Aria, M., & Cuccurullo, C.** (2017). bibliometrix : An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*, 11(4), 959–975. doi:10.1016/j.joi.2017.08.007.
- Avand, M., Moradi, H., & Ramazanzadeh lasboyee, M.** (2022). Predicting temporal and spatial variability in flood vulnerability and risk of rural communities at the watershed scale. *Journal of Environmental Management*, 323, 116261. doi:10.1016/j.jenvman.2022.116261.
- Badgley, G., Anderegg, L. D. L., Berry, J. A., & Field, C. B.** (2019). Terrestrial gross primary production: Using NIRV to scale from site to globe. *Global Change Biology*, 25(11), 3731–3740. doi:https://doi.org/10.1111/gcb.14729.
- Bagherzadeh, J., & Asil, H.** (2019). A review of various semi-supervised learning models with a deep learning and memory approach. *Iran Journal of Computer Science*, 2(2), 65–80. doi:10.1007/s42044-018-00027-6.
- Baig, M. H. A., Abid, M., Khan, M. R., Jiao, W., Amin, M., & Adnan, S.** (2020). Assessing Meteorological and Agricultural Drought in Chitral Kabul River Basin Using Multiple Drought Indices. *Remote Sensing*, 12(9), 1417. doi:10.3390/rs12091417
- Balogun, A. A., Adegoke, J. O., Vezhapparambu, S., Mauder, M., McFadden, J. P., & Gallo, K.** (2009). Surface Energy Balance Measurements Above an Exurban Residential Neighbourhood of Kansas City, Missouri. *Boundary-Layer Meteorology*, 133(3), 299–321. doi:10.1007/s10546-009-9421-3.
- Barba, F. J., Gavahian, M., Es, I., Zhu, Z., Chemat, F., Lorenzo, J. M., & Mousavi Khaneghah, A.** (2019). Solar radiation as a prospective energy source for green and economic processes in the food industry: From waste biomass valorization to dehydration, cooking, and baking. *Journal of Cleaner Production*, 220, 1121–1130. doi:10.1016/j.jclepro.2019.02.175.

- Başakın, E. E.** (2022). Comments on “Groundwater quality modeling using a novel hybrid data-intelligence model based on gray wolf optimization algorithm and multi-layer perceptron artificial neural network: a case study in Asadabad Plain, Hamedan, Iran” Cheraghi, Mehrdad et al. *Environmental Science and Pollution Research*, 29(27), 41869–41871. doi:10.1007/s11356-022-19846-z.
- Başakın, E. E., Ekmekcioğlu, Ö., & Özger, M.** (2021). Drought prediction using hybrid soft-computing methods for semi-arid region. *Modeling Earth Systems and Environment*, 7(4), 2363–2371. doi:10.1007/s40808-020-01010-6.
- Başakın, E. E., Ekmekcioğlu, Ö., & Özger, M.** (2023). Developing a novel approach for missing data imputation of solar radiation: A hybrid differential evolution algorithm based eXtreme gradient boosting model. *Energy Conversion and Management*, 280, 116780. doi:10.1016/j.enconman.2023.116780.
- Başakın, E. E., Stoy, P. C., Demirel, M. C., & Pham, Q. B.** (2024). Spatiotemporal Variability of Gross Primary Productivity in Türkiye: A Multi-Source and Multi-Method Assessment. *Remote Sensing*, 16(11), 1994. doi:10.3390/rs16111994.
- Basheer Shukur, O., & Hisyam Lee, M.** (2015). Imputation of Missing Values in Daily Wind Speed Data Using Hybrid AR-ANN Method. *Modern Applied Science*, 9(11), 1. doi:10.5539/mas.v9n11p1.
- Bashir, F., & Hua-Liang Wei.** (2015). Parametric and non-parametric methods to enhance prediction performance in the presence of missing data. *2015 19th International Conference on System Theory, Control and Computing (ICSTCC)*, 337–342. doi:10.1109/ICSTCC.2015.7321316.
- Bayram, S., & Çitakoğlu, H.** (2023). Modeling monthly reference evapotranspiration process in Turkey: application of machine learning methods. *Environmental Monitoring and Assessment*, 195(1), 67. doi:10.1007/s10661-022-10662-z.
- Bazzaz Abkenar, S., Mahdipour, E., Jameii, S. M., & Haghi Kashani, M.** (2021). A hybrid classification method for Twitter spam detection based on differential evolution and random forest. *Concurrency and Computation: Practice and Experience*, 33(21). doi:10.1002/cpe.6381.
- Beaulieu-Jones, B. K., Lavage, D. R., Snyder, J. W., Moore, J. H., Pendergrass, S. A., & Bauer, C. R.** (2018). Characterizing and Managing Missing Structured Data in Electronic Health Records: Data Analysis. *JMIR Medical Informatics*, 6(1), e11. doi:10.2196/medinform.8960.
- Beguiría, S., Vicente-Serrano, S. M., Reig, F., & Latorre, B.** (2014). Standardized precipitation evapotranspiration index (SPEI) revisited: parameter fitting, evapotranspiration models, tools, datasets and drought monitoring. *International Journal of Climatology*, 34(10), 3001–3023. doi:10.1002/joc.3887.
- Ben Aissia, M. A., Chebana, F., & Ouarda, T. B. M. J.** (2017). Multivariate missing data in hydrology – Review and applications. *Advances in Water Resources*, 110, 299–309. doi:10.1016/j.advwatres.2017.10.002.

- Bendavid, I., Statlender, L., Shvartser, L., Teppler, S., Azullay, R., Sapir, R., & Singer, P.** (2022). A novel machine learning model to predict respiratory failure and invasive mechanical ventilation in critically ill patients suffering from COVID-19. *Scientific Reports*, 12(1), 10573. doi:10.1038/s41598-022-14758-x.
- Berhanu, K. G., Lohani, T. K., & Hatiye, S. D.** (2024). Long-term spatiotemporal dynamics of groundwater storage in the data-scarce region: Tana sub-basin, Ethiopia. *Heliyon*, 10(3), e24474. doi:https://doi.org/10.1016/j.heliyon.2024.e24474.
- Bernacchi, C. J., & VanLoocke, A.** (2015). Terrestrial Ecosystems in a Changing Environment: A Dominant Role for Water. *Annual Review of Plant Biology*, 66(1), 599–622. doi:10.1146/annurev-arplant-043014-114834.
- Bhaga, T. D., Dube, T., Shekede, M. D., & Shoko, C.** (2020). Impacts of Climate Variability and Drought on Surface Water Resources in Sub-Saharan Africa Using Remote Sensing: A Review. *Remote Sensing*, 12(24), 4184. doi:10.3390/rs12244184.
- Bi, W., He, W., Zhou, Y., Ju, W., Liu, Y., Liu, Y., ... Cheng, N.** (2022). A global 0.05° dataset for gross primary production of sunlit and shaded vegetation canopies from 1992 to 2020. *Scientific Data*, 9(1), 213. doi:10.1038/s41597-022-01309-2
- Bi, W., & Zhou, Y.** (2022). A global 0.05° dataset for gross primary production of sunlit and shaded vegetation canopies (1992–2020). [Dataset]. doi:https://doi.org/10.5061/dryad.dfn2z352k.
- Bilgili, B. C., Erşahin, S., Kavakligil, S. S., & Öner, N.** (2020). Net Primary Productivity Of A Mountain Forest Ecosystem As Affected By Climate And Topography. *CERNE*, 26(3), 356–368. doi:10.1590/01047760202026032730.
- Biltekin, D., Eriş, K. K., Çağatay, M. N., Henry, P., & Yakupoğlu, N.** (2023). New records of vegetation and climate changes in the Sea of Marmara during the Marine Isotope Stages 3, 4 and 5 (a-c). *Quaternary International*, 667, 1–18. doi:10.1016/j.quaint.2023.06.005.
- Bonan, G.** (2016). *Ecological Climatology: Concepts and Applications*. Cambridge University Press. https://books.google.com/books?id=AcvZCgAAQBAJ.
- Bounoua, Z., Ouazzani Chahidi, L., & Mechaqrane, A.** (2021). Estimation of daily global solar radiation using empirical and machine-learning methods: A case study of five Moroccan locations. *Sustainable Materials and Technologies*, 28, e00261. doi:10.1016/j.susmat.2021.e00261.
- Bowen, I. S.** (1926). The Ratio of Heat Losses by Conduction and by Evaporation from any Water Surface. *Physical Review*, 27(6), 779–787. doi:10.1103/PhysRev.27.779.

- Bozali, N.** (2020). Assessment of the soil protection function of forest ecosystems using GIS-based Multi-Criteria Decision Analysis: A case study in Adıyaman, Turkey. *Global Ecology and Conservation*, 24, e01271. doi:10.1016/j.gecco.2020.e01271.
- Breiman, L.** (2001). Random Forests. *Machine Learning*, 45(1), 5–32. doi:10.1023/A:1010933404324.
- Burakowski, E., Tawfik, A., Ouimette, A., Lepine, L., Novick, K., Ollinger, S., ... Bonan, G.** (2018). The role of surface roughness, albedo, and Bowen ratio on ecosystem energy balance in the Eastern United States. *Agricultural and Forest Meteorology*, 249, 367–376. doi:10.1016/j.agrformet.2017.11.030.
- Burba, G., Madsen, R., & Feese, K.** (2013). Eddy covariance method for CO₂ emission measurements in CCUS applications: Principles, instrumentation and software. *Energy Procedia*, 40, 329–336. doi:10.1016/j.egypro.2013.08.038.
- Bürger, G.** (2023). Trends? Complicated answers to a simple question. *Hydrological Sciences Journal*, 68(12), 1680–1692. doi:10.1080/02626667.2023.2224922.
- Bushenkova, A., Soares, P. M. M., Johannsen, F., & Lima, D. C. A.** (2024). Towards an improved representation of the urban heat island effect : A multi-scale application of XGBoost for madrid. *Urban Climate*, 55, 101982. doi:10.1016/j.uclim.2024.101982.
- Byun, H. R., & Wilhite, D. A.** (1999). Objective Quantification of Drought Severity and Duration. *Journal of Climate*, 12(9), 2747–2756. doi:10.1175/1520-0442(1999)012<2747:OQODSA>2.0.CO;2.
- Cai, S., Zuo, D., Wang, H., Xu, Z., Wang, G., & Yang, H.** (2023). Assessment of agricultural drought based on multi-source remote sensing data in a major grain producing area of Northwest China. *Agricultural Water Management*, 278, 108142. doi:10.1016/j.agwat.2023.108142.
- Camastra, F., Capone, V., Ciaramella, A., Riccio, A., & Staiano, A.** (2022). Prediction of environmental missing data time series by Support Vector Machine Regression and Correlation Dimension estimation. *Environmental Modelling & Software*, 150, 105343. doi:10.1016/j.envsoft.2022.105343.
- Campos-Taberner, M., Gilabert, M. A., Sánchez-Ruiz, S., Martínez, B., Jiménez-Guisado, A., García-Haro, F. J., & Guanter, L.** (2024). Global Carbon Fluxes Using Multioutput Gaussian Processes Regression and MODIS Products. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17, 11310–11321. doi:10.1109/JSTARS.2024.3413184.
- Camps-Valls, G., Campos-Taberner, M., Moreno-Martínez, Á., Walther, S., Duveiller, G., Cescatti, A., ... Running, S. W.** (2021). A unified vegetation index for quantifying the terrestrial biosphere. *Science Advances*, 7(9), eabc7447. doi:10.1126/sciadv.abc7447.

- Canan, F., Golasi, I., Falasca, S., & Salata, F.** (2020). Outdoor thermal perception and comfort conditions in the Köppen-Geiger climate category BSk. One-year field survey and measurement campaign in Konya, Turkey. *Science of The Total Environment*, 738, 140295. doi:10.1016/j.scitotenv.2020.140295.
- Cao, Q., Liu, Y., Sun, X., & Yang, L.** (2022). Country-level evaluation of solar radiation data sets using ground measurements in China. *Energy*, 241, 122938. doi:10.1016/j.energy.2021.122938.
- Cavus, Y., Stahl, K., & Aksoy, H.** (2023). Drought intensity–duration–frequency curves based on deficit in precipitation and streamflow for water resources management. *Hydrology and Earth System Sciences*, 27(18), 3427–3445. doi:10.5194/hess-27-3427-2023.
- Chang, W., Li, W., Ma, H., Wang, D., Bandala, E. R., Yu, Y., & Rodrigo-Comino, J.** (2022). An integrated approach for shaping drought characteristics at the watershed scale. *Journal of Hydrology*, 604, 127248. doi:10.1016/j.jhydrol.2021.127248.
- Chen, T., & Guestrin, C.** (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 13-17-Aug, 785–794. doi:10.1145/2939672.2939785.
- Chen, T., Wang, Y., & Peng, L.** (2023). Advanced time-lagged effects of drought on global vegetation growth and its social risk in the 21st century. *Journal of Environmental Management*, 347, 119253. doi:10.1016/j.jenvman.2023.119253.
- Chen, X., Wang, B., Yu, M., Jin, J., & Xu, W.** (2016). The interpolation of missing wind speed data based on optimized LSSVM model. *2016 IEEE 8th International Power Electronics and Motion Control Conference, IPEMC-ECCE Asia 2016*, 1, 1448–1451. doi:10.1109/IPEMC.2016.7512504.
- Chen, Y., Gu, H., Wang, M., Gu, Q., Ding, Z., Ma, M., ... Tang, X.** (2019). Contrasting Performance of the Remotely-Derived GPP Products over Different Climate Zones across China. *Remote Sensing*, 11(16), 1855. doi:10.3390/rs11161855.
- Chen, Y., Huang, J., Song, X., Gao, P., Wan, S., Shi, L., & Wang, X.** (2018). Spatiotemporal Characteristics of Winter Wheat Waterlogging in the Middle and Lower Reaches of the Yangtze River, China. *Advances in Meteorology*, 2018, 1–11. doi:10.1155/2018/3542103.
- Chen, Y. Y., Chu, C. R., & Li, M. H.** (2012). A gap-filling model for eddy covariance latent heat flux: Estimating evapotranspiration of a subtropical seasonal evergreen broad-leaved forest as an example. *Journal of Hydrology*, 468–469, 101–110. doi:10.1016/j.jhydrol.2012.08.026.
- Cheng, R.** (2024). Solar-Induced Chlorophyll Fluorescence (SIF): Towards a Better Understanding of Vegetation Dynamics and Carbon Uptake in Arctic-Boreal Ecosystems. *Current Climate Change Reports*, 10(2), 13–32. doi:10.1007/s40641-024-00194-8.

- Cheval, S., Dumitrescu, A., & Amihaesei, V. A.** (2020). Exploratory Analysis of Urban Climate Using a Gap-Filled Landsat 8 Land Surface Temperature Data Set. *Sensors*, 20(18), 5336. doi:10.3390/s20185336.
- Costa, R. L., Barros Gomes, H., Cavalcante Pinto, D. D., da Rocha Júnior, R. L., dos Santos Silva, F. D., Barros Gomes, H., ... Luís Herdies, D.** (2021). Gap Filling and Quality Control Applied to Meteorological Variables Measured in the Northeast Region of Brazil. *Atmosphere*, 12(10), 1278. doi:10.3390/atmos12101278.
- Coulibaly, P., & Evora, N. D.** (2007). Comparison of neural network methods for infilling missing daily weather records. *Journal of Hydrology*, 341(1–2), 27–41. doi:10.1016/j.jhydrol.2007.04.020.
- Dai, H., Huang, G., Zeng, H., & Zhou, F.** (2022). PM2.5 volatility prediction by XGBoost-MLP based on GARCH models. *Journal of Cleaner Production*, 356, 131898. doi:10.1016/j.jclepro.2022.131898.
- Danodia, A., Kushwaha, A., & Patel, N. R.** (2021). Remote sensing-derived combined index for agricultural drought assessment of rabi pulse crops in Bundelkhand region, India. *Environment, Development and Sustainability*, 23(10), 15432–15449. doi:10.1007/s10668-021-01305-3.
- de Oliveira, R. C. G., Cunha, C. L., Tôrres, A. R., & Corrêa, S. M.** (2021). Forecasts of tropospheric ozone in the Metropolitan Area of Rio de Janeiro based on missing data imputation and multivariate calibration techniques. *Environmental Monitoring and Assessment*, 193(8). doi:10.1007/s10661-021-09333-2.
- Deb Burman, P. K., Sarma, D., Morrison, R., Karipot, A., & Chakraborty, S.** (2019). Seasonal variation of evapotranspiration and its effect on the surface energy budget closure at a tropical forest over north-east India. *Journal of Earth System Science*, 128(5), 127. doi:10.1007/s12040-019-1158-x.
- Dechant, B., Ryu, Y., Badgley, G., Köhler, P., Rascher, U., Migliavacca, M., ... Berry, J. A.** (2022). NIRVP: A robust structural proxy for sun-induced chlorophyll fluorescence and photosynthesis across scales. *Remote Sensing of Environment*, 268, 112763. doi:10.1016/j.rse.2021.112763.
- Delbart, N., Kergoat, L., Le Toan, T., Lhermitte, J., & Picard, G.** (2005). Determination of phenological dates in boreal regions using normalized difference water index. *Remote Sensing of Environment*, 97(1), 26–38. doi:10.1016/j.rse.2005.03.011.
- Deligöz, A., Bayar, E., Karatepe, Y., & Genç, M.** (2019). Photosynthetic capacity, nutrient and water status following precommercial thinning in Anatolian black pine. *Forest Ecology and Management*, 451, 117533. doi:10.1016/j.foreco.2019.117533.
- Demir, V., & Citakoglu, H.** (2023). Forecasting of solar radiation using different machine learning approaches. *Neural Computing and Applications*, 35(1), 887–906. doi:10.1007/s00521-022-07841-x.

- Demircan, M., Gürkan, H., Eskioğlu, O., Arabacı, H., & Coşkun, M.** (2017). Climate Change Projections for Turkey: Three Models and Two Scenarios. *Turkish Journal of Water Science and Management*, 1(1), 22–43. doi:10.31807/tjwsm.297183.
- Demirci, U., & Demirel, M. C.** (2023). Effect of Dynamic PET Scaling with LAI and Aspect on the Spatial Performance of a Distributed Hydrologic Model. *Agronomy*, 13(2), 534. doi:10.3390/agronomy13020534.
- Demirhan, H., & Renwick, Z.** (2018). Missing value imputation for short to mid-term horizontal solar irradiance data. *Applied Energy*, 225, 998–1012. doi:10.1016/j.apenergy.2018.05.054.
- Denhard, A., Bandyopadhyay, S., Habte, A., Sengupta, M., Denhard, A., Bandyopadhyay, S., & Habte, A.** (2022). A Comparison of Time-Series Gap-Filling Methods to Impute Solar Radiation Data Preprint A Comparison of Time-Series Gap-Filling Methods to Impute Solar Radiation doi:10.18086/swc.2021.38.03.
- Desai, A. R., Richardson, A. D., Moffat, A. M., Kattge, J., Hollinger, D. Y., Barr, A., Falge, E., ... Stauch, V. J.** (2008). Cross-site evaluation of eddy covariance GPP and RE decomposition techniques. *Agricultural and Forest Meteorology*, 148(6), 821–838. doi:https://doi.org/10.1016/j.agrformet.2007.11.012.
- Dias Júnior, D. A., da Cruz, L. B., Bandeira Diniz, J. O., França da Silva, G. L., Junior, G. B., Silva, A. C., ... Gattass, M.** (2021). Automatic method for classifying COVID-19 patients based on chest X-ray images, using deep features and PSO-optimized XGBoost. *Expert Systems with Applications*, 183. doi:10.1016/j.eswa.2021.115452.
- Dimas, F., Gilani, S., & Aris, M.** (2011). Hourly solar radiation estimation from limited meteorological data to complete missing solar radiation data. *International Conference on Environment Science and Engineering IPCBEE*, 8, 14–18. <http://www.ipcbee.com/vol8/4-S017.pdf>.
- Dong, Y., & Hansen, H.** (2023). Development and design of an affordable field scale weighing lysimeter using a microcontroller system. *Smart Agricultural Technology*, 4, 100147. doi:10.1016/j.atech.2022.100147.
- Du, J., Sun, H., Cao, Y., Liu, Y., Pan, L., & Liu, Y.** (2019). Ensemble interpolation of missing wind turbine nacelle wind speed data in wind farms based on robust particle swarm optimized generalized regression neural network. *International Journal of Green Energy*, 16(14), 1210–1219. doi:10.1080/15435075.2019.1671396.
- Du, Y., Zhu, K., & Zhang, Z.** (2024). CSIF and GOSIF Do Not Accurately Capture the Vegetation Greening During the Spring of 2020. *IEEE Geoscience and Remote Sensing Letters*, 21, 1–5. doi:10.1109/LGRS.2024.3379255.
- Dumitrescu, A., Brabec, M., & Cheval, S.** (2020). Statistical Gap-Filling of SEVIRI Land Surface Temperature. *Remote Sensing*, 12(9), 1423. doi:10.3390/rs12091423.

- Eamus, D., Cleverly, J., Boulain, N., Grant, N., Faux, R., & Villalobos-Vega, R.** (2013). Carbon and water fluxes in an arid-zone acacia savanna woodland: An analyses of seasonal patterns and responses to rainfall events. *Agricultural and Forest Meteorology*, 182–183, 225–238. doi:10.1016/j.agrformet.2013.04.020.
- Ekercin, S., Sertel, E., Dadaser-Celik, F., & Durduran, S.** (2013). *Investigating the Climate Change Impacts on the Water Resources of the Konya Closed Basin Area (Turkey) Using Satellite Remote Sensing Data BT - Causes, Impacts and Solutions to Global Warming* (I. Dincer, C. O. Colpan & F. Kadioglu (eds.); pp. 157–168). Springer New York. doi:10.1007/978-1-4614-7588-0_9.
- Ekmekcioğlu, Ö., Koc, K., Özger, M., & Işık, Z.** (2022). Exploring the additional value of class imbalance distributions on interpretable flash flood susceptibility prediction in the Black Warrior River basin, Alabama, United States. *Journal of Hydrology*, 610, 127877. doi:10.1016/j.jhydrol.2022.127877.
- Emamgholizadeh, S., Bazoobandi, A., Mohammadi, B., Ghorbani, H., & Amel Sadeghi, M.** (2023). Prediction of soil cation exchange capacity using enhanced machine learning approaches in the southern region of the Caspian Sea. *Ain Shams Engineering Journal*, 14(2), 101876. doi:10.1016/j.asej.2022.101876.
- Ergün, E., & Demirel, M. C.** (2023). On the use of distributed hydrologic model for filling large gaps at different parts of the streamflow data. *Engineering Science and Technology, an International Journal*, 37, 101321. doi:10.1016/j.jestch.2022.101321.
- Erşahin, S., Bilgili, B. C., Dikmen, Ü., & Ercanli, İ.** (2016). Net Primary Productivity of Anatolian Forests in Relation to Climate, 2000-2010. *Forest Science*, 62(6), 698–709. doi:10.5849/forsci.15-171.
- Eum, Y., & Yoo, E. H.** (2022). Imputation of missing time-activity data with long-term gaps: A multi-scale residual CNN-LSTM network model. *Computers, Environment and Urban Systems*, 95, 101823. doi:10.1016/j.compenvurbsys.2022.101823.
- Evrendilek, F., Karakaya, N., Aslan, G., & Ertekin, C.** (2011). Using Eddy Covariance Sensors to Quantify Carbon Metabolism of Peatlands: A Case Study in Turkey. *Sensors*, 11(1), 522–538. doi:10.3390/s110100522.
- Falge, E., Tenhunen, J., Baldocchi, D., Aubinet, M., Bakwin, P., Berbigier, P., ... Wofsy, S.** (2002). Phase and amplitude of ecosystem carbon release and uptake potentials as derived from FLUXNET measurements. *Agricultural and Forest Meteorology*, 113(1–4), 75–95. doi:10.1016/S0168-1923(02)00103-X
- Feng, D. C., Wang, W. J., Mangalathu, S., & Taciroglu, E.** (2021). Interpretable XGBoost-SHAP Machine-Learning Model for Shear Strength Prediction of Squat RC Walls. *Journal of Structural Engineering*, 147(11), 1–13. doi:10.1061/(asce)st.1943-541x.0003115.

- Fernani, A.** (2019). Mapping futures studies scholarship from 1968 to present: A bibliometric review of thematic clusters, research trends, and research gaps. *Futures*, 105, 104–123. doi:10.1016/j.futures.2018.09.007.
- Ferreira, L. B.** (2022). *Exploring machine learning and multi-task learning to estimate meteorological data and reference evapotranspiration across Brazil*. 259(May 2021), 0–1. doi:10.1016/j.agwat.2021.107281.
- Ferreira, L. B., da Cunha, F. F., de Oliveira, R. A., & Fernandes Filho, E. I.** (2019). Estimation of reference evapotranspiration in Brazil with limited meteorological data using ANN and SVM – A new approach. *Journal of Hydrology*, 572, 556–570. doi:10.1016/j.jhydrol.2019.03.028.
- Fkiri, S., Rzigui, T., Ghazghazi, H., Khouja, L. M., Khaldi, A., Guibal, F., & Nasr, Z.** (2023). Ecotype effects on photosynthesis performance using A/PFFD among *Pinus nigra* Arn. *Notulae Botanicae Horti Agrobotanici Cluj-Napoca*, 51(3), 12599. doi:10.15835/nbha51312599.
- Foltýnová, L., Fischer, M., & McGloin, R. P.** (2020). Recommendations for gap-filling eddy covariance latent heat flux measurements using marginal distribution sampling. *Theoretical and Applied Climatology*, 139(1–2), 677–688. doi:10.1007/s00704-019-02975-w.
- Forkuor, G., Hounkpatin, O. K. L., Welp, G., & Thiel, M.** (2017). High Resolution Mapping of Soil Properties Using Remote Sensing Variables in South-Western Burkina Faso: A Comparison of Machine Learning and Multiple Linear Regression Models. *PLOS ONE*, 12(1), e0170478. doi:10.1371/journal.pone.0170478.
- Friedlingstein, P., O’Sullivan, M., Jones, M. W., Andrew, R. M., Bakker, D. C. E., Hauck, J., ... Zheng, B.** (2023). Global Carbon Budget 2023. *Earth System Science Data*, 15(12), 5301–5369. doi:10.5194/essd-15-5301-2023.
- Friedman, J. H.** (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5). doi:10.1214/aos/1013203451.
- Fu, Z., Ciais, P., Prentice, I. C., Gentile, P., Makowski, D., Bastos, A., ... Hajima, T.** (2022). Atmospheric dryness reduces photosynthesis along a large range of soil water deficits. *Nature Communications*, 13(1), 989. doi:10.1038/s41467-022-28652-7.
- Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., ... Michaelsen, J.** (2015). The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. *Scientific Data*, 2(1), 150066. doi:10.1038/sdata.2015.66.
- Gan, R., Zhang, Y., Shi, H., Yang, Y., Eamus, D., Cheng, L., ... Yu, Q.** (2018). Use of satellite leaf area index estimating evapotranspiration and gross assimilation for Australian ecosystems. *Ecohydrology*, 11(5). doi:10.1002/eco.1974.

- Gao, B. cai.** (1996). NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space. *Remote Sensing of Environment*, 58(3), 257–266. doi:10.1016/S0034-4257(96)00067-3.
- Gao, H., Tian, H., Zhang, Z., & Xia, X.** (2022). Warming-induced greenhouse gas fluxes from global croplands modified by agricultural practices: A meta-analysis. *Science of The Total Environment*, 820, 153288. doi:10.1016/j.scitotenv.2022.153288.
- Gao, Y., Merz, C., Lischeid, G., & Schneider, M.** (2018). A review on missing hydrological data processing. *Environmental Earth Sciences*, 77(2), 47. doi:10.1007/s12665-018-7228-6.
- García Acevedo, F., Rojas Serrano, J., Vásquez Vega, A., Parra Peñaranda, D., & Castro Becerra, E.** (2016). Estimating missing data in historic series of global radiation through neural network algorithms. *Sistemas y Telemática*, 14(37), 9–22. doi:10.18046/syt.v14i37.2239.
- Gautam, C., & Ravi, V.** (2015). Data imputation via evolutionary computation, clustering and a neural network. *Neurocomputing*, 156, 134–142. doi:10.1016/j.neucom.2014.12.073.
- Gavrilov, M. B., Radaković, M. G., Sipos, G., Mezősi, G., Gavrilov, G., Lukić, T., ... Marković, S. B.** (2020). Aridity in the Central and Southern Pannonian Basin. *Atmosphere*, 11(12), 1269. doi:10.3390/atmos11121269.
- Gebremichael, M., & Barros, A.** (2006). Evaluation of MODIS Gross Primary Productivity (GPP) in tropical monsoon regions. *Remote Sensing of Environment*, 100(2), 150–166. doi:10.1016/j.rse.2005.10.009.
- Goodrich, J. P., Wall, A. M., Campbell, D. I., Fletcher, D., Wecking, A. R., & Schipper, L. A.** (2021). Improved gap filling approach and uncertainty estimation for eddy covariance N2O fluxes. *Agricultural and Forest Meteorology*, 297, 108280. doi:10.1016/j.agrformet.2020.108280.
- Gravesteijn, B. Y., Sewalt, C. A., Venema, E., Nieboer, D., Steyerberg, E. W., Åkerlund, C., ... Zoerle, T.** (2021). Missing Data in Prediction Research: A Five-Step Approach for Multiple Imputation, Illustrated in the CENTER-TBI Study. *Journal of Neurotrauma*, 38(13), 1842–1857. doi:10.1089/neu.2020.7218.
- Gu, Q., Zheng, H., Yao, L., Wang, M., Ma, M., Wang, X., & Tang, X.** (2020). Performance of the Remotely-Derived Products in Monitoring Gross Primary Production across Arid and Semi-Arid Ecosystems in Northwest China. *Land*, 9(9), 288. doi:10.3390/land9090288.
- Gu, Y., Zhang, D., & Bao, Z.** (2021). A new data-driven predictor, PSO-XGBoost, used for permeability of tight sandstone reservoirs: A case study of member of chang 4+5, western Jiyuan Oilfield, Ordos Basin. *Journal of Petroleum Science and Engineering*, 199, 108350. doi:10.1016/j.petrol.2021.108350.
- Güçlü, Y. S.** (2020). Improved visualization for trend analysis by comparing with classical Mann-Kendall test and ITA. *Journal of Hydrology*, 584, 124674. doi:10.1016/j.jhydrol.2020.124674.

- Gulbeyaz, O., Bond-Lamberty, B., Akyurek, Z., & West, T. O.** (2018). A new approach to evaluate the MODIS annual NPP product (MOD17A3) using forest field data from Turkey. *International Journal of Remote Sensing*, 39(8), 2560–2578. doi:10.1080/01431161.2018.1430913.
- Gunda, T., Hornberger, G. M., & Gilligan, J. M.** (2016). Spatiotemporal Patterns of Agricultural Drought in Sri Lanka: 1881–2010. *International Journal of Climatology*, 36(2), 563–575. doi:10.1002/joc.4365.
- Guo, B., Chen, Z., Guo, J., Liu, F., Chen, C., & Liu, K.** (2016). Analysis of the Nonlinear Trends and Non-Stationary Oscillations of Regional Precipitation in Xinjiang, Northwestern China, Using Ensemble Empirical Mode Decomposition. *International Journal of Environmental Research and Public Health*, 13(3), 345. doi:10.3390/ijerph13030345.
- Guo, J., Shi, K., Liu, X., Sun, Y., Li, W., & Kong, Q.** (2019). Singular spectrum analysis of ionospheric anomalies preceding great earthquakes: Case studies of Kaikoura and Fukushima earthquakes. *Journal of Geodynamics*, 124, 1–13. doi:10.1016/j.jog.2019.01.005.
- Guo, X., Zhang, H., Kang, L., Du, J., Li, W., & Zhu, Y.** (2007). Quality control and flux gap filling strategy for Bowen ratio method: Revisiting the Priestley-Taylor evaporation model. *Environmental Fluid Mechanics*, 7(5), 421–437. doi:10.1007/s10652-007-9033-8.
- Gupta, A., Gowda, S., Tiwari, A., & Gupta, A. K.** (2024). XGBoost-SHAP framework for asphalt pavement condition evaluation. *Construction and Building Materials*, 426, 136182. doi:10.1016/j.conbuildmat.2024.136182.
- Gupta, S., Deb Burman, P. K., Tiwari, Y. K., Dumka, U. C., Kumari, N., Srivastava, A., & Raghubanshi, A. S.** (2023). Understanding carbon sequestration trends using model and satellite data under different ecosystems in India. *Science of The Total Environment*, 897, 166381. doi:10.1016/j.scitotenv.2023.166381.
- Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F.** (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, 377(1–2), 80–91. doi:10.1016/j.jhydrol.2009.08.003.
- Haliduola, H. N., Bretz, F., & Mansmann, U.** (2022). Missing data imputation using utility-based regression and sampling approaches. *Computer Methods and Programs in Biomedicine*, 226, 107172. doi:10.1016/j.cmpb.2022.107172.
- Hamed, K. H.** (2008). Trend detection in hydrologic data: The Mann–Kendall trend test under the scaling hypothesis. *Journal of Hydrology*, 349(3–4), 350–363. doi:10.1016/j.jhydrol.2007.11.009.
- Han, J., & Singh, V. P.** (2023). A review of widely used drought indices and the challenges of drought assessment under climate change. *Environmental Monitoring and Assessment*, 195(12), 1438. doi:10.1007/s10661-023-12062-3.

- Hanadé Houmma, I., El Mansouri, L., Hadria, R., Emran, A., & Chehbouni, A.** (2022). Retrospective analysis and version improvement of the satellite-based drought composite index. A semi-arid Tensift-Morocco application. *Geocarto International*, 37(11), 3069–3090. doi:10.1080/10106049.2020.1844314.
- Hanadé Houmma, I., Gadal, S., El Mansouri, L., Garba, M., Gbetkom, P. G., Mamane Barkawi, M. B., & Hadria, R.** (2023). A new multivariate agricultural drought composite index based on random forest algorithm and remote sensing data developed for Sahelian agrosystems. *Geomatics, Natural Hazards and Risk*, 14(1). doi:10.1080/19475705.2023.2223384.
- Hao, C., Zhang, J., & Yao, F.** (2015). Combination of multi-sensor remote sensing data for drought monitoring over Southwest China. *International Journal of Applied Earth Observation and Geoinformation*, 35, 270–283. doi:10.1016/j.jag.2014.09.011.
- Hasan, H. H., Mohd Razali, S. F., Muhammad, N. S., & Ahmad, A.** (2019). Research Trends of Hydrological Drought: A Systematic Review. *Water*, 11(11), 2252. doi:10.3390/w11112252.
- Hasan, M. K., Alam, M. A., Roy, S., Dutta, A., Jawad, M. T., & Das, S.** (2021). Missing value imputation affects the performance of machine learning: A review and analysis of the literature (2010–2021). *Informatics in Medicine Unlocked*, 27, 100799. doi:10.1016/j.imu.2021.100799.
- Hasanpour Kashani, M., & Dinpashoh, Y.** (2012). Evaluation of efficiency of different estimation methods for missing climatological data. *Stochastic Environmental Research and Risk Assessment*, 26(1), 59–71. doi:10.1007/s00477-011-0536-y.
- Hayati Rezvan, P., Lee, K. J., & Simpson, J. A.** (2015). The rise of multiple imputation: a review of the reporting and implementation of the method in medical research. *BMC Medical Research Methodology*, 15(1), 30. doi:10.1186/s12874-015-0022-1.
- He, C., Liu, J., Xu, F., Zhang, T., Chen, S., Sun, Z., ... He, J.** (2020). Improving solar radiation estimation in China based on regional optimal combination of meteorological factors with machine learning methods. *Energy Conversion and Management*, 220, 113111. doi:10.1016/j.enconman.2020.113111.
- He, H., Cao, Y., Cao, Y., & Wen, J.** (2013). Ensemble learning for wind profile prediction with missing values. *Neural Computing and Applications*, 22(2), 287–294. doi:10.1007/s00521-011-0708-1.
- He, M., Ju, W., Zhou, Y., Chen, J., He, H., Wang, S., ... Zhao, F.** (2013). Development of a two-leaf light use efficiency model for improving the calculation of terrestrial gross primary productivity. *Agricultural and Forest Meteorology*, 173, 28–39. doi:10.1016/j.agrformet.2013.01.003.
- He, P., Ma, X., Meng, X., Han, Z., Liu, H., & Sun, Z.** (2022). Spatiotemporal evolutionary and mechanism analysis of grassland GPP in China. *Ecological Indicators*, 143, 109323. doi:10.1016/j.ecolind.2022.109323.

- He, S., Zhang, Y., Ma, N., Tian, J., Kong, D., & Liu, C.** (2022). A daily and 500m coupled evapotranspiration and gross primary production product across China during 2000-2020. *Earth System Science Data*, 14(12), 5463–5488. doi:10.5194/essd-14-5463-2022.
- He, Z., Yang, Y., Fang, R., Zhou, S., Zhao, W., Bai, Y., ... Wang, B.** (2023). Integration of shapley additive explanations with random forest model for quantitative precipitation estimation of mesoscale convective systems. *Frontiers in Environmental Science*, 10. doi:10.3389/fenvs.2022.1057081.
- Hilker, T., Coops, N. C., Wulder, M. A., Black, T. A., & Guy, R. D.** (2008). The use of remote sensing in light use efficiency based models of gross primary production: A review of current status and future requirements. *Science of The Total Environment*, 404(2–3), 411–423. doi:10.1016/j.scitotenv.2007.11.007.
- Hobbins, M. T., Wood, A., McEvoy, D. J., Huntington, J. L., Morton, C., Anderson, M., & Hain, C.** (2016). The Evaporative Demand Drought Index. Part I: Linking Drought Evolution to Variations in Evaporative Demand. *Journal of Hydrometeorology*, 17(6), 1745–1761. doi:10.1175/JHM-D-15-0121.1.
- Hocaoglu, F. O., Oysal, Y., & Kurban, M.** (2009). Missing wind data forecasting with adaptive neuro-fuzzy inference system. *Neural Computing and Applications*, 18(3), 207–212. doi:10.1007/s00521-008-0172-8.
- Hoyos-Gómez, L. S., Ruiz-Muñoz, J. F., & Ruiz-Mendoza, B. J.** (2022). Short-term forecasting of global solar irradiance in tropical environments with incomplete data. *Applied Energy*, 307, 118192. doi:10.1016/j.apenergy.2021.118192.
- Hron, K., Templ, M., & Filzmoser, P.** (2010). Imputation of missing values for compositional data using classical and robust methods. *Computational Statistics & Data Analysis*, 54(12), 3095–3107. doi:10.1016/j.csda.2009.11.023.
- Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q., ... Liu, H. H.** (1998). The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 454(1971), 903–995. doi:10.1098/rspa.1998.0193..
- Huang, X., Xiao, J., Wang, X., & Ma, M.** (2021). Improving the global MODIS GPP model by optimizing parameters with FLUXNET data. *Agricultural and Forest Meteorology*, 300, 108314. doi:10.1016/j.agrformet.2020.108314.
- Huang, Y., Hendricks Franssen, H. J., Herbst, M., Hirschi, M., Michel, D., Seneviratne, S. I., ... & Vereecken, H.** (2020). Evaluation of different methods for gap filling of long-term actual evapotranspiration time series measured by lysimeters. *Vadose Zone Journal*, 19(1), 1–15. doi:10.1002/vzj2.20020.

- Huete, A., Didan, K., Miura, T., Rodriguez, E. ., Gao, X., & Ferreira, L. .** (2002). Overview of the radiometric and biophysical performance of the MODIS vegetation indices. *Remote Sensing of Environment*, 83(1–2), 195–213. doi:10.1016/S0034-4257(02)00096-2.
- Hui, D., Wan, S., Su, B., Katul, G., Monson, R., & Luo, Y.** (2004). Gap-filling missing data in eddy covariance measurements using multiple imputation (MI) for annual estimations. *Agricultural and Forest Meteorology*, 121(1–2), 93–111. doi:10.1016/S0168-1923(03)00158-8.
- Hutley, L. B., Beringer, J., Fatichi, S., Schymanski, S. J., & Northwood, M.** (2022). Gross primary productivity and water use efficiency are increasing in a high rainfall tropical savanna. *Global Change Biology*, 28(7), 2360–2380. doi:10.1111/gcb.16012.
- Iodice D’Enza, A., Markos, A., & Palumbo, F.** (2022). Chunk-wise regularised PCA-based imputation of missing data. *Statistical Methods & Applications*, 31(2), 365–386. doi:10.1007/s10260-021-00575-5.
- Jamshidian, M., Jalal, S., & Jansen, C.** (2014). MissMech: An R Package for Testing Homoscedasticity, Multivariate Normality, and Missing Completely at Random (MCAR). *Journal of Statistical Software*, 56(6). doi:10.18637/jss.v056.i06.
- Jia, L., Wang, H., Jiang, L., & Du, W.** (2023). Weak fault detection of rolling element bearing combining robust EMD with adaptive maximum second-order cyclostationarity blind deconvolution. *Journal of Vibration and Control*, 29(9–10), 2374–2391. doi:10.1177/10775463221080229.
- Jiang, C., Guan, K., Wu, G., Peng, B., & Wang, S.** (2021). A daily, 250 m and real-time gross primary productivity product (2000–present) covering the contiguous United States. *Earth System Science Data*, 13(2), 281–298. doi:10.5194/essd-13-281-2021.
- Jiang, X., Kang, S., Tong, L., Li, F., Li, D., Ding, R., & Qiu, R.** (2014). Crop coefficient and evapotranspiration of grain maize modified by planting density in an arid region of northwest China. *Agricultural Water Management*, 142, 135–143. doi:10.1016/j.agwat.2014.05.006.
- Jiao, W., Wang, L., Novick, K. A., & Chang, Q.** (2019). A new station-enabled multi-sensor integrated index for drought monitoring. *Journal of Hydrology*, 574, 169–180. doi:10.1016/j.jhydrol.2019.04.037.
- Joiner, J., Yoshida, Y., Zhang, Y., Duveiller, G., Jung, M., Lyapustin, A., ... Tucker, C.** (2018). Estimation of Terrestrial Global Gross Primary Production (GPP) with Satellite Data-Driven Models and Eddy Covariance Flux Data. *Remote Sensing*, 10(9), 1346. doi:10.3390/rs10091346.
- Jolliffe, I. T., & Cadima, J.** (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065), 20150202. doi:10.1098/rsta.2015.0202.

- Jung, M., Koirala, S., Weber, U., Ichii, K., Gans, F., Camps-Valls, G., ... Reichstein, M.** (2019). The FLUXCOM ensemble of global land-atmosphere energy fluxes. *Scientific Data*, 6(1), 74. doi:10.1038/s41597-019-0076-8.
- Jung, M., Schwalm, C., Migliavacca, M., Walther, S., Camps-Valls, G., Koirala, S., ... Reichstein, M.** (2020). Scaling carbon fluxes from eddy covariance sites to globe: synthesis and evaluation of the FLUXCOM approach. *Biogeosciences*, 17(5), 1343–1365. doi:10.5194/bg-17-1343-2020.
- Kabir, E., Kumar, P., Kumar, S., Adelodun, A. A., & Kim, K. H.** (2018). Solar energy: Potential and future prospects. *Renewable and Sustainable Energy Reviews*, 82, 894–900. doi:10.1016/j.rser.2017.09.094.
- Kadow, C., Hall, D. M., & Ulbrich, U.** (2020). Artificial intelligence reconstructs missing climate information. *Nature Geoscience*, 13(6), 408–413. doi:10.1038/s41561-020-0582-5.
- Kanthavel, P., Saxena, C. K., & Singh, R. K.** (2022). Integrated Drought Index based on Vine Copula Modelling. *International Journal of Climatology*, 42(16), 9510–9529. doi:10.1002/joc.7840.
- Karimi, M., Shahedi, K., Razi, T., & Miryaghoubzadeh, M.** (2022). Meteorological and agricultural drought monitoring in Southwest of Iran using a remote sensing-based combined drought index. *Stochastic Environmental Research and Risk Assessment*, 36(11), 3707–3724. doi:10.1007/s00477-022-02220-3.
- Keenan, T. F., Luo, X., Stocker, B. D., De Kauwe, M. G., Medlyn, B. E., Prentice, I. C., ... & Zhou, S.** (2023). A constraint on historic growth in global photosynthesis due to rising CO₂. *Nature Climate Change*, 13(12), 1376–1381. doi:10.1038/s41558-023-01867-2.
- Keenan, T. F., Migliavacca, M., Papale, D., Baldocchi, D., Reichstein, M., Torn, M., & Wutzler, T.** (2019). Widespread inhibition of daytime ecosystem respiration. *Nature Ecology & Evolution*, 3(3), 407–415. doi:10.1038/s41559-019-0809-2.
- Kern, S.** (2021). *MODIS Collection 6 global 8-daily Gross Primary Production*. doi:10.25592/UHHFDM.8880.
- Khan, A. M., Stoy, P. C., Joiner, J., Baldocchi, D., Verfaillie, J., Chen, M., & Otkin, J. A.** (2022). The Diurnal Dynamics of Gross Primary Productivity Using Observations From the Advanced Baseline Imager on the Geostationary Operational Environmental Satellite-R Series at an Oak Savanna Ecosystem. *Journal of Geophysical Research: Biogeosciences*, 127(3). doi:10.1029/2021JG006701.
- Kim, D., & Oh, H. S.** (2009). EMD: A Package for Empirical Mode Decomposition and Hilbert Spectrum. *The R Journal*, 1(1), 40. doi:10.32614/RJ-2009-002.
- Kim, D., & Oh, H. S.** (2021). *EMD: Empirical Mode Decomposition and Hilbert Spectral Analysis* (R package version 1.5.9).

- Kim, H. jun, Chandrasekara, S., Kwon, H. han, Lima, C., & Kim, T. woong.** (2023). A novel multi-scale parameter estimation approach to the Hargreaves-Samani equation for estimation of Penman-Monteith reference evapotranspiration. *Agricultural Water Management*, 275(August 2022), 108038. doi:10.1016/j.agwat.2022.108038.
- Kim, K. Y., Haagensohn, R., Kansara, P., Rajaram, H., & Lakshmi, V.** (2024). Augmenting daily MODIS LST with AIRS surface temperature retrievals to estimate ground temperature and permafrost extent in High Mountain Asia. *Remote Sensing of Environment*, 305, 114075. doi:10.1016/j.rse.2024.114075.
- Kim, T., Ko, W., & Kim, J.** (2019). Analysis and impact evaluation of missing data imputation in day-ahead PV generation forecasting. *Applied Sciences (Switzerland)*, 9(1). doi:10.3390/app9010204.
- Kim, T. W., & Ahn, H.** (2009). Spatial rainfall model using a pattern classifier for estimating missing daily rainfall data. *Stochastic Environmental Research and Risk Assessment*, 23(3), 367–376. doi:10.1007/s00477-008-0223-9.
- Kisi, O.** (2010). Comment on ‘Nourani V, Mogaddam AA, Nadiri AO. 2008. An ANN-based model for spatiotemporal groundwater level forecasting. *Hydrological Processes* 22: 5054–5066. *Hydrological Processes*, 24(3), 368–369. doi:10.1002/hyp.7467.
- Kisi, O., & Alizamir, M.** (2018). Agricultural and Forest Meteorology Modelling reference evapotranspiration using a new wavelet conjunction heuristic method: Wavelet extreme learning machine vs wavelet neural networks. *Agricultural and Forest Meteorology*, 263, 41–48. doi:10.1016/j.agrformet.2018.08.007.
- Koc, K.** (2023). Role of Shapley Additive Explanations and Resampling Algorithms for Contract Failure Prediction of Public–Private Partnership Projects. *Journal of Management in Engineering*, 39(5). doi:10.1061/JMENEA.MEENG-5492.
- Kogan, F. N.** (1995). Application of vegetation index and brightness temperature for drought detection. *Advances in Space Research*, 15(11), 91–100. doi:10.1016/0273-1177(95)00079-T.
- Kogan, F. N.** (2001). Operational Space Technology for Global Vegetation Assessment. *Bulletin of the American Meteorological Society*, 82(9), 1949–1964. doi:10.1175/1520-0477(2001)082<1949:OSTFGV>2.3.CO;2.
- Koksal, E. S., Cemek, B., Cetin, S., Gowda, P. H., & Howell, T. A.** (2017). Estimating missing hourly climatic data using artificial neural network for energy balance based ET mapping applications. *Meteorological Applications*, 24(3), 457–465. doi:10.1002/met.1644.
- Kong, D., Yuan, D., Li, H., Zhang, J., Yang, S., Li, Y., Bai, Y., & Zhang, S.** (2023). Improving the Estimation of Gross Primary Productivity across Global Biomes by Modeling Light Use Efficiency through Machine Learning. *Remote Sensing*, 15(8), 2086. doi:10.3390/rs15082086.

- Körner, P., Kronenberg, R., Genzel, S., & Bernhofer, C.** (2018). Introducing Gradient Boosting as a universal gap filling tool for meteorological time series. *Meteorologische Zeitschrift*, 27(5), 369–376. doi:10.1127/metz/2018/0908.
- Koycegiz, C., Sen, O. L., & Buyukyildiz, M.** (2023). An analysis of terrestrial water storage changes of a karstic, endorheic basin in central Anatolia, Turkey. *Ecohydrology & Hydrobiology*, 23(4), 688–702. doi:10.1016/j.ecohyd.2023.07.002.
- Kruskal, W. H., & Wallis, W. A.** (1952). Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association*, 47(260), 583–621. doi:10.1080/01621459.1952.10483441.
- Kumar, R., Samaniego, L., & Attinger, S.** (2013). Implications of distributed hydrologic model parameterization on water fluxes at multiple scales and locations. *Water Resources Research*, 49(1), 360–379. doi:10.1029/2012WR012195.
- Kumari, P., Rehana, S., Singh, S. K., & Inayathulla, M.** (2023). Development of a new agro-meteorological drought index (SPAIE-Agro) in a data-scarce region. *Hydrological Sciences Journal*, 68(9), 1301–1322. doi:10.1080/02626667.2023.2206969.
- Kunwor, S., Starr, G., Loescher, H. W., & Staudhammer, C. L.** (2017). Preserving the variance in imputed eddy-covariance measurements: Alternative methods for defensible gap filling. *Agricultural and Forest Meteorology*, 232, 635–649. doi:10.1016/j.agrformet.2016.10.018.
- Laimighofer, J., & Laaha, G.** (2022). How standard are standardized drought indices? Uncertainty components for the SPI & SPEI case. *Journal of Hydrology*, 613, 128385. doi:10.1016/j.jhydrol.2022.128385.
- Landsberg, J. J., & Waring, R. H.** (1997). A generalised model of forest productivity using simplified concepts of radiation-use efficiency, carbon balance and partitioning. *Forest Ecology and Management*, 95(3), 209–228. doi:https://doi.org/10.1016/S0378-1127(97)00026-1.
- Langkamp, D. L., Lehman, A., & Lemeshow, S.** (2010). Techniques for Handling Missing Data in Secondary Analyses of Large Surveys. *Academic Pediatrics*, 10(3), 205–210. doi:10.1016/j.acap.2010.01.005.
- Lara-Estrada, L., Rasche, L., Sucar, L., & Schneider, U.** (2018). Inferring Missing Climate Data for Agricultural Planning Using Bayesian Networks. *Land*, 7(1), 4. doi:10.3390/land7010004.
- Lasslop, G., Reichstein, M., Papale, D., Richardson, A. D., Arneeth, A., Barr, A., ... Wohlfahrt, G.** (2010). Separation of net ecosystem exchange into assimilation and respiration using a light response curve approach: critical issues and global evaluation. *Global Change Biology*, 16(1), 187–208. doi:10.1111/j.1365-2486.2009.02041.x.
- Lee, D., Kim, J. S., Park, S. W., & Kug, J. S.** (2023). An abrupt shift in gross primary productivity over Eastern China-Mongolia and its inter-model diversity in land surface models. *Scientific Reports*, 13(1), 22971. doi:10.1038/s41598-023-49763-1.

- Lee, H. S.** (2013). Estimation of extreme sea levels along the Bangladesh coast due to storm surge and sea level rise using EEMD and EVA. *Journal of Geophysical Research: Oceans*, 118(9), 4273–4285. doi:10.1002/jgrc.20310.
- Lee, H. S., & Chou, M. T.** (2004). Fuzzy forecasting based on fuzzy time series. *International Journal of Computer Mathematics*, 81(7), 781–789. doi:10.1080/00207160410001712288.
- Lees, K. J., Quaife, T., Artz, R. R. E., Khomik, M., & Clark, J. M.** (2018). Potential for using remote sensing to estimate carbon fluxes across northern peatlands – A review. *Science of The Total Environment*, 615, 857–874. doi:10.1016/j.scitotenv.2017.09.103.
- Leuning, R., Zhang, Y. Q., Rajaud, A., Cleugh, H., & Tu, K.** (2008). A simple surface conductance model to estimate regional evaporation using MODIS leaf area index and the Penman-Monteith equation. *Water Resources Research*, 44(10). doi:10.1029/2007WR006562.
- Li, B., Ryu, Y., Jiang, C., Dechant, B., Liu, J., Yan, Y., & Li, X.** (2023). BESSv2.0: A satellite-based and coupled-process model for quantifying long-term global land–atmosphere fluxes. *Remote Sensing of Environment*, 295, 113696. doi:10.1016/j.rse.2023.113696.
- Li, J., An, X., Li, Q., Wang, C., Yu, H., Zhou, X., & Geng, Y. ao.** (2022). Application of XGBoost algorithm in the optimization of pollutant concentration. *Atmospheric Research*, 276, 106238. doi:10.1016/j.atmosres.2022.106238.
- Li, J., Li, Y., Yin, L., & Zhao, Q.** (2024). A novel composite drought index combining precipitation, temperature and evapotranspiration used for drought monitoring in the Huang-Huai-Hai Plain. *Agricultural Water Management*, 291. doi:10.1016/j.agwat.2023.108626.
- Li, L., Peng, Q., Wang, M., Cao, Y., Gu, X., & Cai, H.** (2024). Quantitative analysis of vegetation drought propagation process and uncertainty in the Yellow River Basin. *Agricultural Water Management*, 295, 108775. doi:10.1016/j.agwat.2024.108775.
- Li, T., Tang, J., Jiang, F., Xu, X., Li, C., Bai, J., & Ding, T.** (2019). Fill Missing Data for Wind Farms Using Long Short-Term Memory Based Recurrent Neural Network. *Proceedings of 2019 IEEE 3rd International Electrical and Energy Conference, CIEEC 2019*, 705–709. doi:10.1109/CIEEC47146.2019.CIEEC-2019284.
- Li, X., & Xiao, J.** (2019). Mapping photosynthesis solely from solar-induced chlorophyll fluorescence: A global, fine-resolution dataset of gross primary production derived from OCO-2. *Remote Sensing*, 11(21). doi:10.3390/rs11212563.
- Li, X., Xiao, J., & He, B.** (2018). Higher absorbed solar radiation partly offset the negative effects of water stress on the photosynthesis of Amazon forests during the 2015 drought. *Environmental Research Letters*, 13(4), 44005. doi:10.1088/1748-9326/aab0b1.

- Li, X., Zhang, W., Vermeulen, A., Dong, J., & Duan, Z.** (2023). Triple collocation-based merging of multi-source gridded evapotranspiration data in the Nordic Region. *Agricultural and Forest Meteorology*, 335, 109451. doi:10.1016/j.agrformet.2023.109451.
- Liang, C., Zhang, M., Wang, Z., Xiang, X., Gong, H., Wang, K., & Liu, H.** (2024). The strengthened impact of water availability at interannual and decadal time scales on vegetation <scp>GPP</scp>. *Global Change Biology*, 30(1). doi:10.1111/gcb.17138.
- Lin, Q., & Wang, J.** (2014). Vertically correlated echelon model for the interpolation of missing wind speed data. *IEEE Transactions on Sustainable Energy*, 5(3), 804–812. doi:10.1109/TSTE.2014.2304971.
- Lin, W. C., & Tsai, C. F.** (2020). Missing value imputation: a review and analysis of the literature (2006–2017). *Artificial Intelligence Review*, 53(2), 1487–1509. doi:10.1007/s10462-019-09709-4.
- Lischeid, G., Webber, H., Sommer, M., Nendel, C., & Ewert, F.** (2022). Machine learning in crop yield modelling: A powerful tool, but no surrogate for science. *Agricultural and Forest Meteorology*, 312, 108698. doi:10.1016/j.agrformet.2021.108698.
- Little, R. J. A., & Rubin, D. B.** (2002). *Statistical Analysis with Missing Data* (John Wiley & Sons (ed.)). John Wiley & Sons, Inc. doi:10.1002/9781119013563.
- Liu, H., Liu, Y., Chen, Y., Fan, M., Chen, Y., Gang, C., ... Wang, Z.** (2023). Dynamics of global dryland vegetation were more sensitive to soil moisture: Evidence from multiple vegetation indices. *Agricultural and Forest Meteorology*, 331, 109327. doi:10.1016/j.agrformet.2023.109327.
- Liu, H., Xu, X., Lin, Z., Zhang, M., Mi, Y., Huang, C., & Yang, H.** (2016). Climatic and human impacts on quasi-periodic and abrupt changes of sedimentation rate at multiple time scales in Lake Taihu, China. *Journal of Hydrology*, 543, 739–748. doi:10.1016/j.jhydrol.2016.10.046.
- Liu, S., & Sun, W.** (2023). Attention mechanism-aided data- and knowledge-driven soft sensors for predicting blast furnace gas generation. *Energy*, 262(PA), 125498. doi:10.1016/j.energy.2022.125498.
- Liu, T., Wei, H., & Zhang, K.** (2018). Wind power prediction with missing data using Gaussian process regression and multiple imputation. *Applied Soft Computing Journal*, 71, 905–916. doi:10.1016/j.asoc.2018.07.027.
- Liu, W., Chen, Z., & Hu, Y.** (2022). XGBoost algorithm-based prediction of safety assessment for pipelines. *International Journal of Pressure Vessels and Piping*, 197, 104655. doi:10.1016/j.ijpvp.2022.104655.
- Liu, W., Li, Z., Li, Y., Ye, T., Chen, S., & Liu, Y.** (2022). Heterogeneous impacts of excessive wetness on maize yields in China: Evidence from statistical yields and process-based crop models. *Agricultural and Forest Meteorology*, 327, 109205. doi:10.1016/j.agrformet.2022.109205.

- Liu, X., Cui, Y., Li, W., Li, M., Li, N., Shi, Z., Dong, J., & Xiao, X.** (2023). Urbanization expands the fluctuating difference in gross primary productivity between urban and rural areas from 2000 to 2018 in China. *Science of The Total Environment*, 901, 166490. doi:10.1016/j.scitotenv.2023.166490.
- Lo Presti, R., Barca, E., & Passarella, G.** (2010). A methodology for treating missing data applied to daily rainfall data in the Candelaro River Basin (Italy). *Environmental Monitoring and Assessment*, 160(1–4), 1–22. doi:10.1007/s10661-008-0653-3.
- Lompar, M., Lalić, B., Dekić, L., & Petrić, M.** (2019). Filling Gaps in Hourly Air Temperature Data Using Debaised ERA5 Data. In *Atmosphere* (Vol. 10, Issue 1). doi:10.3390/atmos10010013.
- López, J. L., Hernández, S., Urrutia, A., López-Cortés, X. A., Araya, H., & Morales-Salinas, L.** (2021). Effect of missing data on short time series and their application in the characterization of surface temperature by detrended fluctuation analysis. *Computers and Geosciences*, 153(March), 104794. doi:10.1016/j.cageo.2021.104794.
- Lu, J., Carbone, G. J., & Gao, P.** (2017). Detrending crop yield data for spatial visualization of drought impacts in the United States, 1895–2014. *Agricultural and Forest Meteorology*, 237–238, 196–208. doi:10.1016/j.agrformet.2017.02.001.
- Lu, L. C., Chiu, S. Y., Chiu, Y. ho, & Chang, T. H.** (2022). Sustainability efficiency of climate change and global disasters based on greenhouse gas emissions from the parallel production sectors – A modified dynamic parallel three-stage network DEA model. *Journal of Environmental Management*, 317, 115401. doi:10.1016/j.jenvman.2022.115401.
- Lu, Q., Liu, H., Wei, L., Zhong, Y., & Zhou, Z.** (2024). Global prediction of gross primary productivity under future climate change. *Science of The Total Environment*, 912, 169239. doi:10.1016/j.scitotenv.2023.169239.
- Lu, Y., Li, T., Hu, H., & Zeng, X.** (2023). Short-term prediction of reference crop evapotranspiration based on machine learning with different decomposition methods in arid areas of China. *Agricultural Water Management*, 279, 108175. doi:10.1016/j.agwat.2023.108175.
- Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., ... Lee, S. I.** (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1), 56–67. doi:10.1038/s42256-019-0138-9.
- Luo, R., Guo, L., Li, X., Tuo, J., Lei, C., & Zhou, Y.** (2021). An eXtreme Gradient Boosting Algorithm Combining Artificial Bee Colony Parameters Optimized Technique for Single Sand Body Identification. *IEEE Access*, 9, 156894–156906. doi:10.1109/ACCESS.2021.3129830.
- Luo, Y. F., & Lu, K. Y.** (2022). An online state of health estimation technique for lithium-ion battery using artificial neural network and linear interpolation. *Journal of Energy Storage*, 52(PC), 105062. doi:10.1016/j.est.2022.105062.

- Lv, Y., Li, X., & Chi, W.** (2023). Maximum Gross Primary Productivity Dominates the Trend in Gross Primary Productivity in China's Deciduous Forest Ecosystems. *Forests*, 14(9), 1–15. doi:10.3390/f14091880.
- Lv, Y., Liu, J., He, W., Zhou, Y., Tu Nguyen, N., Bi, W., ... Chen, H.** (2023). How well do light-use efficiency models capture large-scale drought impacts on vegetation productivity compared with data-driven estimates? *Ecological Indicators*, 146, 109739. doi:10.1016/j.ecolind.2022.109739.
- Ma, J., Xiao, X., Miao, R., Li, Y., Chen, B., Zhang, Y., & Zhao, B.** (2019). Trends and controls of terrestrial gross primary productivity of China during 2000–2016. *Environmental Research Letters*, 14(8), 084032. doi:10.1088/1748-9326/ab31e4.
- Ma, Y., Wang, J., Xiong, J., Sun, M., & Wang, J.** (2024). Risk assessment for cropland abandonment in mountainous area based on AHP and PCA—Take Yunnan Province in China as an example. *Ecological Indicators*, 158, 111287. doi:10.1016/j.ecolind.2023.111287.
- Madani, N., Parazoo, N. C., Kimball, J. S., Ballantyne, A. P., Reichle, R. H., Maneta, M., ... Tagesson, T.** (2020). Recent Amplified Global Gross Primary Productivity Due to Temperature Increase Is Offset by Reduced Productivity Due to Water Constraints. *AGU Advances*, 1(4). doi:10.1029/2020AV000180.
- Mahabbati, A., Beringer, J., Leopold, M., McHugh, I., Cleverly, J., Isaac, P., & Izady, A.** (2021). A comparison of gap-filling algorithms for eddy covariance fluxes and their drivers. *Geoscientific Instrumentation, Methods and Data Systems*, 10(1), 123–140. doi:10.5194/gi-10-123-2021.
- Mäkelä, A., Hari, P., Berninger, F., Hänninen, H., & Nikinmaa, E.** (2004). Acclimation of photosynthetic capacity in Scots pine to the annual cycle of temperature. *Tree Physiology*, 24(4), 369–376. doi:10.1093/treephys/24.4.369.
- Malhi, Y.** (2012). The productivity, metabolism and carbon cycle of tropical forest vegetation. *Journal of Ecology*, 100(1), 65–75. doi:10.1111/j.1365-2745.2011.01916.x.
- Mann, H. B.** (1945). Nonparametric Tests Against Trend. *Econometrica*, 13(3), 245. doi:10.2307/1907187.
- Mann, H. B., & Whitney, D. R.** (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1), 50–60. doi:10.1214/aoms/1177730491.
- Mao, Y., & Jian, M.** (2016). Data completing of missing wind power data based on adaptive BP neural network. *2016 International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)*, 1–6. doi:10.1109/PMAPS.2016.7764057.

- Marzo, A., Trigo-Gonzalez, M., Alonso-Montesinos, J., Martínez-Durbán, M., López, G., Ferrada, P., ... Batlles, F. J.** (2017). Daily global solar radiation estimation in desert areas using daily extreme temperatures and extraterrestrial radiation. *Renewable Energy*, 113, 303–311. doi:10.1016/j.renene.2017.01.061.
- Maselli, F., Barbati, A., Chiesi, M., Chirici, G., & Corona, P.** (2006). Use of remotely sensed and ancillary data for estimating forest gross primary productivity in Italy. *Remote Sensing of Environment*, 100(4), 563–575. doi:10.1016/j.rse.2005.11.010.
- McKee, T. B., Doesken, N. J., & Kleist, J.** (1993). The Relationship of Drought Frequency and Duration to Time Scales. *Proceedings of the 8th Conference on Applied Climatology*, 179–184. <https://www.sciencedirect.com/reference/177494>.
- McMurtrie, R. E., Gholz, H. L., Linder, S., & Gower, S. T.** (1994). Climatic Factors Controlling the Productivity of Pine Stands: A Model-Based Analysis. *Ecological Bulletins*, 43, 173–188. <http://www.jstor.org/stable/20113140>.
- Medlyn, B., Barrett, D., Landsberg, J., Sands, P., & Clement, R.** (2003). Conversion of canopy intercepted radiation to photosynthate: review of modelling approaches for regional scales. *Functional Plant Biology*, 30(2), 153–169. <https://doi.org/10.1071/FP02088>.
- Mehdizadeh, S., Behmanesh, J., & Khalili, K.** (2017). Using MARS, SVM, GEP and empirical equations for estimation of monthly mean reference evapotranspiration. *Computers and Electronics in Agriculture*, 139, 103–114. doi:10.1016/j.compag.2017.05.002.
- Meher, J. K., & Das, L.** (2019). Gridded data as a source of missing data replacement in station records. *Journal of Earth System Science*, 128(3), 58. doi:10.1007/s12040-019-1079-8.
- Mi, X., & Zhao, S.** (2020). Wind speed prediction based on singular spectrum analysis and neural network structural learning. *Energy Conversion and Management*, 216, 112956. doi:10.1016/j.enconman.2020.112956.
- Mirabbasi, R., Ahmadi, F., & Jhajharia, D.** (2020). Comparison of parametric and non-parametric methods for trend identification in groundwater levels in Sirjan plain aquifer, Iran. *Hydrology Research*, 51(6), 1455–1477. doi:10.2166/nh.2020.041.
- Miró, J. J., Caselles, V., & Estrela, M. J.** (2017). Multiple imputation of rainfall missing data in the Iberian Mediterranean context. *Atmospheric Research*, 197, 313–330. doi:10.1016/j.atmosres.2017.07.016.
- Mishra, A. K., & Singh, V. P.** (2010). A review of drought concepts. *Journal of Hydrology*, 391(1–2), 202–216. doi:10.1016/j.jhydrol.2010.07.012.
- Moazenzadeh, R., Mohammadi, B., Duan, Z., & Delghandi, M.** (2022). Improving generalisation capability of artificial intelligence-based solar radiation estimator models using a bio-inspired optimisation algorithm and multi-model approach. *Environmental Science and Pollution Research*, 29(19), 27719–27737. doi:10.1007/s11356-021-17852-1.

- Mohamad, N. B., Lai, A. chow, & Lim, B. han.** (2022). A case study in the tropical region to evaluate univariate imputation methods for solar irradiance data with different weather types. *Sustainable Energy Technologies and Assessments*, 50, 101764. doi:10.1016/j.seta.2021.101764.
- Mohammadi, B., & Moazenzadeh, R.** (2021). Performance Analysis of Daily Global Solar Radiation Models in Peru by Regression Analysis. *Atmosphere*, 12(3), 389. doi:10.3390/atmos12030389.
- Mongeon, P., & Paul-Hus, A.** (2016). The journal coverage of Web of Science and Scopus: a comparative analysis. *Scientometrics*, 106(1), 213–228. doi:10.1007/s11192-015-1765-5.
- Monteith, J. L.** (1972). Solar Radiation and Productivity in Tropical Ecosystems. *Journal of Applied Ecology*, 9(3), 747–766. doi:10.2307/2401901.
- Moore, C. E., Beringer, J., Donohue, R. J., Evans, B., Exbrayat, J. F., Hutley, L. B., & Tapper, N. J.** (2018). Seasonal, interannual and decadal drivers of tree and grass productivity in an Australian tropical savanna. *Global Change Biology*, 24(6), 2530–2544. doi:10.1111/gcb.14072.
- Moradi, I.** (2009). Quality control of global solar radiation using sunshine duration hours. *Energy*, 34(1), 1–6. doi:10.1016/j.energy.2008.09.006.
- Muhammad, R., Mostafa, R. R., Reza, A., Islam, T., Kisi, O., Kuriqi, A., & Heddad, S.** (2021). Estimating reference evapotranspiration using hybrid adaptive fuzzy inferencing coupled with heuristic algorithms. *Computers and Electronics in Agriculture*, 191, 106541. doi:10.1016/j.compag.2021.106541.
- Naeem, S., Zhang, Y., Zhang, X., Rehman, A. U., Tang, Z., Xu, Z., ... Azeem, T.** (2023). Recent change in ecosystem water use efficiency in China mainly dominated by vegetation greening and increased CO₂. *Remote Sensing of Environment*, 298, 113811. doi:10.1016/j.rse.2023.113811.
- Nalbantis, I., & Tsakiris, G.** (2009). Assessment of Hydrological Drought Revisited. *Water Resources Management*, 23(5), 881–897. doi:10.1007/s11269-008-9305-1.
- Narasimhan, B., & Srinivasan, R.** (2005). Development and evaluation of Soil Moisture Deficit Index (SMDI) and Evapotranspiration Deficit Index (ETDI) for agricultural drought monitoring. *Agricultural and Forest Meteorology*, 133(1–4), 69–88. doi:10.1016/j.agrformet.2005.07.012.
- Narvaez, G., Giraldo, L. F., Bressan, M., & Pantoja, A.** (2021). Machine learning for site-adaptation and solar radiation forecasting. *Renewable Energy*, 167, 333–342. doi:10.1016/j.renene.2020.11.089.
- Nash, E., & Sutcliffe, V.** (1970). River flow forecasting Through conceptual models PART I- A Discussion of principles. *Journal of Hydrology*, 10, 282–290..
- Nash, J. E., & Sutcliffe, J. V.** (1970). River flow forecasting through conceptual models part I — A discussion of principles. *Journal of Hydrology*, 10(3), 282–290. doi:10.1016/0022-1694(70)90255-6.

- Nemani, R. R., Keeling, C. D., Hashimoto, H., Jolly, W. M., Piper, S. C., Tucker, C. J., ... Running, S. W. (2003). Climate-Driven Increases in Global Terrestrial Net Primary Production from 1982 to 1999. *Science*, 300(5625), 1560–1563. doi:10.1126/science.1082750.
- Nigatu, Z. M., Fan, D., You, W., & Melesse, A. M. (2021). Hydroclimatic Extremes Evaluation Using GRACE/GRACE-FO and Multidecadal Climatic Variables over the Nile River Basin. *Remote Sensing*, 13(4), 651. doi:10.3390/rs13040651.
- O, S., & Park, S. K. (2024). Global ecosystem responses to flash droughts are modulated by background climate and vegetation conditions. *Communications Earth & Environment*, 5(1), 88. doi:10.1038/s43247-024-01247-4.
- Ogunsola, O. T., & Song, L. (2014). Restoration of long-term missing gaps in solar radiation. *Energy and Buildings*, 82, 580–591. doi:10.1016/j.enbuild.2014.07.088.
- Ollinger, S. V., Richardson, A. D., Martin, M. E., Hollinger, D. Y., Frolking, S. E., Reich, P. B., ... Schmid, H. P. (2008). Canopy nitrogen, carbon assimilation, and albedo in temperate and boreal forests: Functional relations and potential climate feedbacks. *Proceedings of the National Academy of Sciences*, 105(49), 19336–19341. doi:10.1073/pnas.0810021105.
- Olmos-Trujillo, E., González-Trinidad, J., Júnez-Ferreira, H., Pacheco-Guerrero, A., Bautista-Capetillo, C., Avila-Sandoval, C., & Galván-Tejada, E. (2020). Spatio-Temporal Response of Vegetation Indices to Rainfall and Temperature in A Semiarid Region. *Sustainability*, 12(5), 1939. doi:10.3390/su12051939.
- Orhan, O., Yakar, M., & Ekercin, S. (2020). An application on sinkhole susceptibility mapping by integrating remote sensing and geographic information systems. *Arabian Journal of Geosciences*, 13(17), 886. doi:10.1007/s12517-020-05841-6.
- Orieschnig, C., & Cavus, Y. (2024). Spatial characterization of drought through CHIRPS and a station-based dataset in the Eastern Mediterranean. *Proceedings of IAHS*, 385, 79–84. doi:10.5194/piahs-385-79-2024.
- Orimoloye, I. R. (2022). Agricultural Drought and Its Potential Impacts: Enabling Decision-Support for Food Security in Vulnerable Regions. *Frontiers in Sustainable Food Systems*, 6. doi:10.3389/fsufs.2022.838824.
- Osman, M. S., Abu-Mahfouz, A. M., & Page, P. R. (2018). A Survey on Data Imputation Techniques: Water Distribution System as a Use Case. *IEEE Access*, 6, 63279–63291. doi:10.1109/ACCESS.2018.2877269.
- Otkin, J. A., Svoboda, M., Hunt, E. D., Ford, T. W., Anderson, M. C., Hain, C., & Basara, J. B. (2018). Flash Droughts: A Review and Assessment of the Challenges Imposed by Rapid-Onset Droughts in the United States. *Bulletin of the American Meteorological Society*, 99(5), 911–919. doi:10.1175/BAMS-D-17-0149.1.

- Owen, K. E., Tenhunen, J., Reichstein, M., Wang, Q., Falge, E., Geyer, R., ... Vogel, C.** (2007). Linking flux network measurements to continental scale simulations: ecosystem carbon dioxide exchange capacity under non-water-stressed conditions. *Global Change Biology*, 13(4), 734–760. doi:10.1111/j.1365-2486.2007.01326.x.
- Ozcan, O., & Ersoz, F.** (2019). Project and cost-based evaluation of solar energy performance in three different geographical regions of Turkey: Investment analysis application. *Engineering Science and Technology, an International Journal*, 22(4), 1098–1106. doi:10.1016/j.jestch.2019.04.001.
- Ozdogan, M., Woodcock, C. E., Salvucci, G. D., & Demir, H.** (2006). Changes in Summer Irrigated Crop Area and Water Use in Southeastern Turkey from 1993 to 2002: Implications for Current and Future Water Resources. *Water Resources Management*, 20(3), 467–488. doi:10.1007/s11269-006-3087-0.
- Özger, M., Başakın, E. E., Ekmekcioğlu, Ö., & Hacısüleyman, V.** (2020). Comparison of wavelet and empirical mode decomposition hybrid models in drought prediction. *Computers and Electronics in Agriculture*, 179, 105851. doi:10.1016/j.compag.2020.105851.
- Pabon-Moreno, D. E., Migliavacca, M., Reichstein, M., & Mahecha, M. D.** (2022). On the Potential of Sentinel-2 for Estimating Gross Primary Production. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–12. doi:10.1109/TGRS.2022.3152272.
- Pagliacci, F., & Salpina, D.** (2022). Territorial hotspots of exposure to climate disaster risk. The case of agri-food geographical indications in the Veneto Region. *Land Use Policy*, 123, 106404. doi:10.1016/j.landusepol.2022.106404.
- Pan, S., Zheng, Z., Guo, Z., & Luo, H.** (2022). An optimized XGBoost method for predicting reservoir porosity using petrophysical logs. *Journal of Petroleum Science and Engineering*, 208, 109520. doi:10.1016/j.petrol.2021.109520.
- Panahi, M., Rahmati, O., Kalantari, Z., Darabi, H., Rezaie, F., Moghaddam, D. D., ... Lee, S.** (2022). Large-scale dynamic flood monitoring in an arid-zone floodplain using SAR data and hybrid machine-learning models. *Journal of Hydrology*, 611, 128001. doi:10.1016/j.jhydrol.2022.128001.
- Park, I., Kim, H. S., Lee, J., Kim, J. H., Song, C. H., & Kim, H. K.** (2019). Temperature prediction using the missing data refinement model based on a long short-term memory neural network. *Atmosphere*, 10(11), 1–16. doi:10.3390/atmos10110718.
- Parthsarathi Pandya, Gontia N. K., & Parmar H.V.** (2022). Development of PCA-based composite drought index for agricultural drought assessment using remote- sensing. *Journal of Agrometeorology*, 24(4), 384–392. doi:10.54386/jam.v24i4.1738.

- Pasha, S. J., & Mohamed, E. S.** (2022). Advanced hybrid ensemble gain ratio feature selection model using machine learning for enhanced disease risk prediction. *Informatics in Medicine Unlocked*, 32, 101064. doi:10.1016/j.imu.2022.101064.
- Pastorello, G., Trotta, C., Canfora, E., Chu, H., Christianson, D., Cheah, Y. W., ... Papale, D.** (2020). The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data. *Scientific Data*, 7(1), 225. doi:10.1038/s41597-020-0534-3.
- Patakamuri, S. K., & Das, B.** (2022). *trendchange: Innovative Trend Analysis and Time-Series Change Point Analysis* (R package version 1.2). <https://cran.r-project.org/package=trendchange>.
- Pathak, A. A., & Dodamani, B. M.** (2021). Connection between Meteorological and Groundwater Drought with Copula-Based Bivariate Frequency Analysis. *Journal of Hydrologic Engineering*, 26(7). doi:10.1061/(ASCE)HE.1943-5584.0002089.
- Patil, P. P., Jagtap, M. P., Khatri, N., Madan, H., Vadduri, A. A., & Patodia, T.** (2024). Exploration and advancement of NDDI leveraging NDVI and NDWI in Indian semi-arid regions: A remote sensing-based study. *Case Studies in Chemical and Environmental Engineering*, 9, 100573. doi:10.1016/j.csee.2023.100573.
- Pei, Y., Dong, J., Zhang, Y., Yang, J., Zhang, Y., Jiang, C., & Xiao, X.** (2020). Performance of four state-of-the-art GPP products (VPM, MOD17, BESS and PML) for grasslands in drought years. *Ecological Informatics*, 56, 101052. doi:10.1016/j.ecoinf.2020.101052.
- Pei, Y., Dong, J., Zhang, Y., Yuan, W., Doughty, R., Yang, J., ... Xiao, X.** (2022). Evolution of light use efficiency models: Improvement, uncertainties, and implications. *Agricultural and Forest Meteorology*, 317, 108905. doi:https://doi.org/10.1016/j.agrformet.2022.108905.
- Pendergrass, A. G., Meehl, G. A., Pulwarty, R., Hobbins, M., Hoell, A., AghaKouchak, A., ... Woodhouse, C. A.** (2020). Flash droughts present a new challenge for subseasonal-to-seasonal prediction. *Nature Climate Change*, 10(3), 191–199. doi:10.1038/s41558-020-0709-0.
- Pereira, G. M. S., Stonoga, R. L. B., Detzel, D. H. M., Kuster, K. K., Neto, R. A. P., & Catussi Paschoalotto, L. A.** (2018). Analysis and Evaluation of Gap Filling Procedures for Solar Radiation Data. *2018 IEEE 9th Power, Instrumentation and Measurement Meeting (EPIM)*, 1–6. doi:10.1109/EPIM.2018.8756358.
- Piao, S., Wang, X., Park, T., Chen, C., Lian, X., He, Y., ... Myneni, R. B.** (2019). Characteristics, drivers and feedbacks of global greening. *Nature Reviews Earth & Environment*, 1(1), 14–27. doi:10.1038/s43017-019-0001-x.
- Potapov, P., Hansen, M. C., Pickens, A., Hernandez-Serna, A., Tyukavina, A., Turubanova, S., ... Kommareddy, A.** (2022). The Global 2000-2020 Land Cover and Land Use Change Dataset Derived From the Landsat Archive: First Results. *Frontiers in Remote Sensing*, 3. doi:10.3389/frsen.2022.856903.

- Potter, C. S., Randerson, J. T., Field, C. B., Matson, P. A., Vitousek, P. M., Mooney, H. A., & Klooster, S. A.** (1993). Terrestrial ecosystem production: A process model based on global satellite and surface data. *Global Biogeochemical Cycles*, 7(4), 811–841. doi:10.1029/93GB02725.
- Potter, C. S., Randerson, J. T., Field, C. B., Matson, P. A., Vitousek, P. M., Mooney, H. A., & Klooster, S. A.** (1993). Terrestrial ecosystem production: A process model based on global satellite and surface data. *Global Biogeochemical Cycles*, 7(4), 811–841. doi:10.1029/93GB02725.
- Priestley, C. H. B., & Taylor, R. J.** (1972). On the Assessment of Surface Heat Flux and Evaporation Using Large-Scale Parameters. *Monthly Weather Review*, 100(2), 81–92. doi:10.1175/1520-0493(1972)100<0081:OTAOSH>2.3.CO;2.
- Qiu, R., Han, G., Ma, X., Xu, H., Shi, T., & Zhang, M.** (2020). A Comparison of OCO-2 SIF, MODIS GPP, and GOSIF Data from Gross Primary Production (GPP) Estimation and Seasonal Cycles in North America. *Remote Sensing*, 12(2), 258. doi:10.3390/rs12020258.
- Qiu, R., Li, L., Wu, L., Agathokleous, E., Liu, C., & Zhang, B.** (2022). Comparison of machine learning and dynamic models for predicting actual vapour pressure when psychrometric data are unavailable. *Journal of Hydrology*, 610, 127989. doi:10.1016/j.jhydrol.2022.127989.
- Rahman, A. F., Sims, D. A., Cordova, V. D., & El-Masri, B. Z.** (2005). Potential of MODIS EVI and surface temperature for directly estimating per-pixel ecosystem C fluxes. *Geophysical Research Letters*, 32(19). doi:10.1029/2005GL024127.
- Rahman, K. U., Ejaz, N., Shang, S., Balkhair, K. S., Alghamdi, K. M., Zaman, K., ... Hussain, A.** (2024). A robust integrated agricultural drought index under climate and land use variations at the local scale in Pakistan. *Agricultural Water Management*, 295, 108748. doi:10.1016/j.agwat.2024.108748.
- Ranjbar, S., Losos, D., Hoffman, S., Cuntz, M., & Stoy, P. C.** (2024). Geostationary satellite observations can accurately estimate ecosystem carbon uptake and respiration at half hourly time steps at eddy covariance sites. *Available at SSRN*. doi:10.2139/ssrn.4671918.
- Raphael, M. W., Benedict, M. M., & James, M. R.** (2017). Analysis of spatial and temporal drought variability in a tropical river basin using Palmer Drought Severity Index (PDSI). *International Journal of Water Resources and Environmental Engineering*, 9(8), 178–190. doi:10.5897/IJWREE2017.0723.
- Raposo, V. de M. B., Costa, V. A. F., & Rodrigues, A. F.** (2023). A review of recent developments on drought characterization, propagation, and influential factors. *Science of The Total Environment*, 898, 165550. doi:10.1016/j.scitotenv.2023.165550.

- Raza, A., Hussain, I., Ali, Z., Faisal, M., Elashkar, E. E., Shoukry, A. M., Al-Deek, F. F., & Gani, S.** (2021). A seasonally blended and regionally integrated drought index using <scp>Bayesian</scp> network theory. *Meteorological Applications*, 28(3). doi:10.1002/met.1992.
- Reichstein, M., Falge, E., Baldocchi, D., Papale, D., Aubinet, M., Berbigier, P., ... Valentini, R.** (2005). On the separation of net ecosystem exchange into assimilation and ecosystem respiration: review and improved algorithm. *Global Change Biology*, 11(9), 1424–1439. doi:10.1111/j.1365-2486.2005.001002.x.
- Reichstein, M., Stoy, P. C., Desai, A. R., Lasslop, G., & Richardson, A. D.** (2012). Partitioning of Net Fluxes. In M. Aubinet, T. Vesala & D. Papale (Eds.), *Eddy Covariance* (pp. 263–289). Springer Netherlands. doi:10.1007/978-94-007-2351-1_9.
- Reyes-Muñoz, P., D.Kovács, D., Berger, K., Pipia, L., Belda, S., Rivera-Caicedo, ... Verrelst, J.** (2024). Inferring global terrestrial carbon fluxes from the synergy of Sentinel 3 & 5P with Gaussian process hybrid models. *Remote Sensing of Environment*, 305, 114072. doi:https://doi.org/10.1016/j.rse.2024.114072.
- Richardson, A. D., Hollinger, D. Y., Burba, G. G., Davis, K. J., Flanagan, L. B., Katul, G. G., ... Wofsy, S. C.** (2006). A multi-site analysis of random error in tower-based measurements of carbon and energy fluxes. *Agricultural and Forest Meteorology*, 136(1–2), 1–18. doi:10.1016/j.agrformet.2006.01.007.
- Rubin, D. B.** (1976). Inference and missing data. *Biometrika*, 63(3), 581–592. doi:10.1093/biomet/63.3.581.
- Running, S. W., Baldocchi, D. D., Turner, D. P., Gower, S. T., Bakwin, P. S., & Hibbard, K. A.** (1999). A Global Terrestrial Monitoring Network Integrating Tower Fluxes, Flask Sampling, Ecosystem Modeling and EOS Satellite Data. *Remote Sensing of Environment*, 70(1), 108–127. doi:10.1016/S0034-4257(99)00061-9.
- Running, S. W., Nemani, R. R., Heinsch, F. A., Zhao, M., Reeves, M., & Hashimoto, H.** (2004). A continuous satellite-derived measure of global terrestrial primary production. *BioScience*, 54(6), 547–560. doi:10.1641/0006-3568(2004)054[0547:ACSMOG]2.0.CO;2.
- Rwigema, J., Choi, H. R., & Kim, T.** (2019). A differential evolution approach to optimize weights of dynamic time warping for multi-sensor based gesture recognition. *Sensors (Switzerland)*, 19(5), 1–16. doi:10.3390/s19051007.
- Saaban, A., Zainudin, M. L., & Abu Bakar, M. N.** (2016). Piecewise positivity preserving cubic bezier interpolation for estimating solar radiation missing value in Penang, Malaysia. *Journal of Mathematics and Statistics*, 12(4), 302–307. doi:10.3844/jmssp.2016.302.307.

- Saatchi, S. S., Harris, N. L., Brown, S., Lefsky, M., Mitchard, E. T. A., Salas, W., ... Morel, A.** (2011). Benchmark map of forest carbon stocks in tropical regions across three continents. *Proceedings of the National Academy of Sciences*, 108(24), 9899–9904. doi:10.1073/pnas.1019576108.
- Salvia, A. L., Brandli, L. L., Leal Filho, W., & Locatelli Kalil, R. M.** (2019). An analysis of the applications of Analytic Hierarchy Process (AHP) for selection of energy efficiency practices in public lighting in a sample of Brazilian cities. *Energy Policy*, 132, 854–864. doi:10.1016/j.enpol.2019.06.021.
- Samaniego, L., Kumar, R., & Attinger, S.** (2010). Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale. *Water Resources Research*, 46(5). doi:10.1029/2008WR007327.
- Sánchez, J. M., Rodríguez, J. P., & Espitia, H. E.** (2022). Bibliometric analysis of publications discussing the use of the artificial intelligence technique agent-based models in sustainable agriculture. *Heliyon*, 8(12), e12005. doi:10.1016/j.heliyon.2022.e12005.
- Santhosh, M., Venkaiah, C., & Vinod Kumar, D. M.** (2018). Ensemble empirical mode decomposition based adaptive wavelet neural network method for wind speed prediction. *Energy Conversion and Management*, 168(November 2017), 482–493. doi:10.1016/j.enconman.2018.04.099.
- Sarkar, D. P., Uma Shankar, B., & Ranjan Parida, B.** (2024). A novel approach for retrieving GPP of evergreen forest regions of India using random forest regression. *Remote Sensing Applications: Society and Environment*, 33, 101116. doi:10.1016/j.rsase.2023.101116.
- Sattari, M. T., Falsafian, K., Irvem, A., S, S., & Qasem, S. N.** (2020). Potential of kernel and tree-based machine-learning models for estimating missing data of rainfall. *Engineering Applications of Computational Fluid Mechanics*, 14(1), 1078–1094. doi:10.1080/19942060.2020.1803971.
- Schimel, D., Pavlick, R., Fisher, J. B., Asner, G. P., Saatchi, S., Townsend, P., ... Cox, P.** (2015). Observing terrestrial ecosystems and the carbon cycle from space. *Global Change Biology*, 21(5), 1762–1776. doi:10.1111/gcb.12822.
- Schimel, D. S.** (1995). Terrestrial ecosystems and the carbon cycle. *Global Change Biology*, 1(1), 77–91. doi:10.1111/j.1365-2486.1995.tb00008.x.
- Schmidt, A., Wrzesinsky, T., & Klemm, O.** (2008). Gap filling and quality assessment of CO₂ and water vapour fluxes above an urban area with radial basis function neural networks. *Boundary-Layer Meteorology*, 126(3), 389–413. doi:10.1007/s10546-007-9249-7.
- Schwandt, M., Chhatbar, K., Meyer, R., Fross, K., Mitra, I., Vashistha, R., Giridhar, G., Gomathinayagam, S., & Kumar, A.** (2014). Development and Test of Gap Filling Procedures for Solar Radiation Data of the Indian SRRA Measurement Network. *Energy Procedia*, 57, 1100–1109. doi:10.1016/j.egypro.2014.10.096.

- Selek, B., Tuncok, I. K., & Selek, Z.** (2018). Changes in climate zones across Turkey. *Journal of Water and Climate Change*, 9(1), 178–195. doi:10.2166/wcc.2017.121.
- Sen, P. K.** (1968). Estimates of the Regression Coefficient Based on Kendall's Tau. *Journal of the American Statistical Association*, 63(324), 1379–1389. doi:10.1080/01621459.1968.10480934.
- Şen, Z.** (2012). Innovative Trend Analysis Methodology. *Journal of Hydrologic Engineering*, 17(9), 1042–1046. doi:10.1061/(ASCE)HE.1943-5584.0000556.
- Serencam, U., Ekmekcioğlu, Ö., Başakın, E. E., & Özger, M.** (2022). Determining the water level fluctuations of Lake Van through the integrated machine learning methods. *International Journal of Global Warming*, 27(2), 123–142. doi:10.1504/IJGW.2022.123278.
- Sextstone, G. A., Clow, D. W., Stannard, D. I., & Fassnacht, S. R.** (2016). Comparison of methods for quantifying surface sublimation over seasonally snow-covered terrain. *Hydrological Processes*, 30(19), 3373–3389. doi:10.1002/hyp.10864.
- Shah, W. U. H., Hao, G., Yasmeen, R., Yan, H., Shen, J., & Lu, Y.** (2023). Role of China's agricultural water policy reforms and production technology heterogeneity on agriculture water usage efficiency and total factor productivity change. *Agricultural Water Management*, 287(June), 108429. doi:10.1016/j.agwat.2023.108429.
- Sharma, A.** (2024). Nexus of Drought, Relief Finances, and Economic Growth: Evidence from Indian States. *Natural Hazards Review*, 25(4). doi:10.1061/NHREFO.NHENG-1984.
- Shashikant, V., Mohamed Shariff, A. R., Wayayok, A., Kamal, M. R., Lee, Y. P., & Takeuchi, W.** (2021). Utilizing TVDI and NDWI to Classify Severity of Agricultural Drought in Chuping, Malaysia. *Agronomy*, 11(6), 1243. doi:10.3390/agronomy11061243.
- Shen, Z., Zhang, Q., Singh, V. P., Sun, P., Song, C., & Yu, H.** (2019). Agricultural drought monitoring across Inner Mongolia, China: Model development, spatiotemporal patterns and impacts. *Journal of Hydrology*, 571, 793–804. doi:10.1016/j.jhydrol.2019.02.028.
- Shi, X., Ding, H., Wu, M., Shi, M., Chen, F., Li, Y., & Yang, Y.** (2022). A comprehensive drought monitoring method integrating multi-source data. *PeerJ*, 10, e13560. doi:10.7717/peerj.13560.
- Shi, X., Zhang, D., Ding, H., Chen, X., & Zhang, J.** (2024). Distribution, trends and drivers of precipitation use efficiency in the loess plateau. *Hydrological Processes*, 38(2). doi:10.1002/hyp.15102.
- Shtiliyanova, A., Bellocchi, G., Borrás, D., Eza, U., Martin, R., & Carrère, P.** (2017). Kriging-based approach to predict missing air temperature data. *Computers and Electronics in Agriculture*, 142(October), 440–449. doi:10.1016/j.compag.2017.09.033.

- Sidibe, M., Dieppois, B., Mahé, G., Paturel, J. E., Amoussou, E., Anifowose, B., & Lawler, D.** (2018). Trend and variability in a new, reconstructed streamflow dataset for West and Central Africa, and climatic interactions, 1950–2005. *Journal of Hydrology*, 561, 478–493. doi:10.1016/j.jhydrol.2018.04.024.
- Silva-Ramírez, E. L., Pino-Mejías, R., & López-Coello, M.** (2015). Single imputation with multilayer perceptron and multiple imputation combining multilayer perceptron and k-nearest neighbours for monotone patterns. *Applied Soft Computing*, 29, 65–74. doi:10.1016/j.asoc.2014.09.052.
- Sims, D. A., Rahman, A. F., Cordova, V. D., El-Masri, B. Z., Baldocchi, D. D., Flanagan, L. B., ... Xu, L.** (2006). On the use of MODIS EVI to assess gross primary productivity of North American ecosystems. *Journal of Geophysical Research: Biogeosciences*, 111(G4). doi:10.1029/2006JG000162.
- Singh, S. K., Das, A. K., Singh, S. R., & Racherla, V.** (2023). Prediction of rail-wheel contact parameters for a metro coach using machine learning. *Expert Systems with Applications*, 215, 119343. doi:10.1016/j.eswa.2022.119343.
- Singh, V. K., Singh, P., Karmakar, M., Leta, J., & Mayr, P.** (2021). The journal coverage of Web of Science, Scopus and Dimensions: A comparative analysis. *Scientometrics*, 126(6), 5113–5142. doi:10.1007/s11192-021-03948-5.
- Song, C., Dannenberg, M. P., & Hwang, T.** (2013). Optical remote sensing of terrestrial ecosystem primary productivity. *Progress in Physical Geography: Earth and Environment*, 37(6), 834–854. doi:10.1177/0309133313507944.
- Souza, U. B. de, Escola, J. P. L., & Brito, L. da C.** (2022). A survey on Hilbert-Huang transform: Evolution, challenges and solutions. *Digital Signal Processing: A Review Journal*, 120, 103292. doi:10.1016/j.dsp.2021.103292.
- Sovilj, D., Eirola, E., Miche, Y., Björk, K. mikael, Nian, R., Akusok, A., & Lendasse, A.** (2016). Extreme learning machine for missing data using multiple imputations. *Neurocomputing*, 174, 220–231. doi:10.1016/j.neucom.2015.03.108.
- Spinoni, J., Naumann, G., & Vogt, J. V.** (2017). Pan-European seasonal trends and recent changes of drought frequency and severity. *Global and Planetary Change*, 148, 113–130. doi:10.1016/j.gloplacha.2016.11.013.
- Stahlschmidt, S., & Stephen, D.** (2020). *Comparison of Web of Science, Scopus and Dimensions databases*. [http://files/15/Stahlschmidt ve Stephen - Comparison of Web of Science, Scopus and Dimension.pdf](http://files/15/Stahlschmidt%20ve%20Stephen%20-%20Comparison%20of%20Web%20of%20Science,%20Scopus%20and%20Dimension.pdf).
- Stephens, G. L., & L'Ecuyer, T.** (2015). The Earth's energy balance. *Atmospheric Research*, 166, 195–203. doi:10.1016/j.atmosres.2015.06.024.

- Storn, R., & Price, K.** (1997). Differential Evolution – A Simple and Efficient Heuristic for global Optimization over Continuous Spaces. *Journal of Global Optimization*, 11(4), 341–359. doi:10.1023/A:1008202821328.
- Stoy, P. C., Roh, J., & Bromley, G. T.** (2022). It's the Heat and the Humidity: The Complementary Roles of Temperature and Specific Humidity to Recent Changes in the Energy Content of the Near-Surface Atmosphere. *Geophysical Research Letters*, 49(4), e2021GL096628. doi:10.1029/2021GL096628.
- Strike, K., Emam, K. El, & Madhavji, N.** (2001). Software cost estimation with incomplete data. *IEEE Transactions on Software Engineering*, 27(10), 890–908. doi:10.1109/32.962560.
- Sulandari, W., Subanar, S., Lee, M. H., & Rodrigues, P. C.** (2020). Time series forecasting using singular spectrum analysis, fuzzy systems and neural networks. *MethodsX*, 7, 101015. doi:10.1016/j.mex.2020.101015.
- Sun, R., Xiao, Z., Wang, J., Zhu, A., & Wang, M.** (2020). *Algorithm of global Gross and net primary productivity products*. <https://fluxnet.fluxdata.org/data/>.
- Suo, N., Xu, C., Cao, L., Song, L., & Lei, X.** (2024). A copula-based parametric composite drought index for drought monitoring and applicability in arid Central Asia. *CATENA*, 235, 107624. doi:10.1016/j.catena.2023.107624.
- Svoboda, M., LeComte, D., Hayes, M., Heim, R., Gleason, K., Angel, J., ... Stephens, S.** (2002). THE DROUGHT MONITOR. *Bulletin of the American Meteorological Society*, 83(8), 1181–1190. doi:10.1175/1520-0477-83.8.1181.
- Tabari, H., & Talaei, P. H.** (2015). Reconstruction of river water quality missing data using artificial neural networks. *Water Quality Research Journal of Canada*, 50(4), 326–335. doi:10.2166/wqrjc.2015.044.
- Tan, F., Yang, J., & Zhou, C.** (2024). Historical review and synthesis of global carbon neutrality research: A bibliometric analysis based on R-tool. *Journal of Cleaner Production*, 449, 141574. doi:10.1016/j.jclepro.2024.141574.
- Tang, A. C. I., Stoy, P. C., Hirata, R., Musin, K. K., Aeries, E. B., Wenceslaus, J., ... Melling, L.** (2019). The exchange of water and energy between a tropical peat forest and the atmosphere: Seasonal trends and comparison against other tropical rainforests. *Science of the Total Environment*, 683, 166–174. doi:10.1016/j.scitotenv.2019.05.217.
- Tang, X., Feng, Y., Gao, C., Lei, Z., Chen, S., Wang, R., ... Tong, X.** (2023). Entropy-weight-based spatiotemporal drought assessment using MODIS products and Sentinel-1A images in Urumqi, China. *Natural Hazards*, 119(1), 387–408. doi:10.1007/s11069-023-06131-6.
- Tang, X., Li, H., Huang, N., Li, X., Xu, X., Ding, Z., & Xie, J.** (2015). A comprehensive assessment of MODIS-derived GPP for forest ecosystems using the site-level FLUXNET database. *Environmental Earth Sciences*, 74(7), 5907–5918. doi:10.1007/s12665-015-4615-0.

- Tao, H., Ebtehaj, I., Bonakdari, H., Heddami, S., Voyant, C., Al-Ansari, N., Deo, R., & Yaseen, Z. M.** (2019). Designing a New Data Intelligence Model for Global Solar Radiation Prediction: Application of Multivariate Modeling Scheme. *Energies*, 12(7), 1365. doi:10.3390/en12071365.
- Tayanç, M., İm, U., Doğruel, M., & Karaca, M.** (2009). Climate change in Turkey for the last half century. *Climatic Change*, 94(3–4), 483–502. doi:10.1007/s10584-008-9511-0.
- Taylor, K. E.** (2001). Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research: Atmospheres*, 106(D7), 7183–7192. doi:10.1029/2000JD900719.
- Team, R. core.** (2023). *R: A Language and Environment for Statistical Computing* (<https://www.R-project.org/>).
- Thevakaran, A., & Sonnadara, D. U. J.** (2018). Estimating missing daily temperature extremes in Jaffna, Sri Lanka. *Theoretical and Applied Climatology*, 132(1–2), 145–152. doi:10.1007/s00704-017-2082-0.
- Tian, F., & Zhang, Y.** (2020). Spatiotemporal patterns of evapotranspiration, gross primary productivity, and water use efficiency of cropland in agroecosystems and their relation to the water-saving project in the Shiyang River Basin of Northwestern China. *Computers and Electronics in Agriculture*, 172, 105379. doi:10.1016/j.compag.2020.105379.
- Tramontana, G., Jung, M., Schwalm, C. R., Ichii, K., Camps-Valls, G., Ráduly, B., ... Papale, D.** (2016). Predicting carbon dioxide and energy fluxes across global FLUXNET sites with regression algorithms. *Biogeosciences*, 13(14), 4291–4313. doi:10.5194/bg-13-4291-2016.
- Tu, X., Wu, H., Singh, V. P., Chen, X., Lin, K., & Xie, Y.** (2018). Multivariate design of socioeconomic drought and impact of water reservoirs. *Journal of Hydrology*, 566, 192–204. doi:10.1016/j.jhydrol.2018.09.012.
- Tucker, C., Yan, D., Dannenberg, M., Reed, S. C., & Smith, W.** (2018). Science at the frontier: multimethod research to evaluate ecosystem change across multiple scales. *New Phytologist*, 218(4), 1318–1320. doi:10.1111/nph.15195.
- Turkes, M., Turp, M. T., An, N., Ozturk, T., & Kurnaz, M. L.** (2020). *Impacts of Climate Change on Precipitation Climatology and Variability in Turkey* (N. B. Harmancioglu & D. Altinbilek (eds.); pp. 467–491). Springer International Publishing. doi:10.1007/978-3-030-11729-0_14.
- Türkeş, M., Yozgatlıgil, C., Batmaz, İ., İyigün, C., Kartal Koç, E., Fahmi, F., & Aslan, S.** (2016). Has the climate been changing in Turkey? Regional climate change signals based on a comparative statistical analysis of two consecutive time periods, 1950–1980 and 1981–2010. *Climate Research*, 70(1), 77–93. doi:10.3354/cr01410.

- Turner, D. P., Ritts, W. D., Cohen, W. B., Gower, S. T., Running, S. W., Zhao, M., ... Ahl, D. E.** (2006). Evaluation of MODIS NPP and GPP products across multiple biomes. *Remote Sensing of Environment*, 102(3–4), 282–292. doi:10.1016/j.rse.2006.02.017.
- Turner, D. P., Ritts, W. D., Cohen, W. B., Gower, S. T., Zhao, M., Running, S. W., ... Munger, J. W.** (2003). Scaling Gross Primary Production (GPP) over boreal and deciduous forest landscapes in support of MODIS GPP product validation. *Remote Sensing of Environment*, 88(3), 256–270. doi:https://doi.org/10.1016/j.rse.2003.06.005.
- Turrado, C. C., López, M. del C. M., Lasheras, F. S., Gómez, B. A. R., Rollé, J. L. C., & Juez, F. J. de C.** (2014). Missing data imputation of solar radiation data under different atmospheric conditions. *Sensors (Switzerland)*, 14(11), 20382–20399. doi:10.3390/s141120382.
- Tuvdendorj, B., Wu, B., Zeng, H., Batdelger, G., & Nanzad, L.** (2019). Determination of Appropriate Remote Sensing Indices for Spring Wheat Yield Estimation in Mongolia. *Remote Sensing*, 11(21), 2568. doi:10.3390/rs11212568.
- van der Schrier, G., Barichivich, J., Briffa, K. R., & Jones, P. D.** (2013). A scPDSI-based global data set of dry and wet spells for 1901–2009. *Journal of Geophysical Research: Atmospheres*, 118(10), 4025–4048. doi:10.1002/jgrd.50355.
- van Dijk, A. I. J. M., Dolman, A. J., & Schulze, E.** (2005). Radiation, temperature, and leaf area explain ecosystem carbon fluxes in boreal and temperate European forests. *Global Biogeochemical Cycles*, 19(2). doi:10.1029/2004GB002417.
- van Eck, N. J., & Waltman, L.** (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, 84(2), 523–538. doi:10.1007/s11192-009-0146-3.
- van Gorsel, E., Delpierre, N., Leuning, R., Black, A., Munger, J. W., Wofsy, S., ... Wharton, S.** (2009). Estimating nocturnal ecosystem respiration from the vertical turbulent flux and change in storage of CO₂. *Agricultural and Forest Meteorology*, 149(11), 1919–1930. doi:10.1016/j.agrformet.2009.06.020.
- Van Gorsel, E., Leuning, R., Cleugh, H. A., Keith, H., & Suni, T.** (2007). Nocturnal carbon efflux: reconciliation of eddy covariance and chamber measurements using an alternative to the u*-threshold filtering technique. *Tellus B: Chemical and Physical Meteorology*, 59(3), 397. doi:10.1111/j.1600-0889.2007.00252.x.
- Vernay, A., Hasselquist, N., Leppä, K., Klosterhalfen, A., Gutierrez Lopez, J., Stangl, Z. R., ... Marshall, J. D.** (2024). Partitioning gross primary production of a boreal forest among species and strata: A multi-method approach. *Agricultural and Forest Meteorology*, 345, 109857. doi:10.1016/j.agrformet.2023.109857.

- Vicente-Serrano, S. M., Beguería, S., & López-Moreno, J. I.** (2010). A Multiscalar Drought Index Sensitive to Global Warming: The Standardized Precipitation Evapotranspiration Index. *Journal of Climate*, 23(7), 1696–1718. doi:10.1175/2009JCLI2909.1.
- Vignola, F., Derocher, Z., Peterson, J., Vuilleumier, L., Félix, C., Gröbner, J., & Kouremeti, N.** (2016). Effects of changing spectral radiation distribution on the performance of photodiode pyranometers. *Solar Energy*, 129, 224–235. doi:10.1016/j.solener.2016.01.047.
- Walker, A. P., De Kauwe, M. G., Bastos, A., Belmecheri, S., Georgiou, K., Keeling, R. F., ... Zuidema, P. A.** (2021). Integrating the evidence for a terrestrial carbon sink caused by increasing atmospheric CO₂. *New Phytologist*, 229(5), 2413–2445. doi:10.1111/nph.16866.
- Wallace, M. L., Mentch, L., Wheeler, B. J., Tapia, A. L., Richards, M., Zhou, S., ... Buysse, D. J.** (2023). Use and misuse of random forest variable importance metrics in medicine: demonstrations through incident stroke prediction. *BMC Medical Research Methodology*, 23(1), 144. doi:10.1186/s12874-023-01965-x.
- Wang, C., Deng, L., Zhang, Y., Zhao, M., Liang, M., Lee, L. C., ... He, T.** (2024). Farmland phytoremediation in bibliometric analysis. *Journal of Environmental Management*, 351, 119971. doi:10.1016/j.jenvman.2023.119971.
- Wang, F., Shen, Y., Chen, Q., & Wang, W.** (2021). Bridging the gap between GRACE and GRACE follow-on monthly gravity field solutions using improved multichannel singular spectrum analysis. *Journal of Hydrology*, 594, 125972. doi:10.1016/j.jhydrol.2021.125972.
- Wang, H., Li, X., Ma, M., & Geng, L.** (2019). Improving Estimation of Gross Primary Production in Dryland Ecosystems by a Model-Data Fusion Approach. *Remote Sensing*, 11(3), 225. doi:10.3390/rs11030225.
- Wang, H., Yue, W., Wen, S., Xu, X., Haasis, H. D., Su, M., ... & Du, P.** (2022). An improved bearing fault detection strategy based on artificial bee colony algorithm. *CAAI Transactions on Intelligence Technology*, February. doi:10.1049/cit2.12105.
- Wang, H., Zhu, Q., Wang, Y., & Zhang, H.** (2023). Spatio-temporal characteristics and driving factors of flash drought recovery: From the perspective of soil moisture and GPP changes. *Weather and Climate Extremes*, 42, 100605. doi:10.1016/j.wace.2023.100605.
- Wang, J., Sun, R., Zhang, H., Xiao, Z., Zhu, A., Wang, M., Yu, T., & Xiang, K.** (2021). New Global MuSyQ GPP/NPP Remote Sensing Products From 1981 to 2018. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14, 5596–5612. doi:10.1109/JSTARS.2021.3076075.
- Wang, L., & Shi, J.** (2021). A Comprehensive Application of Machine Learning Techniques for Short-Term Solar Radiation Prediction. *Applied Sciences*, 11(13), 5808. doi:10.3390/app11135808.

- Wang, Q., Moreno-Martínez, Á., Muñoz-Marí, J., Campos-Taberner, M., & Camps-Valls, G.** (2023). Estimation of vegetation traits with kernel NDVI. *ISPRS Journal of Photogrammetry and Remote Sensing*, 195, 408–417. doi:<https://doi.org/10.1016/j.isprsjprs.2022.12.019>.
- Wang, S., Lian, J., Peng, Y., Hu, B., & Chen, H.** (2019). Generalized reference evapotranspiration models with limited climatic data based on random forest and gene expression programming in Guangxi, China. *Agricultural Water Management*, 221, 220–230. doi:[10.1016/j.agwat.2019.03.027](https://doi.org/10.1016/j.agwat.2019.03.027).
- Wang, Y., Liu, S., Huang, S., Zhou, Z., & Shi, H.** (2023). Bivariate assessment of socioeconomic drought events based on an improved socioeconomic drought index. *Journal of Hydrology*, 623, 129878. doi:[10.1016/j.jhydrol.2023.129878](https://doi.org/10.1016/j.jhydrol.2023.129878).
- Wang, Z., Wang, Z., Xiong, J., He, W., Yong, Z., & Wang, X.** (2022). Responses of the Remote Sensing Drought Index with Soil Information to Meteorological and Agricultural Droughts in Southeastern Tibet. *Remote Sensing*, 14(23), 6125. doi:[10.3390/rs14236125](https://doi.org/10.3390/rs14236125).
- Waseem, M., Ajmal, M., & Kim, T. W.** (2015). Development of a new composite drought index for multivariate drought assessment. *Journal of Hydrology*, 527, 30–37. doi:[10.1016/j.jhydrol.2015.04.044](https://doi.org/10.1016/j.jhydrol.2015.04.044).
- Wehr, R., & Saleska, S. R.** (2015). An improved isotopic method for partitioning net ecosystem–atmosphere CO₂ exchange. *Agricultural and Forest Meteorology*, 214–215, 515–531. doi:[10.1016/j.agrformet.2015.09.009](https://doi.org/10.1016/j.agrformet.2015.09.009).
- Wei, N., Yin, L., Li, C., Wang, W., Qiao, W., Li, C., ... Fu, L.** (2022). Short-term load forecasting using detrend singular spectrum fluctuation analysis. *Energy*, 256, 124722. doi:[10.1016/j.energy.2022.124722](https://doi.org/10.1016/j.energy.2022.124722).
- Wei, X., He, W., Zhou, Y., Cheng, N., Xiao, J., Bi, W., Liu, Y., Sun, S., & Ju, W.** (2023). Increased Sensitivity of Global Vegetation Productivity to Drought Over the Recent Three Decades. *Journal of Geophysical Research: Atmospheres*, 128(7). doi:[10.1029/2022JD037504](https://doi.org/10.1029/2022JD037504).
- Wells, N., Goddard, S., & Hayes, M. J.** (2004). A Self-Calibrating Palmer Drought Severity Index. *Journal of Climate*, 17(12), 2335–2351. doi:[10.1175/1520-0442\(2004\)017<2335:ASPDSI>2.0.CO;2](https://doi.org/10.1175/1520-0442(2004)017<2335:ASPDSI>2.0.CO;2).
- Wesonga, R.** (2015). On multivariate imputation and forecasting of decadal wind speed missing data. *SpringerPlus*, 4(1). doi:[10.1186/s40064-014-0774-9](https://doi.org/10.1186/s40064-014-0774-9).
- Williams, D. A., Nelsen, B., Berrett, C., Williams, G. P., & Moon, T. K.** (2018). A comparison of data imputation methods using Bayesian compressive sensing and Empirical Mode Decomposition for environmental temperature data. *Environmental Modelling and Software*, 102, 172–184. doi:[10.1016/j.envsoft.2018.01.012](https://doi.org/10.1016/j.envsoft.2018.01.012).

- Williams, D. A., Nelsen, B., Berrett, C., Williams, G. P., & Moon, T. K.** (2018). A comparison of data imputation methods using Bayesian compressive sensing and Empirical Mode Decomposition for environmental temperature data. *Environmental Modelling & Software*, 102, 172–184. doi:10.1016/j.envsoft.2018.01.012.
- Wittstadt, U., Földner, G., Laurenz, E., Warlo, A., Große, A., Herrmann, R., ... Mittelbach, W.** (2017). A novel adsorption module with fiber heat exchangers: Performance analysis based on driving temperature differences. *Renewable Energy*, 110, 154–161. doi:10.1016/j.renene.2016.08.061.
- Wu, X., Xu, H., He, H., Wu, Z., Lu, G., & Liao, T.** (2024). Agricultural Drought Monitoring Using an Enhanced Soil Water Deficit Index Derived from Remote Sensing and Model Data Merging. *Remote Sensing*, 16(12), 2156. doi:10.3390/rs16122156.
- Xiao, J.** (2014). Satellite evidence for significant biophysical consequences of the “Grain for Green” Program on the Loess Plateau in China. *Journal of Geophysical Research: Biogeosciences*, 119(12), 2261–2275. doi:10.1002/2014JG002820.
- Xiao, J., Chevallier, F., Gomez, C., Guanter, L., Hicke, J. A., Huete, A. R., ... Zhang, X.** (2019). Remote sensing of the terrestrial carbon cycle: A review of advances over 50 years. *Remote Sensing of Environment*, 233, 111383. doi:https://doi.org/10.1016/j.rse.2019.111383.
- Xiao, J., & Moody, A.** (2004). Photosynthetic activity of US biomes: responses to the spatial variability and seasonality of precipitation and temperature. *Global Change Biology*, 10(4), 437–451. doi:10.1111/j.1365-2486.2004.00745.x.
- Xiao, J., Ollinger, S. V., Frolking, S., Hurtt, G. C., Hollinger, D. Y., Davis, K. J., Pan, Y., ... Suyker, A. E.** (2014). Data-driven diagnostics of terrestrial carbon dynamics over North America. *Agricultural and Forest Meteorology*, 197, 142–157. doi:10.1016/j.agrformet.2014.06.013.
- Xie, X., Li, A., Tian, J., Wu, C., & Jin, H.** (2023). A fine spatial resolution estimation scheme for large-scale gross primary productivity (GPP) in mountain ecosystems by integrating an eco-hydrological model with the combination of linear and non-linear downscaling processes. *Journal of Hydrology*, 616(9). doi:10.1016/j.jhydrol.2022.128833.
- Xie, Z. X., & Sun, X. F.** (2018). Imputation of missing wind speed data based on low-rank matrix approximation. *2017 2nd International Conference on Power and Renewable Energy, ICPRE 2017*, 397–401. doi:10.1109/ICPRE.2017.8390566.
- Xing, G., Zhao, E. L., Zhang, C., & Wu, J.** (2021). A Decomposition-Ensemble Approach with Denoising Strategy for PM_{2.5} Concentration Forecasting. *Discrete Dynamics in Nature and Society*, 2021. doi:10.1155/2021/5577041.

- Xiong, Y., Chen, X., Tang, L., & Wang, H.** (2022). Comparison of surface renewal and Bowen ratio derived evapotranspiration measurements in an arid vineyard. *Journal of Hydrology*, 613, 128474. doi:10.1016/j.jhydrol.2022.128474.
- Xu, C. D., Wang, J. F., Hu, M. G., & Li, Q. X.** (2013). Interpolation of missing temperature data at meteorological stations using P-BSHADE. *Journal of Climate*, 26(19), 7452–7463. doi:10.1175/JCLI-D-12-00633.1.
- Xu, Q., Xiong, Y., Dai, H., Kumari, K. M., Xu, Q., Ou, H. Y., & Wei, D. Q.** (2017). PDC-SGB: Prediction of effective drug combinations using a stochastic gradient boosting algorithm. *Journal of Theoretical Biology*, 417, 1–7. doi:10.1016/j.jtbi.2017.01.019.
- Xu, X., Gou, X., Zhang, W., Zhao, Y., & Xu, Z.** (2023). A bibliometric analysis of carbon neutrality: Research hotspots and future directions. *Heliyon*, 9(8), e18763. doi:10.1016/j.heliyon.2023.e18763.
- Yan, H., He, Z., Gao, C., Xie, M., Sheng, H., & Chen, H.** (2022). Investment estimation of prefabricated concrete buildings based on XGBoost machine learning algorithm. *Advanced Engineering Informatics*, 54(April), 101789. doi:10.1016/j.aei.2022.101789.
- Yang, P., Zhai, X., Huang, H., Zhang, Y., Zhu, Y., Shi, X., ... Fu, C.** (2023). Association and driving factors of meteorological drought and agricultural drought in Ningxia, Northwest China. *Atmospheric Research*, 289, 106753. doi:10.1016/j.atmosres.2023.106753.
- Yang, R., Wang, J., Zeng, N., Sitch, S., Tang, W., McGrath, M. J., ... Han, P.** (2022). Divergent historical GPP trends among state-of-the-art multi-model simulations and satellite-based products. *Earth System Dynamics*, 13(2), 833–849. doi:10.5194/esd-13-833-2022.
- Yang, W., Peng, Z., Wei, K., Shi, P., & Tian, W.** (2017). Superiorities of variational mode decomposition over empirical mode decomposition particularly in time–frequency feature extraction and wind turbine condition monitoring. *IET Renewable Power Generation*, 11(4), 443–452. doi:10.1049/iet-rpg.2016.0088.
- Yang, Z., Liu, Y., & Li, C.** (2011). Interpolation of missing wind data based on ANFIS. *Renewable Energy*, 36(3), 993–998. doi:10.1016/j.renene.2010.08.033.
- Yang, Z., Xue, F., & Lu, W.** (2021). Handling missing data for construction waste management: machine learning based on aggregated waste generation behaviors. *Resources, Conservation and Recycling*, 175, 105809. doi:10.1016/j.resconrec.2021.105809.
- Yao, T., Shi, Y., & Chen, M.** (2023). Review and research frontier analysis of low carbon economy – Based on bibliometric methods. *Procedia Computer Science*, 221, 1292–1301. doi:10.1016/j.procs.2023.08.118.
- Yin, G., & Zhang, H.** (2023). A new integrated index for drought stress monitoring based on decomposed vegetation response factors. *Journal of Hydrology*, 618, 129252. doi:10.1016/j.jhydrol.2023.129252.

- Yozgatligil, C., Aslan, S., Iyigun, C., & Batmaz, I.** (2013). Comparison of missing value imputation methods in time series: The case of Turkish meteorological data. *Theoretical and Applied Climatology*, 112(1–2), 143–167. doi:10.1007/s00704-012-0723-x.
- Yu-E, M., Wen-Hua, X., & Pi-Feng, L.** (2007). Stem Respiration And Its Controlling Factors In Forest Ecosystems. *Chinese Journal of Plant Ecology*, 31(3), 403–412. doi:10.17521/cjpe.2007.0049.
- Yu, P., Zhang, Y., Liu, P., Zhang, J., Xing, W., Tong, X., ... Meng, P.** (2024). Regulation of biophysical drivers on carbon and water fluxes over a warm-temperate plantation in northern China. *Science of the Total Environment*, 907, 167408. doi:10.1016/j.scitotenv.2023.167408.
- Zainudin, M. L., Saaban, A., & Bakar, M. N. A.** (2015). Estimation of missing values in solar radiation data using piecewise interpolation methods: Case study at Penang city. *AIP Conference Proceedings*, 1691(December). doi:10.1063/1.4937079.
- Zeybekoğlu, U., & Keskin, A. Ü.** (2020). Defining rainfall intensity clusters in Turkey by using the fuzzy c-means algorithm. *Geofizika*, 37(2), 181–195. doi:10.15233/gfz.2020.37.8.
- Zeynolabedin, A., Ghiassi, R., Norooz, R., Najib, S., & Fadili, A.** (2021). Evaluation of geoelectrical models efficiency for coastal seawater intrusion by applying uncertainty analysis. *Journal of Hydrology*, 603(PC), 127086. doi:10.1016/j.jhydrol.2021.127086.
- Zhang, A., & Jia, G.** (2013). Monitoring meteorological drought in semiarid regions using multi-sensor microwave remote sensing data. *Remote Sensing of Environment*, 134, 12–23. doi:10.1016/j.rse.2013.02.023.
- Zhang, G., Band, S. S., Jun, C., Bateni, S. M., Chuang, H. M., Turabieh, H., ... Moslehpour, M.** (2021). Solar radiation estimation in different climates with meteorological variables using Bayesian model averaging and new soft computing models. *Energy Reports*, 7, 8973–8996. doi:10.1016/j.egyr.2021.10.117.
- Zhang, K., Kimball, J. S., Mu, Q., Jones, L. A., Goetz, S. J., & Running, S. W.** (2009). Satellite based analysis of northern ET trends and associated changes in the regional water balance from 1983 to 2005. *Journal of Hydrology*, 379(1–2), 92–110. doi:10.1016/j.jhydrol.2009.09.047.
- Zhang, L., Bing, Z., & Zhang, L.** (2015). A hybrid clustering algorithm based on missing attribute interval estimation for incomplete data. *Pattern Analysis and Applications*, 18(2), 377–384. doi:10.1007/s10044-014-0376-8.
- Zhang, L., Jiao, W., Zhang, H., Huang, C., & Tong, Q.** (2017). Studying drought phenomena in the Continental United States in 2011 and 2012 using various drought indices. *Remote Sensing of Environment*, 190, 96–106. doi:10.1016/j.rse.2016.12.010.

- Zhang, M., Fernández-Torres, M. Á., & Camps-Valls, G.** (2024). Domain knowledge-driven variational recurrent networks for drought monitoring. *Remote Sensing of Environment*, 311, 114252. doi:10.1016/j.rse.2024.114252.
- Zhang, R., E, X., Ma, Z., An, Y., Bao, Q., Wu, Z., Wu, L., & Sun, Z.** (2024). Drought Sensitivity and Vulnerability of Rubber Plantation GPP—Insights from Flux Site-Based Simulation. *Land*, 13(6), 745. doi:10.3390/land13060745.
- Zhang, W., Luo, G., Hamdi, R., Ma, X., Li, Y., Yuan, X., ... De Maeyer, P.** (2023). Can Gross Primary Productivity Products be effectively evaluated in regions with few observation data? *GIScience and Remote Sensing*, 60(1). doi:10.1080/15481603.2023.2213489.
- Zhang, X., Zhang, Y., & Kong, D.** (2023). *Global monthly GPP, ET, Ec, Es and Ei simulated by PML-V2 with AVHRR data at a 0.05 degree resolution over 1982-2014*. doi:10.6084/m9.figshare.14185739.v6.
- Zhang, X., Zhang, Y., Ma, N., Kong, D., Tian, J., Shao, X., & Tang, Q.** (2021). Greening-induced increase in evapotranspiration over Eurasia offset by CO₂-induced vegetational stomatal closure. *Environmental Research Letters*, 16(12). doi:10.1088/1748-9326/ac3532.
- Zhang, Y., Bu, J., Zuo, X., Yu, K., Wang, Q., & Huang, W.** (2024). Vegetation Water Content Retrieval from Spaceborne GNSS-R and Multi-Source Remote Sensing Data Using Ensemble Machine Learning Methods. *Remote Sensing*, 16(15), 2793. doi:10.3390/rs16152793.
- Zhang, Y., & Thorburn, P. J.** (2022). Handling missing data in near real-time environmental monitoring: A system and a review of selected methods. *Future Generation Computer Systems*, 128, 63–72. doi:10.1016/j.future.2021.09.033.
- Zhang, Y., Wang, P., Chen, Y., Yang, J., Wu, D., Ma, Y., Huo, Z., & Liu, S.** (2023). Daily dynamic thresholds of different agricultural drought grades for summer maize based on the Vegetation Water Index. *Journal of Hydrology*, 625, 130070. doi:10.1016/j.jhydrol.2023.130070.
- Zhang, Y., & Ye, A.** (2021). Would the obtainable gross primary productivity (GPP) products stand up? A critical assessment of 45 global GPP products. *Science of The Total Environment*, 783, 146965. doi:10.1016/j.scitotenv.2021.146965.
- Zhang, Y., & Ye, A.** (2022). Uncertainty analysis of multiple terrestrial gross primary productivity products. *Global Ecology and Biogeography*, 31(11), 2204–2218. doi:10.1111/geb.13578.
- Zhao, F., Lu, Y., Li, X., Wang, L., Song, Y., Fan, D., ... Chen, X.** (2022). Multiple imputation method of missing credit risk assessment data based on generative adversarial networks. *Applied Soft Computing*, 126, 109273. doi:10.1016/j.asoc.2022.109273.
- Zhao, Q., Zhang, X., Li, C., Xu, Y., & Fei, J.** (2024). Compound ecological drought assessment of China using a Copula-based drought index. *Ecological Indicators*, 164, 112141. doi:10.1016/j.ecolind.2024.112141.

- Zhao, S., Li, C., Tian, M., Zhang, H., Xiao, Z., & Hu, R.** (2022). Large-Scale Scheduling Model Based on Improved Ant Colony Algorithm. *Mobile Information Systems*, 2022. doi:10.1155/2022/9116121.
- Zheng, Z., Ali, M., Jamei, M., Xiang, Y., Karbasi, M., Yaseen, Z. M., & Farooque, A. A.** (2023). Design data decomposition-based reference evapotranspiration forecasting model: A soft feature filter based deep learning driven approach. *Engineering Applications of Artificial Intelligence*, 121, 105984. doi:10.1016/j.engappai.2023.105984.
- Zhong, C., Li, T., Bi, R., Sanganyado, E., Huang, J., Jiang, S., Zhang, Z., & Du, H.** (2023). A systematic overview, trends and global perspectives on blue carbon: A bibliometric study (2003–2021). *Ecological Indicators*, 148, 110063. doi:10.1016/j.ecolind.2023.110063.
- Zhou, L., Tucker, C. J., Kaufmann, R. K., Slayback, D., Shabanov, N. V., & Myneni, R. B.** (2001). Variations in northern vegetation activity inferred from satellite data of vegetation index during 1981 to 1999. *Journal of Geophysical Research: Atmospheres*, 106(D17), 20069–20083. doi:10.1029/2000JD000115.
- Zhou, X., Bi, S., Yang, Y., Tian, F., & Ren, D.** (2014). Comparison of ET estimations by the three-temperature model, SEBAL model and eddy covariance observations. *Journal Of Hydrology*, 519, 769–776. doi:10.1016/j.jhydrol.2014.08.004.
- Zhu, W., Xie, Z., Zhao, C., Zheng, Z., Qiao, K., Peng, D., & Fu, Y. H.** (2024). Remote sensing of terrestrial gross primary productivity: a review of advances in theoretical foundation, key parameters and methods. *GIScience & Remote Sensing*, 61(1). doi:10.1080/15481603.2024.2318846.
- Zhu, X. J., Yu, G. R., Chen, Z., Zhang, W. K., Han, L., Wang, Q. F., ... Zhu, Z. L.** (2023). Mapping Chinese annual gross primary productivity with eddy covariance measurements and machine learning. *Science of The Total Environment*, 857, 159390. doi:10.1016/j.scitotenv.2022.159390.
- Zhu, Z., Piao, S., Myneni, R. B., Huang, M., Zeng, Z., Canadell, J. G., ... Zeng, N.** (2016). Greening of the Earth and its drivers. *Nature Climate Change*, 6(8), 791–795. doi:10.1038/nclimate3004.
- Zipper, S. C., Qiu, J., & Kucharik, C. J.** (2016). Drought effects on US maize and soybean production: spatiotemporal patterns and historical changes. *Environmental Research Letters*, 11(9), 094021. doi:10.1088/1748-9326/11/9/094021.
- Zitouna-Chebbi, R., Prévot, L., Chakhar, A., Marniche-Ben Abdallah, M., & Jacob, F.** (2018). Observing Actual Evapotranspiration from Flux Tower Eddy Covariance Measurements within a Hilly Watershed: Case Study of the Kamech Site, Cap Bon Peninsula, Tunisia. *Atmosphere*, 9(2), 68. doi:10.3390/atmos9020068.
- Žliobaitė, I., Hollmén, J., & Junninen, H.** (2014). Regression models tolerant to massively missing data: a case study in solar-radiation nowcasting. *Atmospheric Measurement Techniques*, 7(12), 4387–4399. doi:10.5194/amt-7-4387-2014.

Zouzou, Y., & Citakoglu, H. (2022). General and regional cross-station assessment of machine learning models for estimating reference evapotranspiration. *Acta Geophysica*, 71(2), 927–947. doi:10.1007/s11600-022-00939-9.



APPENDICES

APPENDIX A: Additional explanations, figures, and tables for Chapter 6.

APPENDIX B: Additional figures and tables for Chapter 7.



APPENDIX A : Additional explanations, figures, and tables for Chapter 6.

MODIS GPP

Practically, the MODIS GPP algorithm calculates GPP by multiplying APAR (Absorbed Photosynthetically Active Radiation) with the maximum light use efficiency (ϵ_{max}) and adjusting for environmental stress factors represented by temperature and water scalars (T_s , W_s). This formula offers a method for estimating the rate at which an ecosystem absorbs carbon dioxide during photosynthesis, taking into account both the available light energy and environmental conditions affecting plant growth (Huang et al, 2021). The details of the MODIS GPP calculation are given in the equations below.

$$GPP = APAR \times \epsilon_{MAX} \times T_s \times W_s \quad (A.1)$$

$$APAR = PAR \times FPAR \quad (A.2)$$

$$PAR = S \times 0.45 \quad (A.3)$$

ϵ_{MAX} is the maximum Light Use Efficiency (LUE) that varies by plant functional type; PAR stands for Photosynthetically Active Radiation incident, comprising about 45% of incoming shortwave solar radiation; FPAR represents the fraction of PAR absorbed by vegetation; T_s is a temperature stress scalar, and W_s is a water stress scalar. The T_s scalar quantifies the impact of temperature on the efficiency of photosynthesis. It is calculated based on the current temperature conditions relative to an optimal temperature range for photosynthesis. The temperature stress scalar decreases when the temperature moves away from this optimal range, indicating a reduction in photosynthetic efficiency due to temperature stress. The W_s scalar measures the impact of water availability on photosynthetic activity. It is determined based on the moisture content in the soil and/or atmospheric humidity levels relative to the water needs of the vegetation. The water stress scalar decreases with decreasing water availability, reflecting a reduction in photosynthetic efficiency due to water stress. These two scalars are calculated as follows Eq. 4 and 5 (Running et al, 2004).

$$T_s = \begin{cases} 0 & T_{min} < T_{min_{min}} \\ \frac{(T_{min} - T_{min_{min}})}{T_{min_{max}} - T_{min_{min}}} & T_{min_{min}} < T_{min} < T_{min_{max}} \\ 1 & T_{min} > T_{min_{max}} \end{cases} \quad (A.4)$$

Where $T_{min_{min}}$ is the daily minimum temperature when LUE equals zero: $T_{min_{max}}$ is the daily minimum temperature when LUE is at maxima.

$$W_s = \begin{cases} 1 & VPD < VPD_{min} \\ \frac{(VPD_{max} - VPD)}{VPD_{max} - VPD_{min}} & VPD_{min} < VPD < VPD_{max} \\ 0 & VPD > VPD_{max} \end{cases} \quad (A.5)$$

In this context, VPD_{min} represents the minimum Vapor Pressure Deficit (VPD) when the Light Use Efficiency (LUE) is at its maximum; VPD_{max} denotes the maximum VPD where LUE equals zero. BPLUT includes the ϵ_{max} and temperature parameters, as well as the VPD scalars for each Plant Functional Type (PFT). The temperature scalar parameters ($T_{min_{min}}$, $T_{min_{max}}$) and the VPD scalar (VPD_{min} and VPD_{max}) are used to calculate the scalars that weaken ϵ_{MAX} in order to produce the actual light use efficiency (ϵ) for estimating GPP. The weakening scalars (temperature and VPD scalars) are simple linear ramp functions of the daily $T_{min_{min}}$, $T_{min_{max}}$, VPD_{min} , and VPD_{max} , and these parameters are derived from the parameterization of parts of the Biome-BGC algorithm (Running et al, 1999). The flowchart for obtaining MODIS GPP values is given in Figure A1.

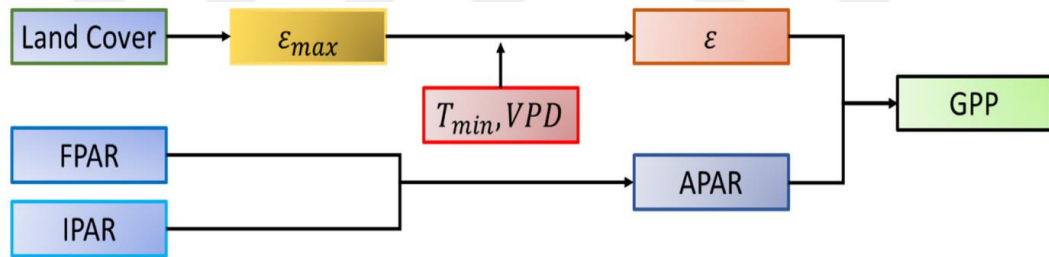


Figure A.1 : MODIS GPP algorithm flow chart.

MuSyQ GPP

The stages for obtaining GPP values using the MuSyQ algorithm, which employs a LUE method to model GPP using remote sensing data with varying spatial and temporal resolutions, along with meteorological data, are outlined below.

$$GPP = LUE \times FPAR \times PAR \quad (A.6)$$

In Eq. 6, LUE represents the actual Light Use Efficiency ($g\ C / MJ$), $FPAR$ is the fraction of photosynthetically active radiation absorbed by vegetation, and PAR is the incoming photosynthetically active radiation ($MJ/m^2/day$). LUE is calculated using a

parameterization approach where the maximum LUE without stress is determined based on vegetation type and the clearness index (CI), and the actual LUE is estimated by multiplying it with the effects of temperature and water stress. CI has been adopted to reflect the impact of the fraction of diffuse light in incoming solar radiation on LUE .

$$LUE = [LUE_{max}^{su} \times CI + LUE_{max}^{sh} \times (1 - CI)] \times f(W) \times f(T) \quad (A.7)$$

Maximum LUE (LUE_{max}) is represented by a weighted calculation of LUE_{max} for sunlit leaves (LUE_{maxsu}) and LUE_{max} for shaded leaves (LUE_{maxsh}). LUE_{maxsu} and LUE_{maxsh} have been optimized using the Shuffled Complex Evolution (SCE-UA) method with data taken from the FLUXNET2015 dataset. CI and $(1-CI)$ are the weighting coefficients. CI represents the clearness index, which is the ratio of solar radiation at the Earth's surface to the extraterrestrial radiation at the top of the atmosphere.

$$CI = \frac{SW_{surface}}{SW_{top}} \quad (A.8)$$

$$SW_{top} = \frac{TD}{\pi} S_0 (w_0 \sin \varphi \sin \delta + \cos \varphi \cos \delta \sin w_0) \quad (A.9)$$

$$S_0 = \bar{S}_0 \times (1 + 0.033 \times \cos(2\pi \times \frac{day}{365}))^2 \quad (A.10)$$

$SW_{surface}$ is from the ERA-Interim dataset. TD is 86400 seconds. $\bar{S}_0$ is the solar constant, equivalent to 1367 W/m². w_0 represents the Solar Horizon at sunrise. φ is the latitude. δ is the solar declination.

$f(T)$ is a temperature stress factor. The growth performance of vegetation is influenced by the average temperature (T , °C) (Potter et al, 1993).

$$f(T) = \frac{1}{[1 + e^{0.5(T_{opt}-10-T)}] \times [1 + e^{0.3(-T_{opt}-10+T)}]} \quad (A.11)$$

T_{opt} (°C) is the optimum growth temperature, which refers to the average temperature at which vegetation grows best.

The water stress factor, denoted as $f(W)$, is calculated using Eq. A.12.

$$f(W) = 0.5 + 0.5 \left(\frac{E}{E_p} \right) \quad (A.12)$$

In this context, E represents actual evapotranspiration (mm), calculated using a modified Penman-Monteith approach utilizing GLASS LAI products (Zhang et al, 2009). E_p signifies potential evapotranspiration (mm) derived from the Priestley and Taylor equation (Priestley and Taylor, 1972). The water stress factor, $f(W)$, is constrained to equal 1.0 when it exceeds 1.0.

FPAR is sourced from the GLASS FPAR product. Photosynthetically Active Radiation (PAR), measured in MJ/m²/day, is calculated from $SW_{surface}$ using a Eq. A.13.

$$PAR = 0.48 \times SW_{surface} \quad (A.13)$$

The flowchart of the MuSyQ algorithm is provided in Figure A.2.

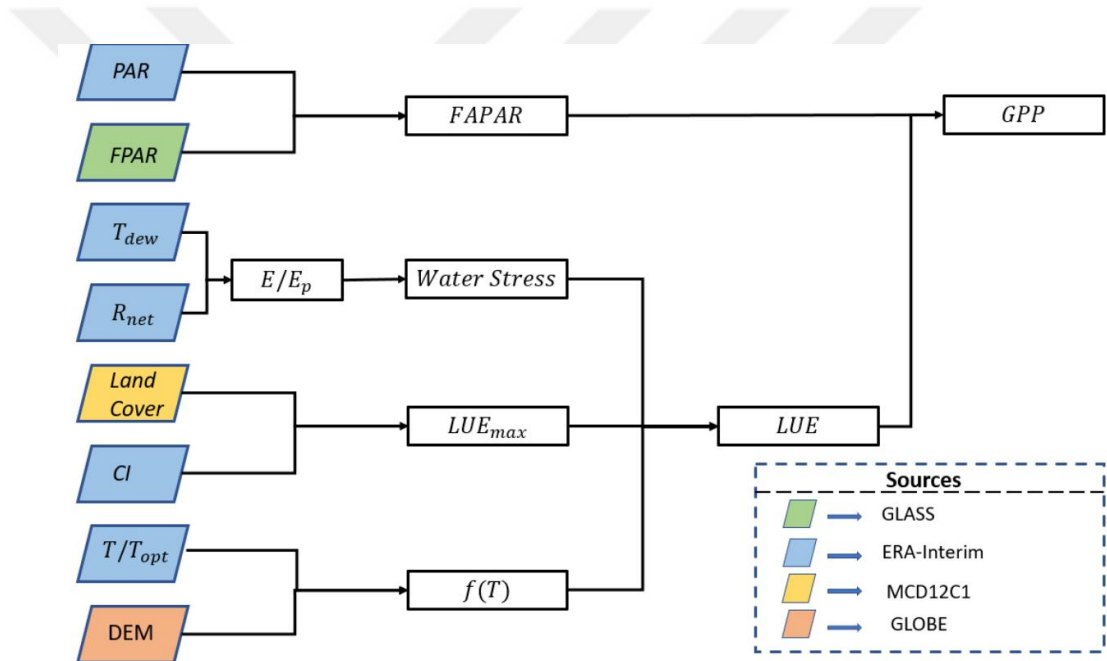


Figure A.2 : The MuSyQ GPP algorithm flowchart.

PMLV2 GPP

The Penman-Monteith-Leuning model (PML) model was initially developed by Leuning et al. (2008) introducing a biophysical model for surface conductance (G_s) based on the Penman-Monteith (PM) equation (Pei et al, 2020). Since then, there have been two improved versions of the PML model: PML-V1 and PML-V2. The PML-V2 employs a biophysical canopy conductance (G_c) model to combine Evapotranspiration (ET) with GPP, as described by Gan et al. (2018). To streamline the model, it retains only essential parameters that are of physiological significance for stomatal response and assimilation (Li et al, 2023). Unlike PML-V1, PML-V2 takes

into account the effects of increasing Leaf Area Index (LAI), rising CO₂ levels, and climate forcing on ET and GPP (Naeem et al, 2023). PML-V2 products consist of five components: GPP, vegetation transpiration (E_c), soil evaporation (E_s), evaporation of intercepted rainfall (E_i), and the evaporation from ice, snow, and water bodies (ET_{water}). In the PML-V2 model, GPP is calculated based photosynthesis (A_{gs}), constrained by the surface vapor pressure deficit (D) (equation. A.14).

$$GPP = A_{gs} f_D \quad (A.14)$$

$$A_{gs} = \frac{P_1 C_a}{k(P_2 + P_4)} \left\{ kLAI + \ln \frac{P_2 + P_3 + P_4}{P_2 + P_3 \exp(kLAI) + P_4} \right\} \quad (A.15)$$

$$P_1 = A_m \beta I_0 \eta, \quad P_2 = A_m \beta I_0, \quad P_3 = A_m \eta C_a, \quad P_4 = \beta I_0 \eta C_a \quad (A.16)$$

$$A_m \cong 0.5 \times V_m \quad (A.17)$$

$$V_m = \frac{V_{m,25} \exp[a(T - 25)]}{1 + \exp[b(T - 41)]} \quad (A.18)$$

$$f_D = \begin{cases} 1 & D < D_{min} \\ \frac{(D_{max} - D)}{D_{max} - D_{min}} & D_{min} < D < D_{max} \\ 0 & D > D_{max} \end{cases} \quad (A.19)$$

where f_D is the D constraint function, D_{min} is the threshold below which there is no vapor pressure constraint, D_{max} is the threshold above which there is no assimilation, D_0 is the reference vapor pressure deficit at stomatal conductance reduction, I_0 is the photosynthetically active radiation (PAR) from shortwave downward radiation, C_a is the atmospheric CO₂ concentration, $V_{m,25}$ is the maximum catalytic capacity of Rubisco per unit leaf area at 25 °C, β is the initial photochemical efficiency, η is the initial value of the slope of CO₂ response curve and m is the stomatal conductance coefficient.

TL-LUE GPP

The total GPP is estimated: The TL-LUE model calculates the GPP of sunlit and shaded leaves separately and sums them along the canopy (He et al, 2013). The revised TL-LUE model adds an atmospheric CO₂ concentration regulation scalar and modifies the air temperature regulation scalar (Bi et al, 2022). The total GPP is estimated:

$$GPP = GPP_{shade} - GPP_{sun} \quad (A.20)$$

$$GPP_{shade} = (\varepsilon_{msh} \times APAR_{sh}) \times T_s \times W_s \times C_s \quad (A.21)$$

$$GPP_{sun} = (\varepsilon_{msu} \times APAR_{su}) \times T_s \times W_s \times C_s \quad (A.22)$$

$$APAR_{sh} = (1 - \alpha) \times \left(\frac{PAR_{dif} - PAR_{dif,u}}{LAI} + C \right) \times LAI_{sh} \quad (A.23)$$

$$APAR_{su} = (1 - \alpha) \times \left(\frac{PAR_{dir} \times \cos\beta}{\cos\theta} + \frac{PAR_{dif} - PAR_{dif,u}}{LAI} + C \right) \times LAI_{su} \quad (A.24)$$

$$LAI_{su} = 2\cos\theta \times \left(1 - e^{-\frac{LAI \times \Omega}{2\cos\theta}} \right) \quad (A.25)$$

$$LAI_{sh} = LAI - LAI_{su} \quad (A.26)$$

$$C = 0.07 \times \Omega \times PAR_{dir} \times (1.1 - 0.1LAI)e^{-\cos\theta} \quad (A.27)$$

$$PAR_{dif} = \times (0.7527 + 3.8453R - 16.316R^2 + 18.962R^3 - 7.0802R^4) \quad (A.28)$$

$$PAR_{dir} = PAR - PAR_{dif} \quad (A.29)$$

$$PAR_{dif,u} = PAR_{dif} \times e^{-\frac{0.5 \times LAI \times \Omega}{\cos\theta}} \quad (A.30)$$

$$\cos\bar{\theta} = 0.537 + 0.025 \times LAI \quad (A.31)$$

These include albedo (α), which reflects the proportion of light that is reflected by a surface, and the solar zenith angle (θ), which indicates the position of the sun in the sky. The leaf angle (β), typically set at 60° , impacts how leaves intercept light. The clumping index (Ω) represents the grouping of leaves, affecting light penetration. Multiple scattered radiation (C) also plays a role in determining light availability. The model takes into account both direct (PAR_{dir}) and diffuse (PAR_{dif}) photosynthetically active radiation above the canopy, essential for photosynthesis. $PAR_{dif,u}$ is the diffuse PAR under the canopy. Leaf Area Index (LAI) is split into sunlit (LAI_{su}) and shaded (LAI_{sh}) portions, indicating the area of leaves receiving direct and indirect sunlight, respectively.

Lastly, the sky clearness index (R) is calculated as the ratio of solar radiation (S) to the product of the solar constant (S_0) and the cosine of the solar zenith angle ($\cos\theta$). This index reflects the atmospheric conditions influencing light penetration. θ is also used as the representative zenith angle for the transmission of diffuse radiation. α is albedo; θ is the solar zenith angle; β is the leaf angle, which is set as 60° ; Ω is the clumping index; C is multiple scattered radiation; PAR_{dir} and PAR_{dif} are the incoming direct and diffuse photosynthetically active radiation above the canopy; $PAR_{dif,u}$ denotes diffuse PAR under the canopy; LAI_{su} and LAI_{sh} are the LAI of sunlit and shaded leaves; R represents the sky clearness index and equals to $S/(S_0 \cos\theta)$, where S is solar radiation, and S_0 is the solar constant; θ is the representative zenith angle for diffuse radiation transmission.

FLUXCOM GPP

FLUXCOM is an initiative formed through the collaboration of various experts, aiming to understand uncertainties in empirical scaling and to present global Gross Primary Production (GPP) values based on machine learning (Tramontana et al, 2016). In FLUXCOM, a variety of machine learning-based regression tools such as model tree ensembles, multiple adaptive regression splines, artificial neural networks, and kernel methods are used. This approach is a significant step in identifying and assessing the limitations and sources of uncertainty in empirical scaling approaches. The study aims to estimate CO_2 and energy fluxes using two different scaling strategies: (1) 8-day average fluxes based solely on remotely sensed data, and (2) daily average fluxes based on remotely sensed and meteorological data. Machine learning (ML) approaches have been applied and extensively cross-validated with flux towers for both setups. The research focuses on CO_2 fluxes (gross primary production, terrestrial ecosystem respiration, net ecosystem exchange) and energy fluxes (latent heat, sensible heat, net radiation) based on these cross-validation results. For the calibration of GPP values in the study, FLUXNET GPP values were used. The following products from remote sensing were used in composing the FLUXCOM dataset: MOD11A2 Land Surface Temperature (LST); MOD13A2 Vegetation Index (Normalized Difference Vegetation Index (NDVI) and Enhanced Vegetation Index (EVI); MOD15A2 Leaf Area Index (LAI) and the fraction of absorbed photosynthetically active radiation (fPAR); and MCD43A2 and MCD43A4 Bidirectional Reflectance Distribution Function (BRDF) corrected surface reflections.

BRDF-corrected surface reflection data were further processed to calculate the Normalized Difference Water Index and the Land Surface Water Index (LSWI). Air temperature (T_{air}), global radiation (R_g), vapor pressure deficit (VPD), and precipitation (measured in situ at the locations of flux towers) were also included. Additionally, long-term time series of these variables from the ERA-Interim dataset for the 1989-2010 period were used; these series underwent bias correction for each site based on the overlap period with in situ measurements. These long-term meteorological data were primarily used to calculate consistent measurements of climatic variables for all sites, considering the temporal scope of the data from different sites. Furthermore, a combination of these ERA-Interim data and in situ measured data was used to create a continuous time series to calculate the soil water availability index (WAI).

Table A.1 : Mann-Kendall and Sens' Slope trend analysis results of Antalya region.

Dataset		Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
TL-LUE	Z	1.78	3.39	4.33	3.92	4.85	5.15	3.47	2.38	2.11	3.2	3.65	2.59
	p	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.03	0.00	0.00	0.01
	Slope	0.006	0.02	0.003	0.04	0.041	0.005	0.03	0.019	0.013	0.021	0.026	0.013
FLUXC OM	Z	1.47	2.28	2.43	1.92	0.8	0.64	0.57	0.71	1.08	1.94	1.29	2.13
	p	0.13	0.02	0.01	0.05	0.42	0.52	0.56	0.47	0.27	0.05	0.19	0.03
	Slope	0.001	0.002	0.002	0.002	0.001	0.003	0.001	0.001	0.001	0.003	0.002	0.002
GOFIS	Z	1.74	2.76	1.41	2.36	2.31	2.7	3.38	3.61	3.55	2.93	3.04	3.32
	p	0.08	0.00	0.15	0.01	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Slope	0.005	0.012	0.014	0.027	0.018	0.022	0.037	0.041	0.021	0.018	0.015	0.013
MODIS	Z	-0.16	-2.82	-3.92	-3.73	-3.14	2.17	3.99	2.3	1.71	4.37	5.22	5.41
	p	0.87	0.00	0.00	0.00	0.00	0.03	0.00	0.02	0.08	0.00	0.00	0.00
	Slope	-	-0.02	-0.06	-0.05	-0.06	0.04	0.14	0.08	0.02	0.06	0.09	0.07
MuSyQ	Z	1.53	4.19	4.14	4.19	5.22	2.58	-1.25	-1.04	0.67	2.28	3.21	1.23
	p	0.12	0.00	0.00	0.00	0.00	0.00	0.20	0.29	0.49	0.02	0.00	0.21
	Slope	0.004	0.012	0.015	0.019	0.022	0.018	-	-	0.002	0.009	0.014	0.003
PMLV2	Z	1.33	1.59	2.3	3.74	5.82	4.99	3.63	2.95	3.49	1.56	2.84	2.95
	p	0.18	0.11	0.021	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.004	0.003
	Slope	0.004	0.008	0.011	0.021	0.05	0.039	0.013	0.019	0.019	0.005	0.01	0.012

Table A.2 : Mann-Kendall and Sens' Slope trend analysis results of Erzincan region.

Dataset		Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
TL-LUE	Z	-0.17	1.65	3.28	3.39	4.18	5.72	3.09	-1.78	-1.92	-1.03	1.55	7.81
	p	0.86	0.09	0.001	0.00	0.00	0.00	0.001	0.07	0.053	0.3	0.11	0.043
	Slope	0.00	0.00	0.006	0.019	0.046	0.057	0.029	-0.01	-0.00	-	0.001	0.00
FLUXC OM	Z	-3.14	1.17	1.5	1.82	0.82	2.34	-1.31	-0.45	-0.64	-0.12	0.1	0.31
	p	0.7	0.23	0.13	0.06	0.4	0.019	0.18	0.64	0.52	0.89	0.99	0.75
	Slope	0.00	0.00	0.00	0.001	0.001	0.002	-	-0.00	-0.00	-0.00	-0.00	0.00
GOFIS	Z	0.62	-0.28	0.64	1.52	3.27	2.76	1.52	1.18	2.36	2.7	2.14	0.84
	p	0.53	0.77	0.51	0.12	0.001	0.005	0.12	0.23	0.017	0.006	0.032	0.39
	Slope	0.001	-0.00	0.001	0.012	0.044	0.04	0.021	0.009	0.008	0.007	0.004	0.001
MODIS	Z	1.33	-2.79	-2.56	-3.47	-4.37	-3.27	4.63	4.5	4.63	4.96	4.63	4.57
	p	0.18	0.005	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Slope	0.01	-	-0.008	-0.026	-0.09	-0.056	0.07	0.098	0.082	0.073	0.048	0.015
MuSyQ	Z	1.38	3.38	2.53	1.96	4.14	4.92	4.34	3.19	3.24	2.46	2.53	0.9
	p	0.16	0.00	0.011	0.049	0.00	0.00	0.00	0.001	0.001	0.013	0.01	0.363
	Slope	0.00	0.001	0.002	0.004	0.019	0.033	0.023	0.013	0.007	0.004	0.002	0.00
PMLV2	Z	0.65	0.51	2.47	2.3	4.31	4.68	3.23	2.38	3.69	2.98	2.92	1.47
	p	0.51	0.6	0.013	0.021	0.00	0.00	0.001	0.017	0.002	0.003	0.003	0.13
	Slope	0.00	0.000	0.003	0.009	0.026	0.036	0.016	0.007	0.013	0.004	0.008	0.001

Table A.3 : Mann-Kendall and Sens' Slope trend analysis results of Izmir region.

Dataset		Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
TL-LUE	Z	1.66	2.45	4.44	4.33	4.032	3.58	2.94	-0.13	-1.89	0.65	2.53	2.94
	p	0.09	0.013	0.00	0.00	0.000	0.00	0.003	0.89	0.058	0.511	0.011	0.003
	Slope	0.012	0.025	0.059	0.056	0.049	0.078	0.05	-	0.017	0.005	0.023	0.019
FLUXC OM	Z	8.15	0.5	1.82	0.64	0.61	1.33	0.47	-2.15	-0.66	2.94	2.66	2.38
	p	0.9	0.61	0.067	0.52	0.53	0.18	0.63	0.031	0.5	0.003	0.007	0.016
	Slope	0.00	0.00	0.003	0.000	0.000	0.009	0.002	-0.00	0.005	0.004	0.004	0.002
GOFIS	Z	1.91	3.38	2.48	2.82	1.46	3.61	3.61	2.53	1.86	0.28	1.07	1.8
	p	0.055	0.00	0.013	0.004	0.14	0.00	0.00	0.011	0.06	0.77	0.28	0.07
	Slope	0.025	0.046	0.040	0.048	0.021	0.056	0.066	0.019	0.013	0.004	0.013	0.018
MODIS	Z	-0.48	-2.49	-3.53	-2.3	-1.2	2.88	3.73	0.74	-3.6	0.22	4.63	5.09
	p	0.62	0.012	0.00	0.021	0.22	0.003	0.00	0.455	0.00	0.82	0.00	0.00
	Slope	-	-	-0.07	-0.05	-0.023	0.08	0.17	0.018	0.00	0.006	0.078	0.088
MuSyQ	Z	3.34	5.05	4.97	4.85	5.48	2.91	-0.57	-1.68	0.82	3.29	4.22	2.96
	p	0.00	0.00	0.00	0.00	0.00	0.003	0.56	0.09	0.40	0.00	0.00	0.003
	Slope	0.014	0.031	0.039	0.049	0.047	0.024	-	-	0.005	0.025	0.028	0.011
PMLV2	Z	1.93	1.42	3.01	4.28	4.94	4.62	2.81	1.7	3.23	3.91	3.09	4.03
	p	0.053	0.155	0.002	0.00	0.00	0.00	0.004	0.088	0.00	0.00	0.001	0.00
	Slope	0.008	0.007	0.019	0.025	0.048	0.036	0.027	0.012	0.02	0.015	0.016	0.016

Table A.4 : Mann-Kendall and Sens' Slope trend analysis results of Kirklareli region.

Dataset		Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
TL-LUE	Z	1.66	1.74	2.75	2.79	4.82	4.74	2.64	1.18	-0.09	1.78	2.19	2.68
	p	0.09	0.08	0.005	0.005	0.000	0.00	0.008	0.23	0.95	0.074	0.028	0.007
	Slope	0.003	0.009	0.031	0.059	0.085	0.094	0.064	0.018	-0.00	1.65	0.012	0.008
FLUXCOM	Z	0.29	2.41	2.8	1.87	-0.5	-0.5	-0.29	-0.73	-1.11	3.08	1.36	-0.89
	p	0.77	0.015	0.004	0.06	0.61	0.61	0.77	0.46	0.25	0.02	0.17	0.36
	Slope	0.00	0.002	0.004	0.005	-0.002	-	-	-	-	0.003	0.002	-0.00
GOFIS	Z	1.91	2.03	1.01	1.41	3.21	3.1	3.21	2.25	1.18	0.73	1.52	1.91
	p	0.055	0.042	0.3	0.15	0.013	0.001	0.001	0.024	0.23	0.46	0.12	0.055
	Slope	0.001	0.018	0.018	0.038	0.005	0.058	0.1	0.043	0.01	0.007	0.011	0.008
MODIS	Z	0.22	-1.74	-2.88	-3.53	-3.27	1.45	3.14	3.99	4.44	4.44	4.25	4.57
	p	0.82	0.074	0.003	0.000	0.001	0.14	0.00	0.00	0.00	0.00	0.00	0.00
	Slope	0.001	-0.019	-0.04	-0.09	-0.1	0.05	0.098	0.11	0.088	0.1	0.089	0.046
MuSyQ	Z	1.23	2.31	2.08	2.33	5.1	5.75	5.9	5.72	6.16	4.29	4.04	1.3
	p	0.21	0.02	0.036	0.019	0.00	.00	0.00	0.00	0.00	0.00	0.00	0.19
	Slope	0.001	0.007	0.012	0.029	0.054	0.068	0.081	0.065	0.047	0.023	0.012	0.00
PMLV2	Z	0.79	0.79	1.56	3.38	3.66	2.24	3.74	3.63	2.78	3.4	2.55	0.79
	p	0.426	0.426	0.11	0.00	0.00	0.024	0.00	0.00	0.005	0.00	0.01	0.426
	Slope	0.02	0.004	0.013	0.035	0.022	0.018	0.03	0.021	0.014	0.014	0.008	0.001

Table A.5 : Mann-Kendall and Sens' Slope trend analysis results of Konya region.

Dataset		Jan	Feb	Mar	Apr	May	Ju	Jul	Au	Sep	Oct	No	De
TL-LUE	Z	1.78	2.86	3.09	3.35	2.86	2.83	4.07	4.67	3.2	2	2.45	1.66
	p	0.07	0.004	0.019	0.007	0.004	0.004	0.000	0.000	0.001	0.044	0.013	0.095
	Slope	0.00	0.003	0.01	0.032	0.037	0.024	0.022	0.015	0.007	0.002	0.002	0.001
FLUXCOM	Z	-1.06	1.31	1.38	0.38	-1.15	-1.38	-0.31	-1.16	3.49	4.07	0.66	-0.78
	p	0.28	0.18	0.165	0.7	0.24	0.16	0.73	0.9	0.9	0.68	0.5	0.43
	Slope	-0.00	0.00	0.012	0.00	-0.00	-0.00	-0.00	0.00	0.00	0.00	0.00	-0.00
GOFIS	Z	0.39	0.73	0.84	1.57	0	1.63	5.02	5.13	4.34	3.38	0.95	0.95
	p	0.69	0.46	0.39	0.11	1	0.1	0.00	0.00	0.00	0.00	0.33	0.33
	Slope	0.001	0.006	0.013	0.03	0.00	0.035	0.087	0.067	0.029	0.011	0.002	0.003
MODIS	Z	0.22	-1.71	-1.74	-2.49	-1.78	2.43	4.18	3.34	3.73	4.83	4.63	3.99
	p	0.82	0.08	0.07	0.012	0.07	0.01	0.00	0.00	0.00	0.00	0.00	0.00
	Slope	0.001	-0.007	-0.013	-0.04	-0.031	0.056	0.087	0.051	0.023	0.037	0.034	0.02
MuSyQ	Z	1.33	3.19	3.24	4.12	4.65	4.47	4.62	4.9	5.12	4.82	4.45	1.59
	p	0.18	0.001	0.001	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11
	Slope	0.001	0.004	0.009	0.027	0.042	0.030	0.02	0.019	0.016	0.017	0.005	0.001
PMLV2	Z	0.99	1.9	2.78	4.65	4.23	3.35	4.31	5.9	5.96	0.85	1.98	0.96
	p	0.32	0.057	0.005	0.00	0.00	0.00	0.00	0.00	0.00	0.39	0.046	0.33
	Slope	0.001	0.003	0.008	0.017	0.032	0.017	0.016	0.014	0.001	0.002	0.002	0.001

Table A.6 : Mann-Kendall and Sens' Slope trend analysis results of Samsun region.

Dataset		Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
TL-LUE	Z	1.59	2.04	2.75	2.15	4.59	5.23	3.43	0.65	-1.18	-0.58	2.02	1.4
	p	0.11	0.04	0.005	0.03	0.00	0.00	0.00	0.51	0.23	0.56	0.04	0.15
	Slope	0.000	0.01	0.021	0.05	0.088	0.1	0.064	0.01	-0.00	-0.00	0.008	0.004
FLUXCOM	Z	-5.82	1.68	0.52	1.75	-0.85	0.4	-0.08	-1.1	-1.2	0.01	0.38	-0.45
	p	0.9	0.09	0.6	0.078	0.39	0.68	0.93	0.26	0.2	0.99	0.7	0.64
	Slope	0.00	0.001	0.00	0.003	-0.00	0.001	-0.00	-	-	0.00	0.00	-0.00
GOFIS	Z	1.91	2.03	1.74	1.46	2.7	2.03	3.49	3.15	1.41	0.78	0.84	3.1
	p	0.055	0.042	0.08	0.14	0.006	0.042	0.00	0.001	0.15	0.42	0.39	0.001
	Slope	0.004	0.013	0.022	0.038	0.062	0.046	0.009	0.066	0.024	0.01	0.007	0.012
MODIS	Z	0.94	-1.97	-2.23	-4.31	-3.53	-2.88	3.01	4.44	4.31	4.63	4.44	5.22
	p	0.34	0.047	0.025	0.00	0.00	0.003	0.002	0.00	0.00	0.00	0.00	0.00
	Slope	0.003	-0.015	-0.031	-0.1	-0.17	-	0.081	0.178	0.154	0.173	0.115	0.006
MuSyQ	Z	0.4	3.29	3.59	3.26	4.45	3.99	4.75	5.38	4.62	3.92	3.24	0.57
	p	0.68	0.00	0.00	0.001	0.00	0.00	0.00	0.00	0.00	0.00	0.001	0.56
	Slope	0.000	0.007	0.011	0.035	0.05	0.03	0.03	0.04	0.03	0.02	0.008	0.00
PMLV2	Z	3.38	3.26	3.09	4.62	5.11	4.37	3.74	4.06	3.35	3.63	3.77	2.1
	p	0.00	0.001	0.002	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.035
	Slope	0.01	0.013	0.018	0.048	0.064	0.07	0.031	0.033	0.019	0.02	0.014	0.003

Table A.7 : Mann-Kendall and Sens' Slope trend analysis results of Sanliurfa region.

Dataset		Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
TL-LUE	Z	0.8	1.96	2.94	3.92	2.08	0.13	2.45	4.4	5.72	4.33	2.3	0.84
	p	0.41 9	0.048	0.003	0.00	0.037	0.89	0.01 3	0.00	0.00	0.00	0.02 1	0.39
	Slope	0.00 1	0.012	0.045	0.05 8	0.03	0.00	0.00 5	0.01 3	0.02 3	0.01 5	0.00 4	0.00 2
FLUXCOM	Z	2.34	2.55	2.41	-2.57	-0.89	-1.2	-1.99	-1.78	-3.34	-1.43	0.64	0.33
	p	0.01 9	0.01	0.015	0.01	0.36	0.23	0.04	0.07	0.00	0.15	0.52	0.73
	Slope	0.00	0.001	0.002	-0.00	-0.00	-0.00	- 0.00 1	- 0.00 1	- 0.00 1	-0.00	0.00	0.00
GOFIS	Z	0.95	1.24	0.84	1.18	1.12	4	4.17	4.68	5.3	4.68	1.91	1.18
	p	0.33	0.21	0.39	0.23	0.25	0.00	0.00	0.00	0.00	0.00	0.05 5	0.23
	Slope	0.00 6	0.017	0.031	0.03 1	0.028	0.02 6	0.04 8	0.07 3	0.06 7	0.03 3	0.01	0.00 5
MODIS	Z	0.61	-1.84	-3.6	-2.69	2.88	3.92	4.12	-1.07	-3.53	1.26	4.31	4.25
	p	0.53	0.064	0.003	0.00 7	0.003	0.00	0.00	0.28	0.00	0.2	0.00	0.00
	Slope	0.00 2	-0.018	-0.068	-0.06	0.059	0.12	0.03 7	- 0.00 7	- 0.01 8	0.00 6	0.03 3	0.03
MuSyQ	Z	2.86	3.64	3.44	3.44	3.74	4.57	6.26	6.46	7.16	6.78	5.43	2.96
	p	0.00 4	0.00	0.005	0.00 5	0.001	0.00	0.00	0.00	0.00	0.00	0.00	0.00 3
	Slope	0.00 4	0.013	0.028	0.03 1	0.033	0.02 1	0.02 4	0.03 1	0.02 8	0.01 6	0.01	0.00 3
PMLV2	Z	0.25	1.07	3.57	3.55	0.11	-2.38	2.07	4.8	5.45	1.81	2.58	1.5
	p	0.79	0.28	0.00	0.00	0.9	0.01 7	0.03 8	0.00	0.00	0.06 9	0.00 9	0.13 2
	Slope	0.01 1	0.007	0.021	0.03	0.001	-0.01	0.00 8	0.01 7	0.02 8	0.00 3	0.00 4	0.00 3

Table A.8 : Variance contribution rate of IMFs.

Region	Dataset	IMF1 (%)	IMF2 (%)	IMF3 (%)	IMF4 (%)	IMF5 (%)	IMF6 (%)	IMF7 (%)	IMF8 (%)	TREND (%)
Antalya	TL-LUE	68,2	28,4	0,62	0,14	0,42	-	-	-	2,12
	FLUXCOM	97	2,2	0,38	0,07	0,14	-	-	-	0,13
	GOFIS	78,8	5	0,67	1,7	3,56	1,24	3,39	-	5,5
	MODIS	40,4	54,8	1,89	0,42	0,13	-	-	-	2,3
	MuSyQ	52,5	40,6	1	0,69	1	-	-	-	4,1
	PMLV2	12,6	80	1,9	0,8	0,6	-	-	-	3,6
Erzincan	TL-LUE	80	8,5	0,48	2,39	3,9	-	-	-	4,1
	FLUXCOM	85	12,2	1,63	0,34	0,53	-	-	-	0,04
	GOFIS	32,7	65,3	0,64	0,19	0,19	-	-	-	0,9
	MODIS	40,4	54,8	1,89	0,42	0,13	-	-	-	2,3
	MuSyQ	86,8	8,6	1,16	0,7	0,07	0,21	-	-	2,3
	PMLV2	69,7	24,7	0,87	0,43	0,63	-	-	-	4
Izmir	TL-LUE	24,6	61,2	3,28	1,68	0,38	0,55	-	-	8,17
	FLUXCOM	85	1	0,33	7,1	4,77	-	-	-	1,7
	GOFIS	60,6	31,8	2,13	0,38	1,47	-	-	-	3,6
	MODIS	36	41	20,6	1	0,34	-	-	-	0,8
	MuSyQ	22,1	60	1,38	0,32	0,34	6	-	-	10
	PMLV2	28,5	61	3,4	2,1	1,24	-	-	-	3,2
Kirkklareli	TL-LUE	68,6	25,3	1,1	0,34	0,08	0,4	-	-	4,1
	FLUXCOM	83,1	11	3	1,81	0,57	0,1	-	-	0,2
	GOFIS	67,1	25,1	1,72	1,75	1,72	-	-	-	2,5
	MODIS	35	62,1	1,24	0,4	0,23	-	-	-	1
	MuSyQ	72,6	15,7	1	1,62	1,2	1,45	-	-	6,5
	PMLV2	39,2	53	3,06	0,32	0,8	1	-	-	2,5
Konya	TL-LUE	57,5	25,4	5,7	2,39	0,38	1,28	-	-	7,5
	FLUXCOM	83,1	11	3,1	1,81	0,57	0,1	-	-	0,25
	GOFIS	34,8	47	9,29	1,4	1	1,47	-	-	5
	MODIS	28,9	57,8	7,8	1,39	0,62	-	-	-	3,5
	MuSyQ	57	16,1	7	2,1	1,2	-	-	-	16,45
	PMLV2	40,6	39,8	6,9	1	2,13	-	-	-	10
Samsun	TL-LUE	64,5	10,5	0,75	2,13	5,93	5,5	1,74	-	9
	FLUXCOM	97	2,21	0,38	0,07	0,14	-	-	-	0,13
	GOFIS	86	8,51	1	1,13	-	-	-	-	3,5
	MODIS	76,5	19,7	1,43	1,7	0,37	-	-	-	0,17
	MuSyQ	79,4	12,7	4,39	0,45	0,9	-	-	-	2,1
	PMLV2	56,9	34	1,22	1,4	1,23	2	-	-	3,1
Sanliurfa	TL-LUE	50,3	42,1	2,63	0,93	-	-	-	-	4
	FLUXCOM	45,5	46,7	2,8	2,22	1	1,1	-	-	0,3
	GOFIS	41,3	4,84	1,7	1,7	0,67	-	-	-	10
	MODIS	40,1	55,5	1,39	0,9	0,51	-	-	-	1,5
	MuSyQ	34,9	49,3	1,7	0,72	1,5	-	-	-	12
	PMLV2	38,2	49	6,8	1,77	1,7	-	-	-	2,5

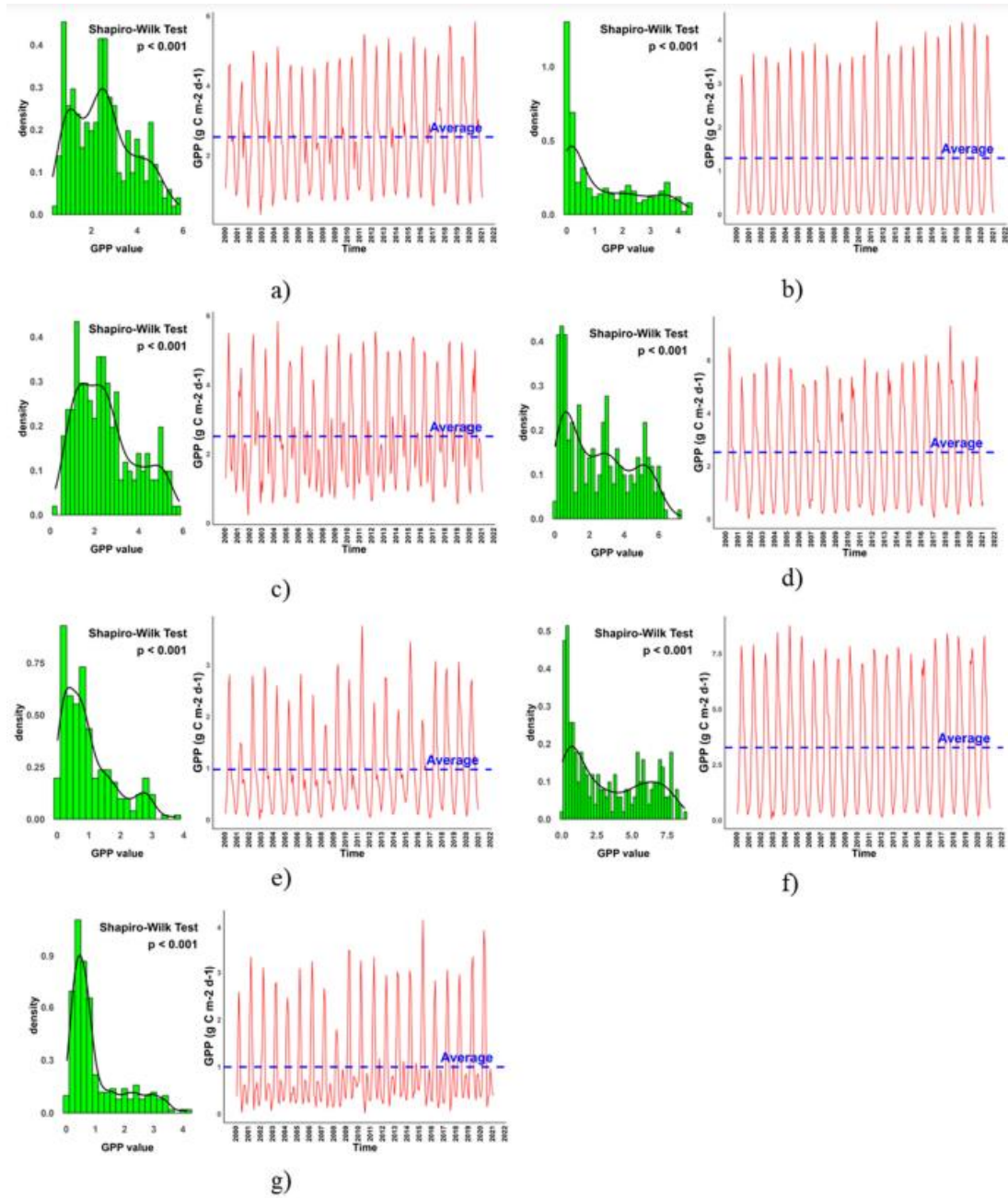


Figure A.3 : Histograms and time series of MODIS datasets (a) Antalya, (b) Erzincan, (c) Izmir, (d) Kirlareli, (e) Konya, (f) Samsun, (g) Sanliurfa.

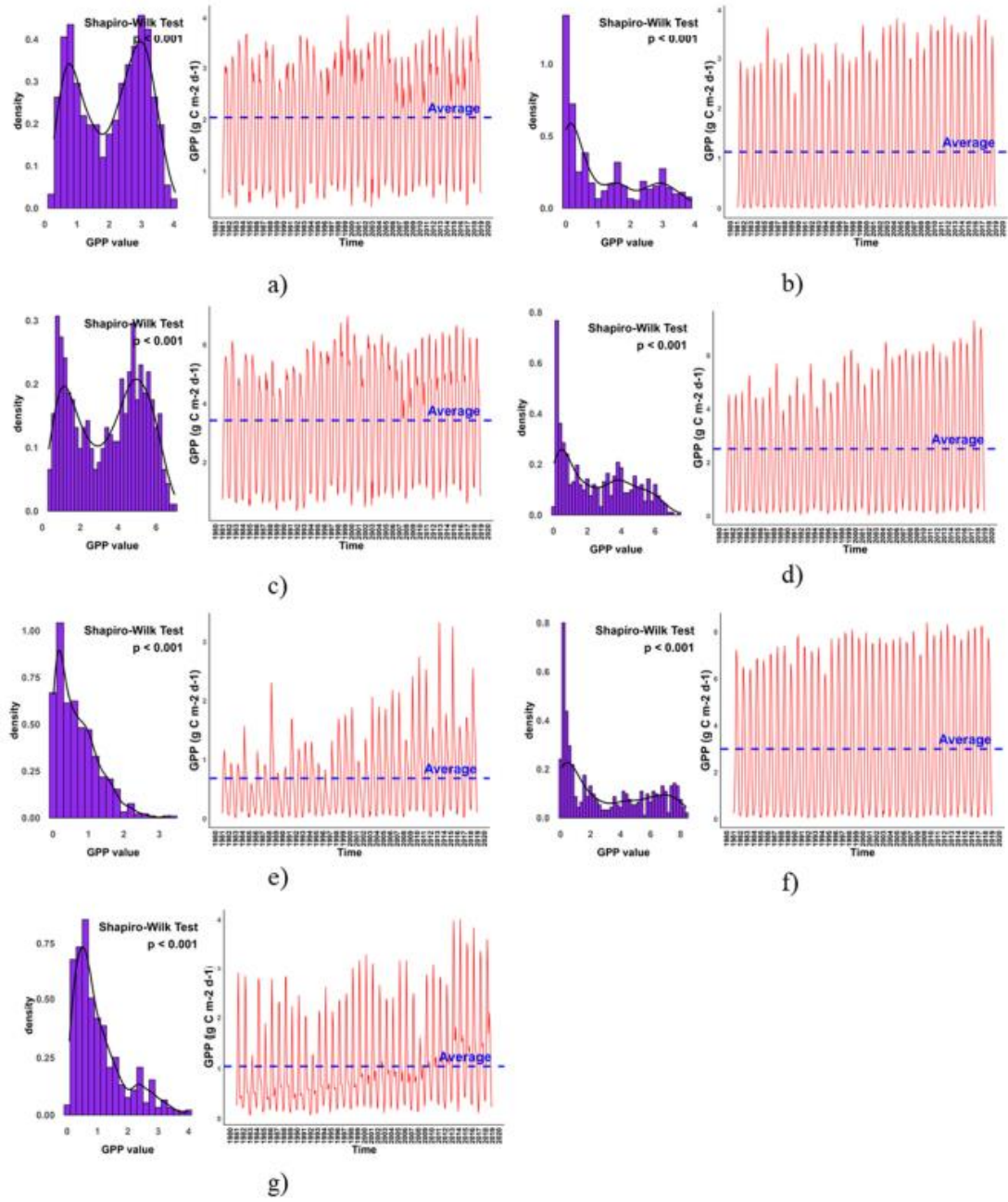


Figure A.4 : Histograms and time series of MuSyQ datasets (a) Antalya, (b) Erzincan, (c) Izmir, (d) Kırklareli, (e) Konya, (f) Samsun, (g) Sanliurfa.

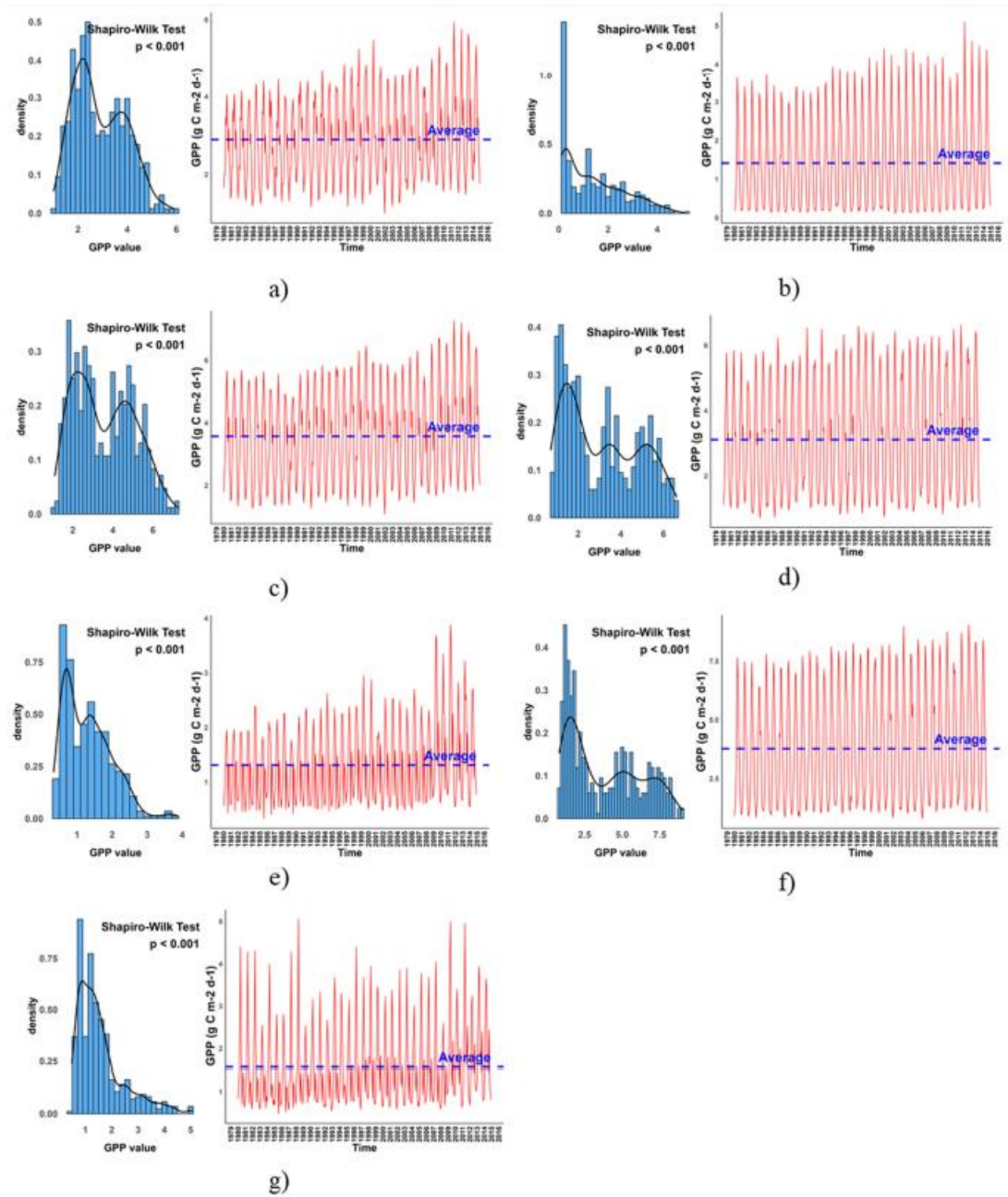


Figure A.5 : Histograms and time series of PMLV2 datasets (a) Antalya, (b) Erzincan, (c) Izmir, (d) Kirlareli, (e) Konya, (f) Samsun, (g) Sanliurfa.

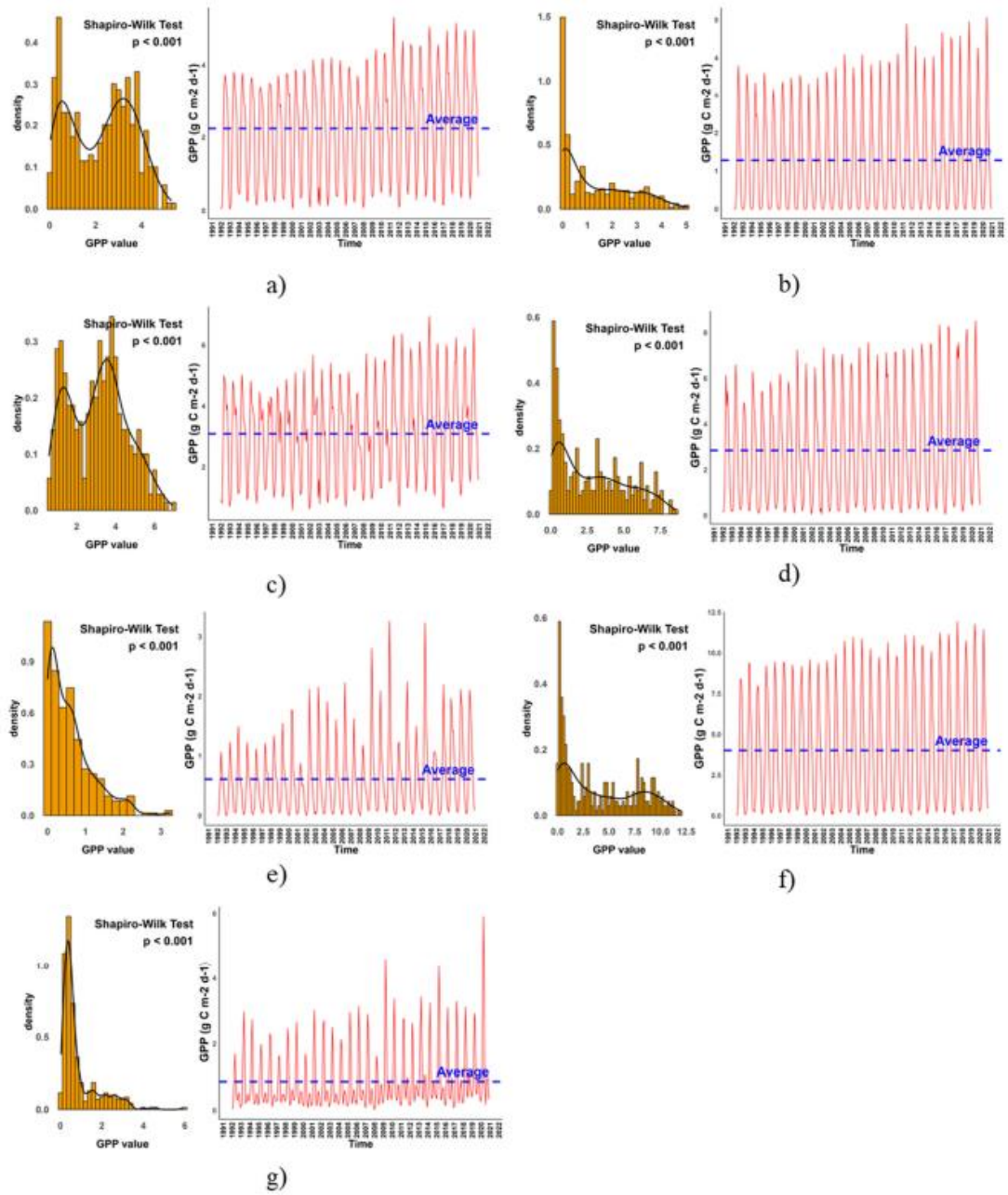


Figure A. 6 : Histograms and time series of TL-LUE datasets (a) Antalya, (b) Erzincan, (c) Izmir, (d) Kırklareli, (e) Konya, (f) Samsun, (g) Sanliurfa.

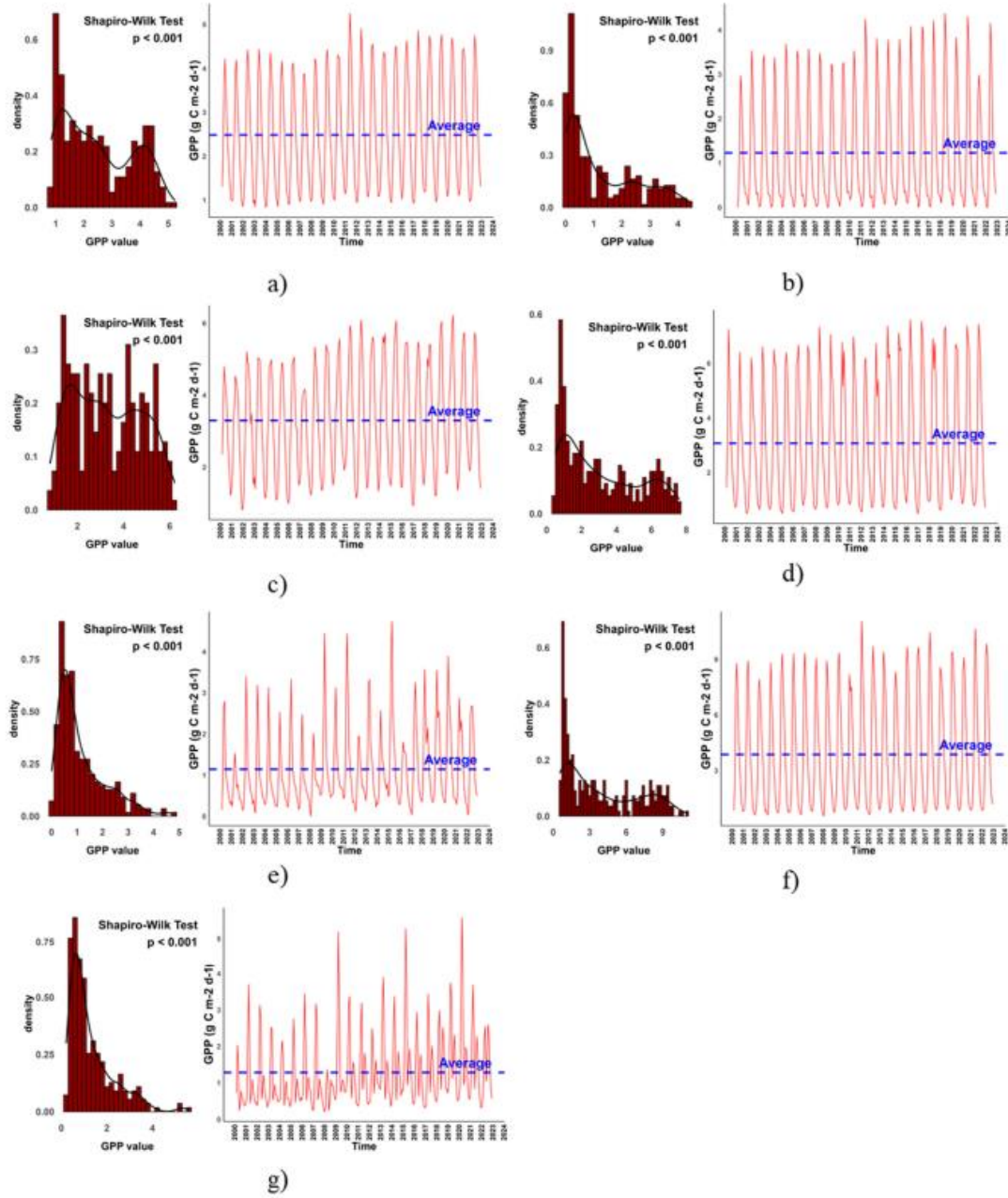


Figure A.7 : Histograms and time series of GOSIF datasets (a) Antalya, (b) Erzincan, (c) Izmir, (d) Kirlareli, (e) Konya, (f) Samsun, (g) Sanliurfa.

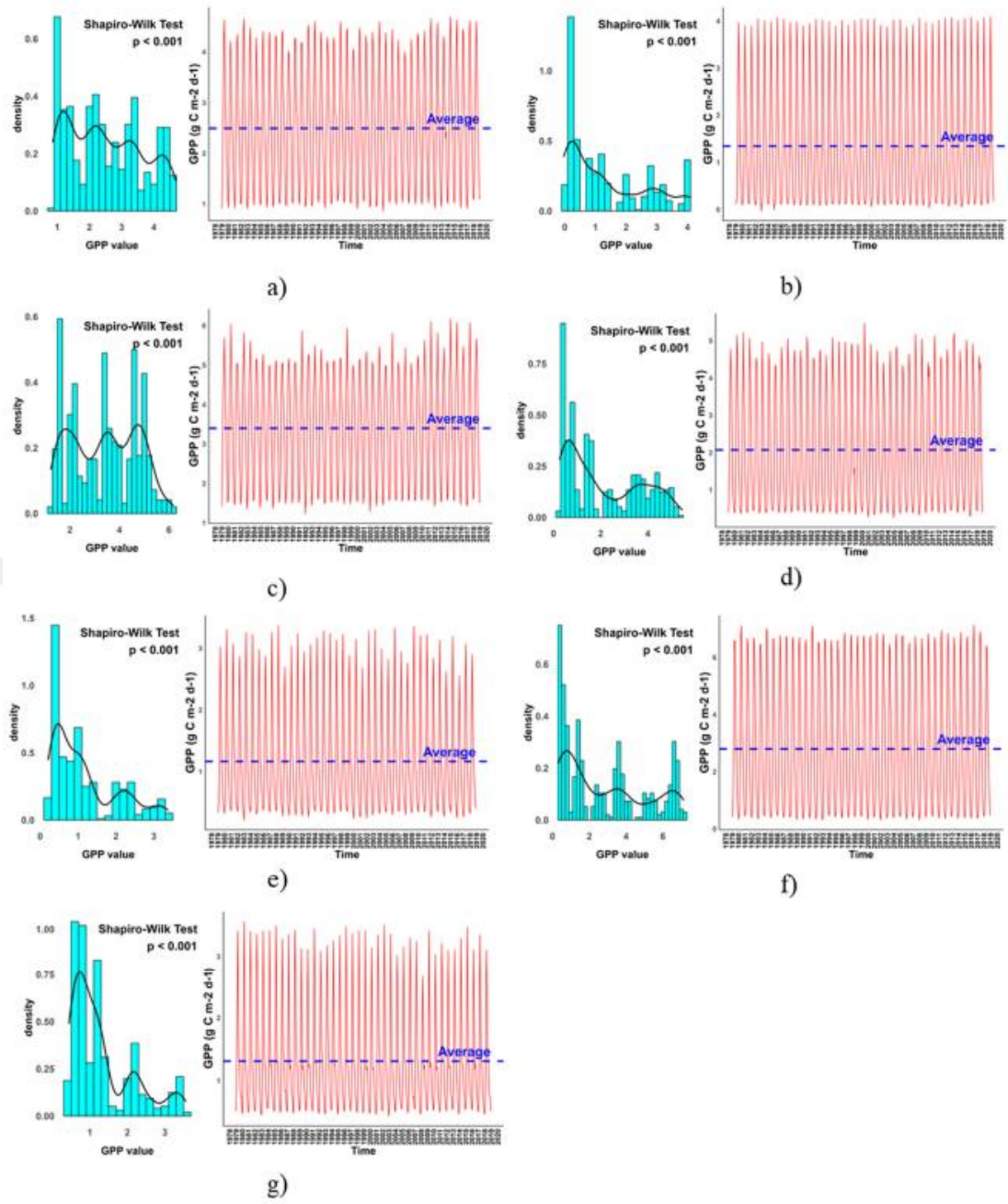


Figure A.8 : Histograms and time series of FLUXCOM datasets (a) Antalya, (b) Erzincan, (c) Izmir, (d) Kirlklareli, (e) Konya, (f) Samsun, (g) Sanliurfa.

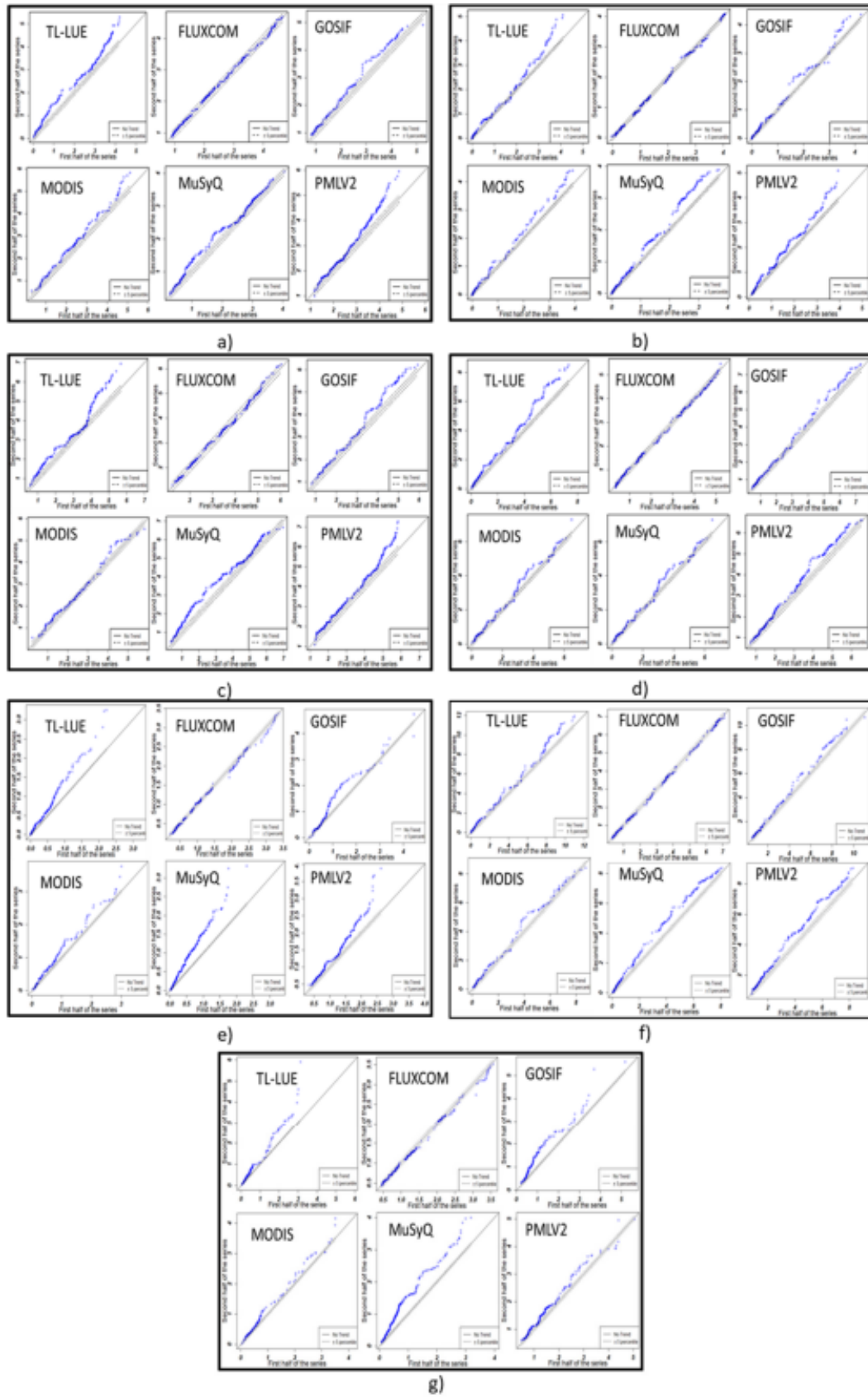


Figure A.9 : ITA trend analysis results (horizontal and vertical axes show GPP values in $\text{g C m}^2 \text{d}^{-1}$), (a) Antalya, (b) Erzincan, (c) Izmir, (d) Kırklareli, (e) Konya, (f) Samsun, (g) Sanliurfa.

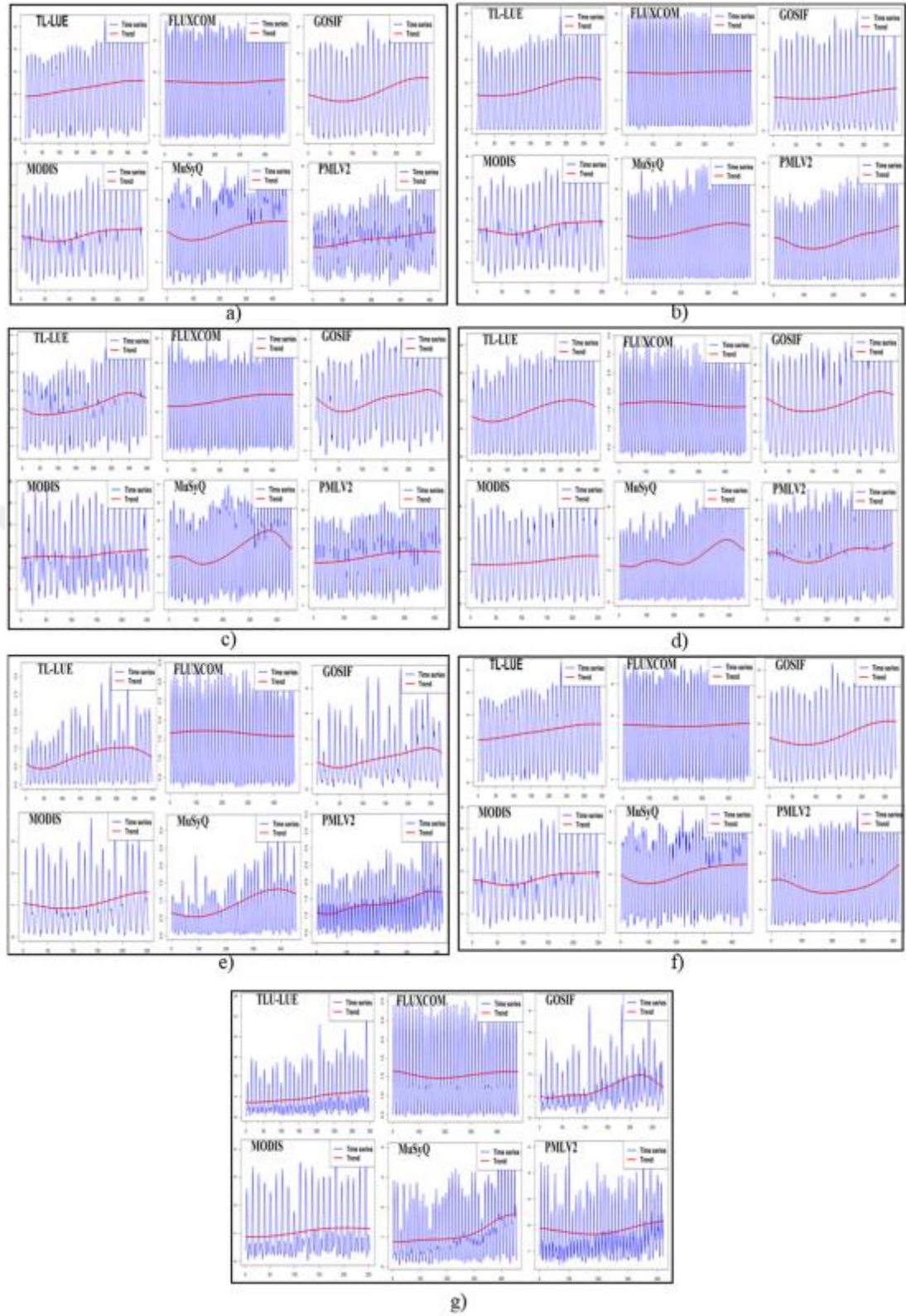


Figure A.10 : EMD trend analysis results (the horizontal axis shows time; the vertical axis shows GPP values), (a) Antalya, (b) Erzincan, (c) Izmir, (d) Kırklareli, (e) Konya, (f) Samsun, (g) Sanliurfa.



APPENDIX B : Additional figures and tables for Chapter 7.

Table B1. The mean [minimum: maximum] winter wheat yield (t/ha) per district in the study area from 2001 to 2020.

District	Mean [Min: Max]	District	Mean [Min: Max]	District	Mean [Min: Max]
Agacoren	2.32[0.85:3.19]	Derbent	1.91[1.27:3.18]	Kulu	2.18[0.64:3.89]
Akoren	1.67[0.34:2.8]	Doganhisar	2.23[1.45:2.83]	Meram	2.41[1.43:3.75]
Aksaray_Merkez	2.61[1.27:3.5]	Emirgazi	2.09[0.41:3.44]	Nigde_Merkez	2.58[1.45:4.28]
Aksehir	2.34[1.02:3.47]	Eregli	2.59[0.8:3.62]	Ortakoy	2.37[0.79:3.4]
Altunhisar	3.48[0.85:5.46]	Eskil	3.19[1.27:5.42]	Sarayonu	2.07[1.27:2.83]
Altinekin	1.75[1.03:2.51]	Gulagac	2.1[1.36:2.83]	Sarıyahsi	2.12[1.38:2.83]
Ayranci	1.37[0.6:1.89]	Guneysinir	2.22[0.81:3.53]	Selcuklu	2.42[1.29:4.2]
Beysehir	2.92[1.71:3.89]	Guzelyurt	1.74[1.03:2.52]	Seydisehir	3.18[2.04:4.91]
Bor	2.12[1.17:3.33]	Huyuk	2.37[1.44:3.99]	Tuzlukcu	2.25[1.02:3.53]
Camardi	1.86[1.19:2.6]	Ilgin	2.45[1.75:3.53]	Ulukisla	2.18[1.45:2.94]
Celtik	2.66[1.19:4.36]	Kadinhani	2.75[0.76:4.42]	Yalihuyuk	2.7[1.2:4.77]
Ciftlik	2.61[1.49:3.97]	Karaman_Merkez	2.33[1.54:2.86]	Yunak	2.27[1.03:3.79]
Cihanbeyli	2.29[0.95:3.4]	Karapinar	3.07[1.13:4.91]		
Cumra	2.74[1.44:4.1]	Karatay	2.67[0.64:4.92]		

Table B2. XGBoost hyperparameter range and optimum values.

Hyperparameter	Explanation	Range	Optimum Value
num_round	The number of rounds to run the training	[1, ∞]	100
alpha	L1 regularization term on weights.	[0, ∞]	0.01
lambda	L2 regularization term on weights.	[0, ∞]	0.03
max_depth	Maximum depth of a tree.	[1, ∞]	10
eta	To prevent overfitting, step size shrinkage is used in updates.	[0, 1]	0.01
early_stopping_rounds	The model continues to train until the validation score no longer improves.	[1, ∞]	30
subsample	Subsampling ratio of the training instance.	[0, 1]	0.8

Table B3. XGBoost model performance evaluation results.

Number	Combination	R		RMSE		AIC
		Train	Test	Train	Test	Train
1	TCI_PCI_VCI_GPP_NDWI_SMCi	0.820	0.687	30.491	31.488	4103.93
2	PCI_VCI_GPP_NDWI_SMCi	0.814	0.678	30.523	31.683	4107.51
3	TCI_PCI_VCI_NDWI_SMCi	0.813	0.657	30.550	31.554	4108.23
4	TCI_PCI_GPP_NDWI_SMCi	0.812	0.666	30.565	32.040	4108.84
5	TCI_PCI_VCI_GPP_NDWI	0.809	0.676	30.630	31.889	4110.40
6	PCI_VCI_NDWI_SMCi	0.806	0.668	30.762	31.983	4115.53
7	PCI_VCI_GPP_NDWI	0.805	0.664	30.858	31.943	4119.29
8	TCI_PCI_GPP_NDWI	0.805	0.663	30.879	31.502	4120.1
9	PCI_GPP_NDWI_SMCi	0.804	0.678	31.004	31.952	4124.96
10	TCI_VCI_GPP_NDWI_SMCi	0.803	0.676	31.023	31.582	4125.68
11	TCI_PCI_VCI_NDWI	0.803	0.666	31.049	31.525	4126.71
12	TCI_PCI_VCI_GPP	0.802	0.598	31.123	31.796	4129.55
13	TCI_VCI_NDWI_SMCi	0.801	0.662	31.257	31.642	4134.71
14	TCI_PCI_GPP_SMCi	0.799	0.613	31.277	31.541	4135.47
15	TCI_PCI_VCI_GPP_SMCi	0.795	0.626	31.342	32.944	4137.97
16	TCI_PCI_VCI_SMCi	0.795	0.623	31.348	31.723	4138.21
17	TCI_PCI_NDWI_SMCi	0.795	0.669	31.391	33.512	4139.83
18	TCI_GPP_NDWI_SMCi	0.792	0.660	31.449	31.658	4142.05
19	PCI_VCI_NDWI	0.792	0.660	31.508	32.079	4144.31
20	VCI_GPP_NDWI_SMCi	0.792	0.659	31.551	33.120	4145.92
21	TCI_VCI_GPP_NDWI	0.790	0.676	31.569	33.151	4146.61
22	PCI_GPP_NDWI	0.787	0.650	31.696	32.012	4151.45
23	TCI_VCI_NDWI	0.786	0.670	31.788	32.521	4154.92
24	PCI_NDWI_SMCi	0.785	0.660	31.846	31.959	4157.09
25	GPP_NDWI_SMCi	0.784	0.658	31.897	31.977	4159.01
26	VCI_GPP_NDWI	0.783	0.655	31.952	32.948	4161.08
27	TCI_PCI_GPP	0.783	0.579	31.995	32.391	4162.72
28	TCI_VCI_GPP	0.782	0.594	32.020	31.764	4163.63
29	TCI_VCI_GPP_SMCi	0.782	0.621	32.072	33.806	4165.60
30	PCI_VCI_GPP_SMCi	0.782	0.537	32.169	32.095	4169.21
31	TCI_PCI_VCI	0.782	0.597	32.192	32.339	4170.09
32	VCI_NDWI_SMCi	0.781	0.639	32.314	33.766	4174.62
33	TCI_GPP_SMCi	0.777	0.616	32.402	34.547	4177.87
34	TCI_GPP_NDWI	0.777	0.669	32.468	31.927	4180.33
35	TCI_VCI_SMCi	0.774	0.620	32.529	33.123	4182.58
36	TCI_PCI_NDWI	0.774	0.646	32.535	33.224	4182.79
37	PCI_GPP_SMCi	0.769	0.519	32.543	33.929	4183.09
38	TCI_PCI_SMCi	0.769	0.604	32.544	33.511	4183.13
39	PCI_VCI_SMCi	0.767	0.525	32.891	34.840	4195.87
40	TCI_GPP	0.766	0.530	33.069	34.866	4202.32
41	VCI_GPP_SMCi	0.761	0.513	33.226	33.129	4208.03
42	PCI_NDWI	0.760	0.628	33.419	32.814	4214.97
43	VCI_NDWI	0.760	0.626	33.422	35.791	4215.05
44	TCI_VCI	0.758	0.582	33.540	32.928	4219.29
45	PCI_VCI_GPP	0.757	0.473	33.642	35.015	4222.93
46	TCI_NDWI_SMCi	0.751	0.648	33.659	33.634	4223.56
47	GPP_NDWI	0.750	0.604	33.675	35.167	4224.11
48	TCI_PCI	0.741	0.508	33.745	33.477	4226.60
49	TCI_SMCi	0.739	0.557	33.748	32.720	4226.71
50	NDWI_SMCi	0.738	0.615	33.815	35.161	4229.11
51	PCI_SMCi	0.727	0.460	33.826	34.560	4229.50
52	VCI_SMCi	0.727	0.467	33.842	34.089	4230.07
53	GPP_SMCi	0.721	0.427	34.026	34.544	4236.55
54	TCI_NDWI	0.720	0.626	34.513	35.799	4253.62
55	PCI_GPP	0.712	0.379	34.713	36.625	4260.56
56	PCI_VCI	0.710	0.396	34.737	35.703	4261.37
57	VCI_GPP	0.695	0.363	34.922	36.228	4267.75
58	NDWI	0.693	0.605	35.135	36.002	4275.03
59	TCI	0.689	0.523	35.197	36.965	4277.16
60	SMCi	0.671	0.415	35.533	37.124	4288.57
61	VCI	0.657	0.400	35.819	36.644	4298.19
62	GPP	0.640	0.281	35.881	37.565	4300.27
63	PCI	0.636	0.311	36.031	37.943	4305.27

R is Sperman's Correlation Coefficient, RMSE is Root Mean Square Error, AIC is Akaike Information Criteria

Table B4. CDI weights obtained from PCA of the training set.

District	PCI	TCI	VCI	SMCI	GPP	NDWI
Agacoren	0.11	0.20	0.25	0.10	0.18	0.16
Akoren	0.09	0.20	0.25	0.11	0.17	0.19
Aksaray_Merkez	0.10	0.21	0.24	0.11	0.16	0.17
Ayranci	0.10	0.19	0.23	0.08	0.20	0.19
Celtik	0.10	0.19	0.23	0.10	0.23	0.15
Ciftlik	0.13	0.20	0.24	0.16	0.09	0.18
Cumra	0.07	0.23	0.23	0.08	0.17	0.23
Derbent	0.11	0.16	0.25	0.11	0.23	0.14
Doganhisar	0.11	0.18	0.24	0.10	0.24	0.13
Emirgazi	0.10	0.21	0.22	0.13	0.15	0.18
Eregli	0.08	0.21	0.23	0.11	0.16	0.21
Eskil	0.06	0.23	0.21	0.11	0.19	0.20
Gulagac	0.14	0.23	0.20	0.18	0.01	0.24
Guneysinir	0.09	0.21	0.26	0.09	0.14	0.22
Guzelyurt	0.14	0.20	0.23	0.18	0.06	0.19
Huyuk	0.11	0.19	0.25	0.10	0.24	0.12
Kadinhani	0.09	0.19	0.23	0.10	0.23	0.16
Karaman_merkez	0.09	0.21	0.23	0.09	0.18	0.21
Karatay	0.06	0.23	0.23	0.09	0.21	0.19
Kulu	0.09	0.21	0.24	0.10	0.19	0.17
Meram	0.10	0.19	0.25	0.09	0.21	0.16
Nigde_merkez	0.12	0.21	0.22	0.13	0.13	0.19
Ortakoy	0.10	0.20	0.26	0.09	0.18	0.17
Sarayonu	0.14	0.17	0.22	0.10	0.27	0.10
Sariyahi	0.10	0.19	0.24	0.10	0.24	0.14
Seydisehir	0.11	0.19	0.24	0.12	0.23	0.11
Tuzlukcu	0.13	0.18	0.24	0.09	0.23	0.13
Ulukisla	0.13	0.20	0.20	0.15	0.11	0.21
Yalihuyuk	0.13	0.18	0.25	0.11	0.22	0.11
Yunak	0.12	0.19	0.24	0.10	0.22	0.14
Mean	0.10	0.20	0.23	0.10	0.18	0.17

Table B5. Cumulative variance explanation values of the principal components.

District	PC1	PC2	PC3	PC4	PC5	PC6
Agacoren	0.37827	0.60014	0.79012	0.89038	0.94722	1
Akoren	0.37964	0.59817	0.77144	0.88502	0.95235	1
Aksaray_Merkez	0.35659	0.60437	0.7922	0.89357	0.95284	1
Aksehir	0.38757	0.62111	0.78816	0.89806	0.9531	1
Altunhisar	0.44851	0.65333	0.83335	0.92149	0.99945	1
Altinekin	0.4282	0.64025	0.79591	0.90434	0.95385	1
Ayranci	0.40097	0.64281	0.79632	0.89818	0.95475	1
Beysehir	0.36282	0.61065	0.77973	0.89307	0.95559	1
Bor	0.37337	0.60589	0.77766	0.88352	0.94712	1
Camardi	0.33861	0.62365	0.78129	0.88327	0.95033	1
Celtik	0.41946	0.64257	0.7959	0.89283	0.95176	1
Ciftlik	0.34696	0.62536	0.79831	0.89858	0.95811	1
Cihanbeyli	0.40538	0.62014	0.79222	0.89885	0.95041	1
Cumra	0.38495	0.61096	0.79047	0.90009	0.95204	1
Derbent	0.38832	0.65965	0.80318	0.90319	0.95989	1
Doganhisar	0.39545	0.64397	0.79229	0.90881	0.9575	1
Emirgazi	0.35625	0.61432	0.77783	0.87943	0.94283	1
Eregli	0.39541	0.61914	0.79046	0.89503	0.95003	1
Eskil	0.42697	0.66389	0.82336	0.91405	0.9608	1
Gulagac	0.32869	0.59249	0.78783	0.88982	0.95349	1
Guneysinir	0.36384	0.58569	0.77724	0.88414	0.94538	1
Guzelyurt	0.32679	0.61355	0.78935	0.89317	0.95291	1
Huyuk	0.35886	0.61545	0.78083	0.90095	0.9599	1
Ilgin	0.41073	0.63481	0.79614	0.90241	0.95603	1
Kadinhani	0.43016	0.64754	0.80746	0.90301	0.95664	1
Karaman_Merkez	0.39086	0.62359	0.80394	0.90332	0.95325	1
Karapinar	0.4122	0.63506	0.81199	0.91407	0.9619	1
Karatay	0.42968	0.64214	0.80804	0.91251	0.96232	1
Kulu	0.39755	0.61013	0.78862	0.89499	0.95073	1
Meram	0.38076	0.61418	0.78257	0.89144	0.95092	1
Nigde_Merkez	0.33724	0.61227	0.79721	0.89742	0.95735	1
Ortakoy	0.38318	0.6013	0.79289	0.89361	0.94818	1
Sarayonu	0.42007	0.62269	0.79726	0.90003	0.95428	1
Sariyahsi	0.40518	0.61789	0.79633	0.89622	0.95153	1
Selcuklu	0.42175	0.63906	0.79508	0.89852	0.95584	1
Seydisehir	0.33699	0.605	0.7813	0.88593	0.95052	1
Tuzlukcu	0.41472	0.62503	0.78151	0.88928	0.94961	1
Ulukisla	0.33136	0.60431	0.77192	0.87614	0.94966	1
Yalihuuyuk	0.31934	0.57048	0.77229	0.87673	0.94937	1

PC is Principal Component

Table B6. Correlation test results between drought indices and crop yield.

District	CDI (SHAP)	CDI (PCA)	CDI (MEAN)	SPEI_1	SPEI_3	SPEI_6	SPEI_9	scPDSI
Agacoren	0.0075	-0.0391	-0.0105	0.0586	0.0812	0.1338	0.1925	0.1534
Akoren	0.4391	0.4165	0.4376	0.3624	0.3684	0.3293	0.2060	0.0632
Aksaray_Merkez	0.6812	0.6361	0.6436	0.5368	0.5744	0.5654	0.4481	0.2301
Aksehir	0.5835	0.3188	0.3895	0.3880	0.5534	0.3714	0.1925	0.2752
Altınekin	0.5474	0.3534	0.3474	0.3729	0.3774	0.4180	0.4301	0.2932
Altunhisar	0.7383	0.7188	0.7158	0.5729	0.6617	0.6737	0.6045	0.3774
Ayranci	0.2692	0.0075	0.0075	0.0902	0.1113	0.2827	0.4256	0.1338
Beysehir	0.6226	0.5910	0.6135	0.2180	0.3068	0.1519	0.1203	0.2000
Bor	0.6436	0.5744	0.5654	0.4556	0.4556	0.3353	0.2586	0.1714
Camardi	0.5038	0.3203	0.3053	0.4767	0.4286	0.3053	0.2752	0.2797
Celtik	0.7008	0.4872	0.4932	0.5053	0.6165	0.6586	0.5820	0.4797
Ciftlik	0.4075	0.4752	0.4827	0.2105	0.2586	0.2556	0.2767	0.2150
Cihanbeyli	0.7624	0.6887	0.6917	0.6692	0.6647	0.5323	0.4241	0.3940
Cumra	0.5579	0.4977	0.4797	0.3805	0.4030	0.2842	0.1158	0.1940
Derbent	0.5729	0.3038	0.3669	0.2286	0.3308	0.2271	0.1835	0.2030
Doganhisar	0.5203	0.4767	0.4165	0.3218	0.3278	0.2000	0.1940	0.2797
Emirgazi	0.8000	0.7008	0.7128	0.6015	0.6120	0.5203	0.3504	0.3233
Eregli	0.2541	0.1549	0.1669	0.0992	0.1594	0.2150	0.2000	0.0316
Eskil	0.5053	0.2677	0.3323	0.2256	0.3955	0.3820	0.3263	0.1744
Gulagac	0.7353	0.6000	0.6617	0.6090	0.6556	0.5729	0.5579	0.4060
Gunesinir	0.1820	0.1203	0.0797	0.0331	-0.0602	-0.1850	-0.2226	0.0767
Guzelyurt	0.3398	-0.0556	0.0361	0.1203	0.1624	-0.0090	-0.0165	-0.1098
Huyuk	0.6887	0.7233	0.7083	0.3850	0.4015	0.3263	0.2722	0.3008
Ilgin	0.7940	0.7594	0.7053	0.5609	0.6286	0.5564	0.4647	0.4677
Kadinhani	0.8286	0.7519	0.7398	0.4932	0.5489	0.4511	0.2917	0.2827
Karaman_merkez	0.7293	0.5489	0.5564	0.3489	0.4511	0.4211	0.4030	0.1233
Karapinar	0.6346	0.4526	0.4541	0.5368	0.7023	0.6767	0.5429	0.4571
Karatay	0.5714	0.4451	0.4677	0.5038	0.4301	0.3940	0.2556	0.2436
Kulu	0.3308	0.4015	0.4301	0.2647	0.3038	0.2060	0.0617	0.1910
Meram	0.4556	0.3654	0.3850	0.3233	0.3759	0.3459	0.3429	0.2556
Nigde_merkez	0.5474	0.4376	0.4406	0.3549	0.4406	0.3865	0.3008	0.1654
Ortakoy	0.5654	0.4436	0.5158	0.2346	0.3098	0.2647	0.2120	0.1293
Sarayonu	0.3774	0.3113	0.2872	0.0150	0.0421	-0.0917	-0.2602	-0.0962
Sarıyahsi	0.6000	0.5805	0.5383	0.5038	0.4180	0.4602	0.5053	0.4226
Selcuklu	0.6647	0.6030	0.5759	0.3774	0.3805	0.2722	0.1609	0.1504
Seydisehir	0.0015	-0.1098	-0.1143	-0.1579	-0.0105	-0.0541	-0.0060	-0.0662
Tuzlukcu	0.5579	0.5053	0.5504	0.4571	0.5188	0.3233	0.1519	0.2647
Ulukisla	0.6722	0.5925	0.6180	0.5353	0.6105	0.5353	0.3850	0.1805
Yalihuyuk	0.1278	0.0752	0.1805	-0.0045	0.0406	-0.0737	-0.1218	-0.1293
Yunak	0.6406	0.6391	0.6271	0.5910	0.6331	0.5940	0.4662	0.4541

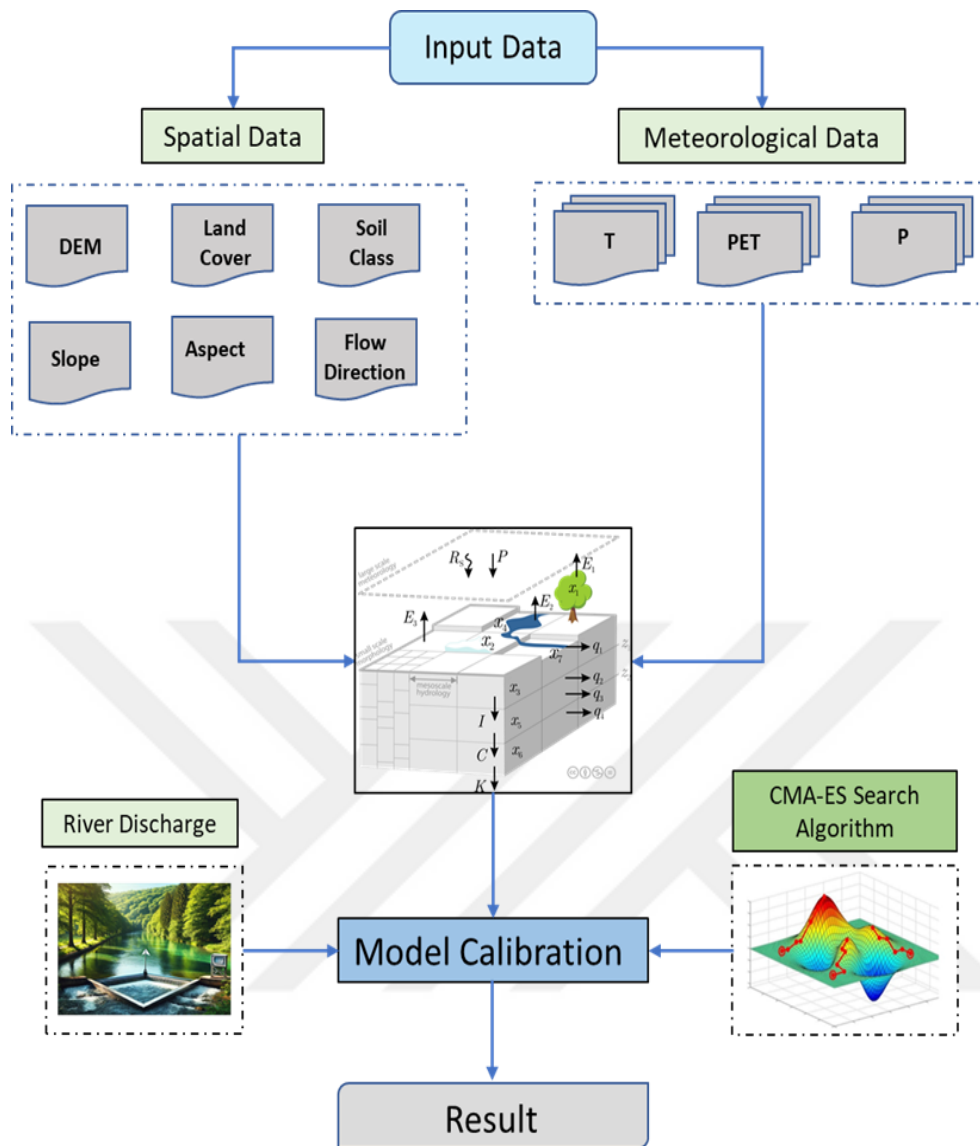


Figure B1. The mHM workflow.

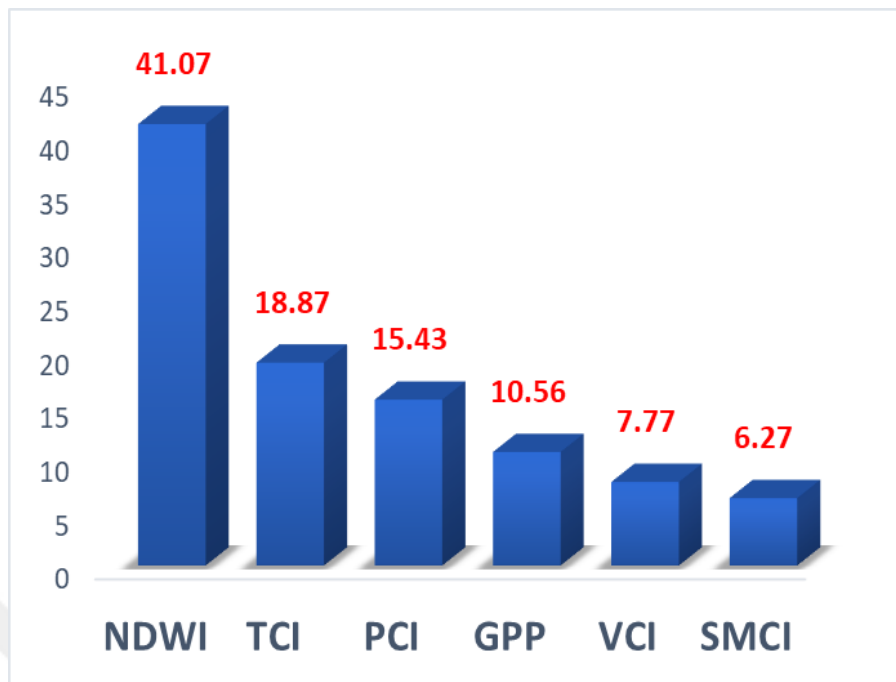


Figure B2. SHAP feature importance graph.



CURRICULUM VITAE

Name Surname : Eyyup Ensar BAŞAKIN

EDUCATION :

- **B.Sc.** : 2016, Erzincan University, Engineering Faculty, Civil Engineering Department
- **M.Sc.** : 2019, Istanbul Technical University, Graduate School Science Engineering and Technology, Hydraulics and Water Resources Engineering Programme

PROFESSIONAL EXPERIENCE AND REWARDS:

- 2018- Istanbul Technical University (Research Assistant)
- 2214-A - International Research Fellowship Programme for PhD Students (2023-2024)

PUBLICATIONS, PRESENTATIONS AND PATENTS ON THE THESIS:

- **Başakın, E. E.**, Stoy, P. C., Demirel, M. C., Ozdogan, M., and Otkin, J. A. 2024. Combined Drought Index Using High-Resolution Hydrological Models and Explainable Artificial Intelligence Techniques in Türkiye, *Remote Sensing*, 16(20), 3799. doi:10.3390/rs16203799
- **Başakın, E. E.**, Stoy, P. C., Demirel, M. C., and Pham, Q. B. 2024. Spatiotemporal Variability of Gross Primary Productivity in Türkiye: A Multi-Source and Multi-Method Assessment, *Remote Sensing*, 16(11), 1994. doi: 10.3390/rs16111994
- **Başakın, E. E.**, Ekmekcioğlu, Ö., and Özger, M. 2023. Providing a comprehensive understanding of missing data imputation processes in evapotranspiration-related research: a systematic literature review, *Hydrological Sciences Journal*, 68(14), 2089-2104. doi: 10.1080/02626667.2023.2249456
- **Başakın, E. E.**, Ekmekcioğlu, Ö., Stoy, P. C. and Özger, M. 2023. Estimation of daily reference evapotranspiration by hybrid singular spectrum analysis-based stochastic gradient boosting, *MethodsX*, 10, 102163. doi: 10.1016/j.mex.2023.102163
- **Başakın, E. E.**, Ekmekcioğlu, Ö., and Özger, M. 2023. Developing a novel approach for missing data imputation of solar radiation: A hybrid differential evolution algorithm based eXtreme gradient boosting model, *Energy Conversion and Management*, 280, 116780. doi: 10.1016/j.enconman.2023.116780

- **Başakın, E. E.,** and Özger, M. 2022. Missing Data Imputation for Solar Radiation by Deep Neural Network, *European Journal of Science and Technology*, 35, 548-555. doi: 10.31590/ejosat.1085022
- **Başakın, E.E.,** Stoy, P.C., Ranjbar, S., Demirel M.C. 2024. Gross Primary Productivity research: Ongoing trends and future trajectories, Carbon Fluxes and Biophysical, ESLEVIER [Accepted]
- **Başakın, E. E.,** and Demirel, M. C. 2024: Comparison of Satellite Rainfall Data with Local Gauge Observations Over Aksaray. 4. International Trakya Scientific Research and Innovation Congress, July 20-21, 2024. Edirne, Türkiye.
- **Başakın, E. E.,** and Özger, M. 2023: Nasa Power Sisteminde Sunulan Uzaktan Algılama Verileri Kullanılarak Referans Evapotranspirasyon Tahmini. Karadeniz 12th International Conference on Applied Sciences, March 3-5, 2023 Rize, Türkiye.
- **Başakın, E. E.,** and Özger, M. 2023: Nasa Power Sisteminde Sunulan Uzaktan Algılama Verilerinin Yer Gözlem Verileri ile Uyumunun İncelenmesi. Karadeniz 12th International Conference on Applied Sciences, March 3-5, 2023 Rize, Türkiye.
- **Başakın, E. E.,** and Özger, M. 2022: Using Neighboring Stations to Impute Missing Solar Radiation Data. 4. International Sciences and Innovation Congress, February 21-22 2022, Ankara, Türkiye.
- **Başakın, E. E.,** and Özger, M. 2022: Investigation of The Methods Utilized to Complete the Solar Radiation Data. 4. International Sciences and Innovation Congress, February 21-22 2022, Ankara, Türkiye.

OTHER PUBLICATIONS, PRESENTATIONS AND PATENTS:

- Şaylan, L., Kimura, R., **Başakın, E.E.,** and Kurosaki, Y. 2024 Variables controlling growing season carbon dynamics in desert shrub ecosystem, *Theoretical and Applied Climatology*, 155, 4797–4812. doi: 10.1007/s00704-024-04924-8
- Mouhoumed, R.M., Ekmekcioğlu, Ö., **Başakın, E.E.,** and Özger, M. 2023 Integrated Fuzzy AHP-TOPSIS Model for Assessing Managed Aquifer Recharge Potential in a Hot Dry Region: A Case Study of Djibouti at a Country Scale, *Water*, 15(14), 2534. doi: 10.3390/w15142534
- Ekmekcioğlu, Ö., **Başakın, E.E.,** Altınbaş, N., Özger, M. and Şaylan, L. 2023 A multi-perspective input selection strategy for daily net ecosystem exchange predictions based on machine learning methods, *Theoretical and Applied Climatology*, 151, 81-98. doi: 10.1007/s00704-022-04265-4
- Ekmekcioğlu, Ö., **Başakın, E.E.,** and Özger, M. 2023 Exploring the practical application of genetic programming for stormwater drain inlet hydraulic efficiency estimation, *International journal of Environmental Science and Technology*, 20, 1489–1502. doi: 10.1007/s13762-022-04035-9
- Ekmekcioğlu, Ö., **Başakın, E.E.,** and Özger, M. 2022 Developing meta-heuristic optimization based ensemble machine learning algorithms for hydraulic efficiency assessment of storm water grate inlets, *Urban Water Journal*, 19(10), 1093-1108. doi: 10.1080/1573062x.2022.2134806

- Serencam, U., Ekmekcioğlu, Ö., **Başakın, E. E.**, and Özger, M. 2022. Determining the water level fluctuations of Lake Van through the integrated machine learning methods, *International Journal of Global Warming*, 27(2), 123-142. doi: 10.1504/IJGW.2022.123278
- **Başakın, E. E.**, 2022. Comments on “Groundwater quality modeling using a novel hybrid data-intelligence model based on gray wolf optimization algorithm and multi-layer perceptron artificial neural network: a case study in Asadabad Plain, Hamedan, Iran” Cheraghi, Mehrdad et al. (10.1007/s11356-021-16300-4), *Environmental Science and Pollution Research*, 29, 41869–41871. doi: 10.1007/s11356-022-19846-z
- Altunkaynak, A., and **Başakın, E. E.**, 2022. Estimation of Wheat Yield Based on Precipitation and Evapotranspiration Using Soft Computing Methods, In: Uddin, M.S., Bansal, J.C. (eds) *Computer Vision and Machine Learning in Agriculture, Volume 2. Algorithms for Intelligent Systems*. Springer, Singapore. doi: 10.1007/978-981-16-9991-7_6
- Yiğiter, Ş. Y., Sarı, S. S., and **Başakın, E. E.**, 2021. Comparative Time Series Analysis with Sukuk Data of Companies Traded in Turkey, *The journal of Social Sciences Institute of Ataturk University*, 25(3), 450-466. doi: 10.53487/ataunisosbil.644418
- **Başakın, E.E.**, Ekmekcioğlu, Ö., Çıtakoğlu, H., and Özger, M. 2021 A new insight to the wind speed forecasting: robust multi-stage ensemble soft computing approach based on pre-processing uncertainty assessment, *Neural Computing and Applications*, 34, 783-812. doi: 10.1007/s00521-021-06424-6
- **Başakın, E.E.**, Ekmekcioğlu, Ö., Özger, M. and Çıtakoğlu, H. 2021 Determination of Intensity-Duration-Frequency Relation by Particle Swarm Optimization and Genetic Programming, Conference: II. International Applied Statistics Conference, 1-8.
- Sarı, S.S., and **Başakın, E.E.**, 2021 Analysis of Bist Bank Index with Data Based Models, *Journal of Productivity*, 3, 147-163. doi: 10.51551/verimlilik.691193
- **Başakın, E.E.**, Ekmekcioğlu, Ö., Özger, M., Altınbaş, N. and Şaylan, L. 2021 Estimation of measured evapotranspiration using data-driven methods with limited meteorological variables, *Italian Journal of Agrometeorology*, 1, 63-80. doi: 10.36253/ijam-1055
- Sarı, S.S., and **Başakın, E.E.**, 2021 Estimating of Real Sector Confidence Index with Empiricalmode Decomposition Fuzzy Logic Method, 2nd International 5 Ocak Congress on Social Sciences and Humanities, Adana, Turkey
- Ekmekcioğlu, Ö., **Başakın, E.E.**, and Özger, M. 2021 Numerical Analysis of Leakage Flowrate Parameters in a Pipe Opening to Atmosphere, *Duzce University Journal of Science and Technology*, 9(1), 265-277. doi: 10.29130/dubited.779467
- **Başakın, E.E.**, and Özger, M. 2021 Monthly River Discharge Prediction by Wavelet Fuzzy Time Series Method, *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 29(1), 17-35. doi: 10.1142/S0218488521500021

- **Başakın, E.E.**, and Ekmekcioğlu, Ö., 2021 Letter to the Editor "Estimation of global solar radiation data based on satellite-derived atmospheric parameters over the urban area of Mashhad, Iran", *Environmental Science and Pollution Research*, 28, 19530–19532. doi: 10.1007/s11356-021-13201-4
- Başakın, B.G., and **Başakın, E.E.**, 2021 Türkiye Obezite Prevalansı Değerlerinin Anfis İle Tahmin Edilmesi, 5th International Congress on Nutrition Obesity and Community Health, Istanbul, Turkey
- **Başakın, E.E.**, Ekmekcioğlu, Ö., and Özger, M. 2021 Drought prediction using hybrid soft-computing methods for semi-arid region, *Modeling Earth Systems and Environment*, 7, 2363–2371. doi: 10.1007/s40808-020-01010-6
- **Başakın, E.E.** and Ekmekcioğlu, Ö. 2020 Comment on “Comparison of the Ability of ARIMA, WNN and SVM Models for Drought Forecasting in the Sanjiang Plain, China” by Yuhu Zhang, Huirong Yang, Hengjian Cui, and Qihua Chen, in *Natural Resources Research* DOI: 10.1007/s11053-019-09512-6, *Natural Resources Research*, 29, 1465-1467. doi: 10.1007/s11053-020-09638-y
- **Başakın, E.E.**, Ekmekcioğlu, Ö., and Mohammedi, B. 2020 Letter to the editor “comparing artificial intelligence techniques for chlorophyll-a prediction in US lakes”, *Environmental Science and Pollution Research*, 27, 22131-22134. doi: 10.1007/s11356-020-08666-8
- Ekmekcioğlu, Ö., **Başakın, E.E.**, and Özger, M. 2020 Tree-based nonlinear ensemble technique to predict energy dissipation in stepped spillways, *European Journal of Environmental and Civil Engineering*, 26(8), 3547–3565. doi: 10.1080/19648189.2020.1805024
- Ekmekcioğlu, Ö., **Başakın, E.E.**, and Özger, M. 2020 Discharge coefficient equation to calculate the leakage from pipe networks, *Journal of the Institute of Science and Technology*, 10(3), 1737-1746. doi: 10.21597/jist.675015
- Taşova, M., Erdoğan, M. and **Başakın, E.E.** 2020 Determination of Drying Machine Potential by Different Estimation Methods, *Yuzuncu Yıl University Journal of the Institute of Natural and Applied Sciences*, 25(2), 49-55.
- Özger, M., **Başakın, E.E.**, Ekmekcioğlu, Ö., and Hacısüleyman, V. 2020 Comparison of wavelet and empirical mode decomposition hybrid models in drought prediction, *Computers and Electronics in Agriculture*, 179, 105851. doi: 10.1016/j.compag.2020.105851
- Sarı, S.S., and **Başakın, E.E.**, 2020 Estimating the Index of Real Estate Investment Trusts with Tree Based Regression Models, 5. Uluslararası Marmara Sosyal Bilimler Kongresi, Kocaeli, Turkey
- Altunkaynak, A., **Başakın, E.E.**, and Kartal, E., 2020 Air Pollution Prediction with Wavelet K-Nearest Neighbour Method, *Uludağ University Journal of The Faculty of Engineering*, 25(3), 1547-1556. doi: 10.17482/uumfd.809938
- **Başakın, E.E.**, Ekmekcioğlu, Ö., Özger, M., and Çelik, A., 2020 Prediction of Turkey Wheat Yield by Wavelet Fuzzy Time Series and Gray Prediction Methods, *Turkish Journal of Agricultural Research*, 7(3), 246-252. doi: 10.19159/tutad.685342
- **Başakın, E. E.**, Özger, M. and Ünal, N. E. 2019. Water Consumption Model of Istanbul City by Gray Prediction Method, *Journal of Polytechnic*, 22(3), 755-761. doi: 10.2339/politeknik.422727

- **Başakın, E. E.,** Ekmekcioğlu, Ö., **Kartal, E.,** and Özger, M. 2019. Prediction of Air Pollution by Machine Learning Methods, 3rd International UNIDOKAP Black Sea Symposium “Sustainable Agriculture and Environment”, July 2019 Tokat, Türkiye
- Erdoğan, M., Yürekli, K., **Başakın, E. E.,** and Cömert, M. M. 2019. Estimate of Tokat Province Drought with Adaptive Neuro-Fuzzy Inference System, 3rd International UNIDOKAP Black Sea Symposium “Sustainable Agriculture and Environment”, July 2019 Tokat, Türkiye
- **Başakın, E. E.,** Sağ, A.B., Kartal, E., Avcuoğlu, M.A. and Demirel, M.C. 2019. Calibration of Conceptual Hydrological Model Parameters using Heuristic Algorithms, 3rd International UNIDOKAP Black Sea Symposium “Sustainable Agriculture and Environment”, July 2019 Tokat, Türkiye
- Özger, M., **Başakın, E.E.,** Ekmekcioğlu, Ö., and Demirel, E. 2019 Günlük Akımların Görgül Kip Ayırışimli Bulanık Mantık Modeli ile Tahmini, Uluslararası Marmara Fen ve Sosyal Bilimler Kongresi (Bahar), June 2019 Kocaeli, Türkiye
- **Başakın, E. E.,** Ekmekcioğlu, Ö., and Özger, M. 2019. Drought analysis with machine learning methods, *Pamukkale University Journal of Engineering Sciences*, 25(8), 985-991. doi: 10.5505/pajes.2019.34392
- Demirel, M. C., Özen, A., Orta, S., Toker, E., Demir, H.K. Ekmekcioğlu, Ö., Tayşi, H., Eruçar, S., Sağ, A.B., Sarı, Ö., Tuncer, E, Hancı, H., Özcan, T.İ., Erdem, H., Koşucu, M.M., **Başakın, E.E.,** Ahmed, K., Anwar, A., Avcuoğlu, M.B., Vanlı, Ö., Stisen, S. and Booij, M.J. 2019 Additional value of using satellite-based soil moisture and two sources of groundwater data for hydrological model calibration, *Water*, 11(10), 2083. doi: 10.3390/w11102083
- Altunkaynak, A. and **Başakın E.E.** 2018 Prediction of River Flow Using Time Series and Comparison with Different Methods, *Erzincan University Journal of Science and Technology*, 11(1), 92-101. doi: 10.18185/erzifbed.339781
- **Başakın, E. E.,** Sarı, S.S. and Ekmekcioğlu, Ö. 2018. Prediction of The Electricity Consumption of Turkey by Using Machine Learning Methods, International Congress of Energy, Economy and Security, April 2018 Istanbul, Türkiye
- Ekmekcioğlu, Ö., **Başakın, E. E.,** and Özger, M. 2018. Modeling of Efficiency of Grate Inlets by Using Modern Modelling Techniques, International Symposium on Urban Water and Wastewater Management, October 2018 Denizli, Türkiye
- Özer, A., Sarı, S.S., and **Başakın, E. E.,** 2018. Stock Market Index Prediction with Artificial Neural Networks and Fuzzy Logic: Example of Developed and Developing Countries, *Journal of the Social Sciences Institute of Hittite University*, 11(1), 99-124. doi: 10.17218/hititsosbil.390490
- Yiğiter, Ş. Y., Sarı, S. S., Karabulut, T. and **Başakın, E. E.,** 2018. Estimation of Lease Certificate Price Evaluation Through Machine Learning Method, *International Journal of Islamic Economics and Finance Studies*, 4(3), 74-82. doi: 10.25272/ijisef.412760

- **Başakın, E. E.**, and Sarı, S.S. 2018. Mühendislik ve Sosyal Bilimlerde Bulanık Mantık Yaklaşımı, 31. Ulusal Matematik Sempozyumu, September 2018 Erzincan, Türkiye
- Yiğiter, Ş. Y., Sarı, S. S., Karabulut, T. and **Başakın, E. E.**, 2017. Estimation of Rent Certificate Values by Using K- Nearest Neighborhood Algorithm, International Congress Onpolitical, Economic and Socialstudies, November 9-11, 2017 Anakara, Türkiye
- Yiğiter, Ş. Y., Sarı, S. S. and **Başakın, E. E.**, 2017. Estimations of Closing Stock Prices Through Artificial Neural Networks and Fuzzy Inference Systems, *Kahramanmaraş Sütçü İmam Üniversitesi İktisadi Ve İdari Bilimler Fakültesi Dergisi*, 7(1) 1-22.

