Air Temperature-based Spatial Modeling of Remote Sensing Data Using Machine Learning Approaches: a Systematic Literature Review

David Sampelan[a, b,*], Anggitya Pratiwi[a], Anas Baihaqi[a, c], Suci Agustiarini[a]
[a]Climatological Station of West Nusa Tenggara, BMKG, West Nusa Tenggara, 83362, Indonesia,
[b]Department of Geomatics Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, 60111, Indonesia,
[c]Department of Mathematics, Institut Teknologi Sepuluh Nopember, Surabaya, 60111, Indonesia

*Corresponding author: david.sampelan@bmkg.go.id

Abstract. This study presents a systematic review of spatial air temperature modeling based on remote sensing data using machine learning approaches during the period 2016–2025. Using the PRISMA framework, we conducted literature searches in Google Scholar (998 articles) and Scopus (489 articles).. After merging the datasets, removing duplicates, and applying inclusion–exclusion criteria, 12 articles were retained for in-depth analysis. The findings indicate a marked increase in publications since 2021, reflecting growing global interest in integrating remote sensing and machine learning for air temperature estimation. Ensemble algorithms such as Random Forest and XGBoost dominate due to their balance of accuracy and computational efficiency, while temporal deep learning approaches such as LSTM and TCN are emerging as powerful tools for capturing complex atmospheric dynamics. Among remote sensing predictors, Land Surface Temperature (LST) is the most frequently used, often complemented by NDVI, albedo, and elevation to improve spatial accuracy. Geographical context strongly influences methodological performance. XGBoost proves effective in heterogeneous urban areas, Random Forest performs well in mountainous regions, and artificial neural networks demonstrate higher adaptability in extreme environments such as the Greenland ice sheet. Nonetheless, limited ground-based observations and sparse station networks remain key challenges, particularly across tropical and archipelagic regions. This review identifies three major directions for future research: (1) expanding studies to underrepresented tropical regions, (2) leveraging temporal deep learning methods for detecting extreme events, and (3) integrating multisensor data with innovative validation strategies to enhance the robustness and reliability of air temperature modeling.

Keywords: Air Temperature, Land Surface Temperature, Machine Learning, Remote Sensing, PRISMA.

I. INTRODUCTION

Air temperature is a key climatic parameter with significant relevance for environmental studies, public health, and spatial planning [1], [2]. However, weather station data are often limited in both number and spatial coverage, making them insufficient to capture fine-scale variability—particularly in tropical archipelagic regions [1], [3], [4], [5], [6]. Remote sensing provides an alternative with wide and continuous spatial coverage, yet the relationship between satellite-derived variables and air temperature is complex, requiring more adaptive analytical approaches. In this context, machine learning has emerged as a promising method, as it can model non-linear relationships and improve the accuracy of air temperature estimation. Nevertheless, challenges remain, including variable selection, model adaptation to tropical climates, and limitations in ground-based validation [2], [3], [4], [6].

Unlike previous reviews that predominantly focus on temperate regions, this study highlights tropical and archipelagic contexts, which remain underexplored. By integrating the PRISMA framework [7] with a broad literature search from Scopus and Google Scholar (2016–2025), this review analyzes publication trends, the dominant machine learning algorithms, and the key remote sensing variables applied in air temperature modeling. Furthermore, it emphasizes the potential of temporal deep learning approaches and multisensor data integration as future research directions to achieve more accurate, context-specific, and climate-relevant air temperature models.

The research questions (RQ), as outlined in Table 1, were formulated to ensure coherence between the background, objectives, and the systematic review approach applied in this study.

TABLE 1. RESEARCH QUESTION

Research Question	Objective	Method
What are the research trends in air temperature modeling (2016–2025)?	Describe publication trends, domains, and contexts	Trend and context analysis
Which ML algorithms are most widely applied, with their strengths and limitations?	Evaluate effectiveness of ML methods	Algorithm categorization & comparison
Which remote sensing variables are most used, and what are the main validation challenges?	Assess relevance of variables and validation limits	Input analysis & critical review

II. METHODOLOGY

2.1 Search Strategy

The initial stage of this review was conducted through a systematic literature search in two primary databases: Scopus and Google Scholar (the latter accessed via *Publish or Perish* software). These databases were selected because of their complementary strengths: Scopus applies rigorous peer-review standards for article inclusion, while Google Scholar offers broader coverage, including recent studies that may not yet be indexed in Scopus [8], [9], [10], [11], [12].

The search was carried out on 3 September 2025 using a combination of keywords in both English and Indonesian: ("air temperature" OR "suhu udara") AND ("remote sensing" OR "penginderaan jauh") AND ("machine learning"). The use of bilingual terms ensured that the search captured literature in multiple languages and reflected the diverse terminology employed by researchers in climatology and remote sensing.

The search scope differed across databases: Scopus covered publications from 2016–2025 to capture medium-term trends, while Google Scholar was limited to 2021–2025 to retrieve the most recent studies that may not yet appear in Scopus. All results were exported in CSV/RIS format, merged, and cleaned of duplicates before being screened according to the predefined inclusion and exclusion criteria. The number of articles retrieved from each database is presented separately in the PRISMA flow diagram (Figure 3), prior to being consolidated into the final dataset for analysis.

To illustrate the identification and selection process, screenshots of the search results from *Publish or Perish* (Google Scholar, Figure 2) and Scopus (Figure 1) are provided. These visuals show the study counts and the application of inclusion – exclusion criteria prior to screening.

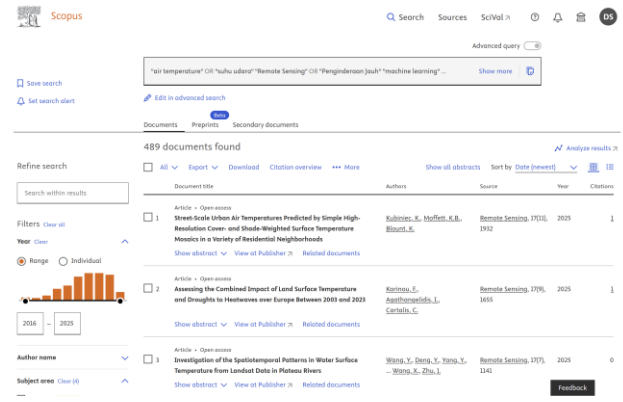

Figure 1. Screenshot of literature search results retrieved from the Scopus website, showing the total number of articles identified before duplicate removal and screening.

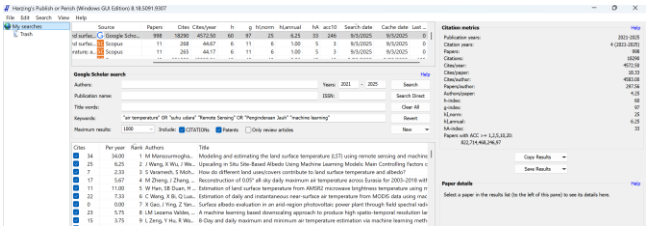

Figure 2. Screenshot of literature search results retrieved from the Google Scholar using Harzing's Publish or Perish, showing the total number of articles identified before duplicate removal and screening.

2.2 Selection Criteria

The inclusion and exclusion criteria were carefully defined to ensure that only literature directly relevant to the research objectives was retained.

Inclusion criteria :

1. Studies that explicitly address air temperature modeling using machine learning with remote sensing data.
2. Articles published in peer-reviewed international journals, reputable international conference proceedings, or relevant review papers.

3. Publication years aligned with database coverage: 2016–2025 for Scopus and 2021–2025 for Google Scholar.
4. Clear methodological description, including the machine learning algorithms applied.
5. Full-text availability for comprehensive analysis.

Exclusion criteria :

1. Grey literature such as technical reports, books, dissertations, or theses.
2. Articles mentioning air temperature only in general terms without explicit use of machine learning and remote sensing.
3. Studies inaccessible in full text.

2.3 Screening Process

The screening process was conducted in multiple stages, beginning with title and abstract review, followed by full-text assessment to ensure relevance. Prior to screening, duplicate records were removed using DOI and title (case-insensitive) matching in Zotero. Automatically detected duplicates were further verified manually to avoid residual redundancy. Only articles meeting all inclusion criteria were advanced to the data extraction stage. The PRISMA flow diagram (Figure 3) illustrates the step-by-step screening process, including the number of records retained or excluded at each stage and the reasons for exclusion.

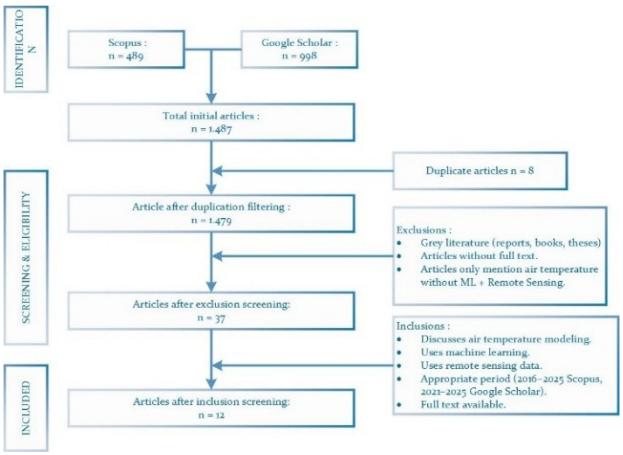

Figure 3. PRISMA flow diagram of the literature search and screening process

2.4 Data Extraction

From the initial 1,459 publications, only 12 studies met all inclusion criteria, namely a focus on air temperature modeling using remote sensing–based machine learning, availability in full text, and explicit reporting of methods and performance metrics. Although relatively small in number, this sample is still representative as it covers diverse geographical contexts (agricultural, urban, mountainous, and polar regions), multiple sensors (MODIS, Landsat, Sentinel, ERA5, and geostationary data), and a range of analytical approaches (RF, XGBoost, ANN, temporal deep learning, and ML–geostatistical hybrids). For each selected article, key information was extracted, including authorship and publication year, research objectives, applied machine learning algorithms, remote sensing input variables, and main findings. A standardized extraction form was employed to ensure consistency and comparability across studies.

2.5 Quality assessment of the literature

To evaluate the strength of evidence, each study was assessed in terms of methodological clarity, adequacy of data, and transparency of reporting. The assessment adapted the PROBAST framework across four domains: data, modeling, validation, and metric reporting [13], [14], [15], [16]. This approach provided a more systematic evaluation than purely narrative descriptions, while keeping the emphasis on methodological trends rather than statistical meta-analysis.

In general, data quality ranged from very limited (e.g., only four in-situ stations) to massive datasets comprising millions of samples. Modeling approaches revealed a shift from regression-based methods and Random Forest toward ML–geostatistical hybrids and temporal deep learning. Validation strategies were commonly based on k-fold cross-validation, though some studies relied on simple train–test splits or external validation by station or temporal series. Reporting of performance metrics was relatively consistent, with RMSE, MAE, and R^2 as the main indicators, while several studies added NSE, KGE, precision, recall, and F1-scores to suit specific application contexts.

TABLE 2. SUMMARY OF RISK OF BIAS ASSESSMENT (ADAPTED FROM PROBAST) ACROSS THE 12 INCLUDED STUDIES

Domain	Low Risk	High Risk	Unclear
Data	9	3	0
Modeling	8	4	0
Validation	7	5	0
Metric reporting	10	2	0

A summary of the risk of bias assessment is presented in Table 2, while the full table with detailed information for each study (including data quality, applied algorithms, validation techniques, and reported metrics) is provided in Appendix A1.

In Table 2, ‘Low Risk’ refers to transparent and well-documented validation procedures, regardless of whether they were based on random k-fold cross-validation or hold-out methods. Nonetheless, we acknowledge that spatial or temporal hold-out strategies provide more realistic accuracy than random cross-validation, which may lead to overoptimistic estimates.

2.6 Comparative Performance of Models

To complement the quality assessment, a comparative analysis of model performance was conducted based on the reviewed studies. The summary highlights study area, data sources, algorithms, validation strategies, and key performance metrics (RMSE, MAE, R^2), enabling cross-study comparison of algorithm effectiveness and the influence of validation design. The complete performance table for all studies is provided in Appendix A2.

Overall, Random Forest (RF) remained dominant, consistently achieving strong results across diverse contexts [3], [17], [18], [19], [20], while XGBoost demonstrated superior performance in urban environments [21], [22], [23], [24], Temporal deep learning approaches such as LSTM, TCN, and N-BEATS were particularly effective for large datasets with strong temporal dynamics [1], [25]. Hybrid ML–geostatistical models also proved successful for temperature downscaling [21], [26], [27]. In terms of validation, most studies relied on random cross-validation, which tends to yield optimistic results, whereas spatial or temporal hold-out approaches produced lower but more realistic accuracy for predicting new locations [28], [29], [30], [31].

III. RESULT AND DISCUSSION

3.1 Literature Review

The selected studies reveal substantial diversity in geographical contexts, applied machine learning algorithms, and remote sensing variables for air temperature modeling. Each study tailored its methodological design and input variables to the specific characteristics of its setting, ranging from agricultural landscapes and densely populated urban areas to extreme environments such as the Greenland Ice Sheet [32]. Table 3 summarizes the key articles, including author, year, study location, methods, predictor variables, and main findings, thereby providing a comprehensive overview of the methodological variations and the relative effectiveness of different algorithms in estimating air temperature from remote sensing data.

3.2 Case Studies

To illustrate how machine learning approaches are applied under diverse geographical conditions, several representative studies were selected. These examples highlight the range of challenges encountered, from densely populated urban environments to mountainous regions with complex topography, as well as extreme ecosystems such as the Greenland Ice Sheet. The following summaries emphasize the geographical setting, algorithms employed, and key findings of each study, providing a concrete picture of the effectiveness and limitations of remote sensing–based machine learning methods for air temperature modeling.

1. In urban areas, Hassani et al. (2024) [33] compared XGBoost with the numerical WRF model and Kriging interpolation for temperature mapping in Warsaw, Poland. Their results demonstrated that XGBoost delivered the best performance (RMSE = 1.06 °C), significantly outperforming the other methods. This finding highlights the potential of machine learning as a more accurate and computationally efficient alternative for urban temperature estimation.
2. In mountainous environments, Pradhan et al. (2024) [34] evaluated temperature modeling in Himachal Pradesh, India, where complex topography poses major challenges. The study reported that Multilayer Perceptron (MLP) achieved high accuracy (RMSE = 1.13 °C at Bhuntar station), while also revealing a substantial bias in ERA5 reanalysis data, underscoring the importance of local verification. Similar observations were made by Joy et al. (2025) [35], who found that model errors (RMSE) tend to be higher in mountainous and arid regions.
3. Under extreme environmental conditions, Che et al. (2022) [32] investigated the Greenland Ice Sheet and found that Neural Networks (NN) were the most reliable approach, achieving R = 0.96 with RMSE = 2.67 °C, outperforming other ML algorithms. The study emphasized the critical role of albedo as a predictor variable, showing that its exclusion significantly reduced model accuracy.

3.3 Publication Trends in Air Temperature Modeling Studies

The analysis of the selected articles (Figure 4) reveals clear patterns in publication output, methodological diversity, and the remote sensing variables employed for air temperature modeling. As shown in Figure 4, the annual trend indicates a steady increase since 2021, with a marked surge during 2022–2024 [35], [36], [37]. This growth highlights the rising prominence of integrating remote sensing and machine learning in air temperature estimation, confirming it as an emerging area of focus in recent scientific literature.

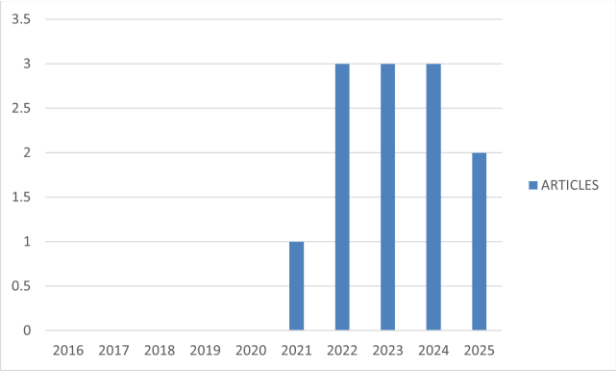

Figure 4. Publication trends in air temperature modeling studies (2016–2025)

3.4 Distribution of Machine Learning Algorithms

The distribution of algorithms (Figure 5) shows the dominance of Random Forest, which was applied in 60% of the reviewed studies [3], [38], [39], followed by XGBoost and Artificial Neural Networks (ANN), each used in approximately 30% of cases [33], [35]. These three algorithms emerged as the primary choices due to their capacity to handle high-dimensional datasets while delivering stable accuracy across diverse geographical contexts.

TABLE 3. SELECTED STUDIES BASED ON SCREENING CRITERIA

Author & Year	Geography	ML Methods	Key Variables	Main Findings
Xu et al., 2023 [38]	Henan, China (wheat-growing, dense vegetation)	TVX + RF, DBN, MLR	LST, NDVI, soil moisture, albedo, elevation	Integrating TVX with RF achieved best performance (R^2 = 0.971, RMSE = 1.62 °C) in agricultural areas.
Andriambololonaharisoamalala et al., 2025 [40]	Perth, Australia (heterogeneous urban)	GeoML (RF + Kriging)	Landsat LST, ECOSTRESS LST, NDVI, NDBI, NDWI, albedo, elevation	Hybrid GeoML outperformed alternatives (Pearson r = 0.85, RMSE = 2.7 °C, MAE < 2.2 °C) for urban downscaling.
Karagiannidis et al., 2025 [37]	Greece (complex topography, sparse stations)	RF, XGBoost, ANN	All-sky LST, heat fluxes, albedo, slope, curvature, elevation, lat/lon	RF with 20 predictors was most efficient (MAE = 0.96 °C, R^2 = 0.976). Performance less sensitive to topography/clouds.
Jangho Lee, 2025 [41]	Illinois, USA (mixed agriculture–urban, temperate)	LSTM, TCN, N-BEATS	GOES LST, NDVI, NDBI, temporal scales (month, hour, lag)	Temporal models outperformed static ones (RMSE reduced from 2.6–2.8 °C to <1.8 °C). TCN best for extremes (highest F1).
Wang et al., 2022 [3]	Jingjinji, China (complex terrain, urbanized)	RF, DT, FNN, GLM, SVM	LST, elevation, solar declination, NDVI, land cover, DSR	RF best algorithm (daily RMSE = 1.29 °C). LST, solar declination, and DSR were key predictors.
Joy et al., 2025 [35]	India (varied: mountains, plains, coastal)	XGBoost, ANN, GAM, MLR	LST, Julian day, elevation, land cover, coastal distance, lat/lon	XGBoost achieved best accuracy (RMSE = 1.79 °C, R^2 = 0.90). LST, Julian day, elevation most influential.
Che et al., 2022 [32]	Greenland Ice Sheet (extreme environment)	NN, GPR, SVM, RF	MODIS LST, albedo, wind speed, humidity, elevation, lat/lon, month	NN was most accurate (r = 0.96, RMSE = 2.67 °C). Albedo was a critical predictor.
Zheng et al., 2022 [36]	Eurasia (large-scale, diverse ecosystems)	HGB, RF, ET, DBN	Day/night LST, radiation, LAI, EVI, albedo, DEM	HGB was most accurate (R^2 = 0.984, RMSE = 1.74 °C). Regional modeling improved by dividing Eurasia into seven homogeneous zones.
Buo et al., 2021 [39]	Estonia (temperate, mixed landscapes)	RF	Land cover, elevation, NDVI, X–Y coordinates, DOY	RF effective for gap-filling cloudy LST (RMSE = 1.37 °C). Time and vegetation were key predictors.
Babaei et al., 2023 [42]	Lake Urmia, Iran (shrinking hypersaline lake)	RF, XGBoost, SVM, KNN, etc.	LST, NDVI, NDWI, NDBI, climate data	ML predicted future LST, showing lake's cooling effect declined sharply (water level drop from 1050 m in 1998 to 90 m in 2023).
Hassani et al., 2024 [33]	Warsaw, Poland (urban)	XGBoost, WRF, OK	LST, ERA5 Ta, calendar	XGBoost outperformed WRF and Kriging (RMSE = 1.06 °C, R^2 = 0.94). ML viable alternative to numerical models.
Pradhan et al., 2024 [34]	Himachal Pradesh, India (mountainous)	MLP, GAM, MLR	LST, DEM, SAA, SE, IL	MLP achieved best performance (RMSE = 1.13 °C at Bhuntar station). ERA5 reanalysis showed large bias in mountains.

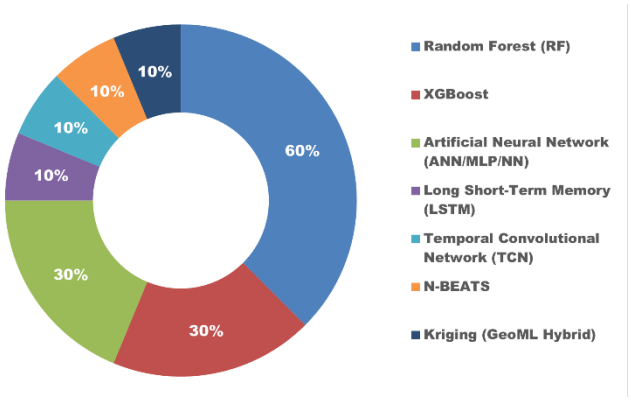

Figure 5. Distribution of machine learning algorithms used in the selected studies

3.5 Remote Sensing Variables as Key Predictors

The frequency analysis (Figure 6) highlights Land Surface Temperature (LST) as the most widely used input variable, appearing in nearly all studies (100%) [3], [36], [38], confirming its critical role in air temperature modeling. Other variables such as NDVI (58%) [35] and elevation/DEM (50%) [37] were also prominent, reflecting the importance of integrating vegetation and topographic parameters to improve estimation quality. These findings emphasize that although LST is the primary predictor, the inclusion of supplementary variables remains essential to enhance spatial representation.

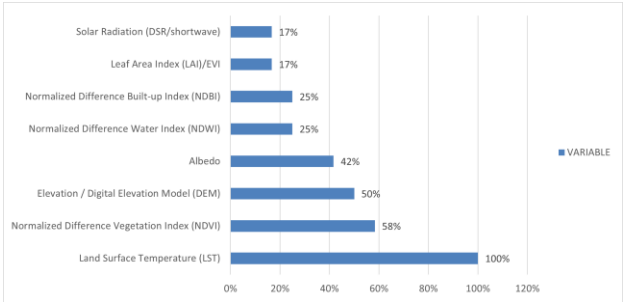

Figure 6. Distribution of remote sensing predictor variables used in the selected studies

TABLE 4. COMPARISON OF MACHINE LEARNING METHODS AND GEOGRAPHICAL CONTEXTS

Geographical Context	Algorithms Applied	Rationale / Strengths	Representative Studies
Urban (heterogeneous, mixed land use)	XGBoost, GeoML (RF + Kriging)	Accurately captures high spatial variability; computationally efficient	Hassani et al., 2024 (Warsaw); Andriambololonaharisoamalala et al., 2025 (Perth)
Mountainous (complex topography, sparse stations)	Random Forest, ANN/MLP	Robust to topographic variables; capable of modeling non-linearities	Pradhan et al., 2024 (India); Karagiannidis et al., 2025 (Greece)
Agricultural/Plains	Random Forest, MLR + TVX	Well-suited for dense vegetation; interpretable with vegetation and climate predictors	Xu et al., 2023 (Henan, China)
Extreme environments (ice sheets, shrinking lakes)	ANN/NN, RF–XGBoost combinations	Adaptive to unique conditions; effective at capturing complex interactions	Che et al., 2022 (Greenland); Babaei et al., 2023 (Lake Urmia)
Large-scale / continental (diverse ecosystems)	RF, HGB, DBN	Effective for multi-ecosystem data; high accuracy at broad scales	Zheng et al., 2022 (Eurasia)
Temporal analysis (dynamic prediction)	LSTM, TCN, N-BEATS	Captures time-series patterns and extreme events	Lee, 2025 (Illinois, USA)

3.6 Comparison of Methods and Contexts

As shown in Table 4 the literature review indicates that no single machine learning algorithm consistently outperforms others across all geographical settings. Random Forest (RF) emerged as the most frequently applied method due to its robustness in handling high-dimensional datasets and its capability to assess variable importance [3], [38]. However, its main limitation lies in the inability to capture temporal dynamics, which reduces its performance in short-term forecasting and the detection of extreme events [41].

In contrast, XGBoost, applied in around 30% of the studies, demonstrated both computational efficiency and high accuracy, particularly in regions with high spatial complexity such as India and Poland [33], [35]. It proved effective in detecting extreme variations, including daily maximum temperature [35], although interpreting variable importance is more challenging compared to RF.

Neural network–based models, such as Artificial Neural Networks (ANN), Multilayer Perceptron (MLP), and Neural Networks (NN), were widely employed in areas with strong non-linear dynamics, such as Greenland [32]. These models excel at capturing complex interactions among predictors but require large datasets and face risks of overfitting if not properly optimized [36]. Meanwhile, temporal deep learning methods—Long Short-Term Memory (LSTM), Temporal Convolutional Networks (TCN), and N-BEATS—have only been applied in a single study so far [41], yet results were highly promising, reducing RMSE by up to 30% compared with static models, especially in detecting extreme events.

Geographical context also strongly influenced methodological choices. Urban areas often relied on XGBoost and hybrid approaches such as GeoML [40] which are better suited to capturing land-use heterogeneity. In mountainous regions like India and Greece, RF and ANN were favored [34], [37] to address sparse station coverage and topographic complexity. Under extreme environments such as the Greenland Ice Sheet [32] and the hypersaline Lake Urmia in Iran [42], ANN and RF–XGBoost combinations proved more adaptive in modeling unique ecological conditions that are difficult to capture with traditional approaches.

Overall, RF and XGBoost can be considered the backbone of spatial air temperature modeling, offering a balance between accuracy, stability, and interpretability [3], [35], [38]. Nevertheless, recent trends suggest that temporal deep learning models are emerging as a promising direction for future research, particularly for capturing rapidly changing atmospheric dynamics in tropical and extreme environments [41].

3.7 Research Gaps and Future Directions

The literature review reveals that although air temperature modeling based on remote sensing and machine learning has advanced considerably, several research gaps remain. First, most studies have focused on temperate and subtropical regions [3], [37], while tropical and archipelagic areas such as Southeast Asia remain underrepresented. This highlights the need to develop models tailored to the highly dynamic conditions of tropical climates.

Second, algorithm use is still dominated by ensemble models such as Random Forest and XGBoost [35], [38]. While accurate, studies employing temporal deep learning approaches (e.g., LSTM, TCN, N-BEATS) remain scarce, despite their strong potential for capturing time-series dynamics and extreme events, especially in regions vulnerable to rapid climate change.

Third, predictor variables have been largely limited to Land Surface Temperature (LST), NDVI, and topographic parameters [35], [36]. Integration with other variables now available from new satellite missions, such as soil moisture, aerosol concentration, or high-resolution radiation data remains underexplored. Multisensor and multiresolution data fusion is therefore an important avenue to improve modeling accuracy.

Finally, validation practices remain a common weakness. Limited distribution of weather stations often introduces bias in model evaluation. Strengthening validation strategies through alternative sources, such as IoT sensors, community-based monitoring networks, or locally corrected reanalysis data should be prioritized.

Looking forward, future research can be directed along three main pathways: (1) expanding study coverage to tropical and archipelagic regions, (2) harnessing the potential of deep learning for temporal modeling and extreme event detection, and (3) integrating multisensor data and innovative validation strategies to produce air temperature mapping that is more representative, accurate, and applicable for climate change adaptation.

IV. CONCLUSION

This systematic review demonstrates that spatial air temperature modeling using remote sensing and machine learning has progressed rapidly over the period 2016–2025. Publication trends reveal a marked increase since 2021, reflecting growing scientific attention to this field. Tree-based algorithms, particularly Random Forest and XGBoost, continue to dominate due to their balance of high accuracy and computational efficiency, while temporal deep learning approaches such as LSTM and TCN are beginning to show considerable promise for capturing complex atmospheric dynamics.

With respect to predictor variables, Land Surface Temperature (LST) has consistently been the most critical input, often combined with vegetation indices (NDVI), albedo, and topographic parameters to improve spatial representation. However, the reliance on traditional predictors suggests untapped potential for integrating emerging multisensor data, including soil moisture, aerosols, and high-resolution radiation products.

Cross-context comparisons indicate that algorithm selection is closely shaped by local conditions, XGBoost tends to outperform in heterogeneous urban settings, Random Forest and ANN are more stable in mountainous areas, and Neural Network models show greater adaptability in extreme environments. Nevertheless, validation remains a key challenge, particularly due to the sparse distribution of weather stations in tropical and archipelagic regions.

Overall, this review highlights three main directions for future research: (1) expanding coverage to underexplored tropical regions, (2) integrating temporal deep learning approaches to better capture extreme temperature dynamics, and (3) leveraging multisensor data and innovative validation strategies. Addressing these gaps will enable air temperature modeling to become more accurate, context-specific, and applicable for supporting climate change adaptation efforts.

REFERENCES

[1] H. Shen, Y. Jiang, T. Li, Q. Cheng, C. Zeng, and L. Zhang, "Deep learning-based air temperature mapping by fusing remote sensing, station, simulation and socioeconomic data," *Remote Sens Environ*, vol. 240, p. 111692, Apr. 2020, doi: 10.1016/j.rse.2020.111692.

[2] P. Su, T. Abera, Y. Guan, and P. Pellikka, "Image-to-Image Training for Spatially Seamless Air Temperature Estimation With Satellite Images and Station Data," *IEEE J Sel Top Appl Earth Obs Remote Sens*, vol. 16, pp. 3353–3363, 2023, doi: 10.1109/JSTARS.2023.3256363.

[3] C. Wang, X. Bi, Q. Luan, and Z. Li, "Estimation of Daily and Instantaneous Near-Surface Air Temperature from MODIS Data Using Machine Learning Methods in the Jingjinji Area of China," *Remote Sens (Basel)*, vol. 14, no. 8, p. 1916, Apr. 2022, doi: 10.3390/rs14081916.

[4] D. Parsons, D. Stern, D. Ndanguza, and M. B. Sylla, "Evaluation of Satellite-Based Air Temperature Estimates at Eight Diverse Sites in Africa," *Climate*, vol. 10, no. 7, p. 98, Jun. 2022, doi: 10.3390/cli10070098.

[5] S. Liu, H. Su, J. Tian, and W. Wang, "An analysis of spatial representativeness of air temperature monitoring stations," *Theor Appl Climatol*, vol. 132, no. 3–4, pp. 857–865, May 2018, doi: 10.1007/s00704-017-2133-6.

[6] I. A. Adeniran, M. Nazeer, M. S. Wong, and P.-W. Chan, "An improved machine learning-based model for prediction of diurnal and spatially continuous near surface air temperature," *Sci Rep*, vol. 14, no. 1, p. 27342, Nov. 2024, doi: 10.1038/s41598-024-78349-8.

[7] M. J. Page *et al.*, "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews," *BMJ*, p. n71, Mar. 2021, doi: 10.1136/bmj.n71.

[8] M. E. Falagas, E. I. Pitsouni, G. A. Malietzis, and G. Pappas, "Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses," *The FASEB Journal*, vol. 22, no. 2, pp. 338–342, Feb. 2008, doi: 10.1096/fj.07-9492LSF.

[9] A. Martín-Martín, M. Thelwall, E. Orduna-Malea, and E. Delgado López-Cózar, "Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: a multidisciplinary comparison of coverage via citations," *Scientometrics*, vol. 126, no. 1, pp. 871–906, Jan. 2021, doi: 10.1007/s11192-020-03690-4.

[10] A. Martín-Martín, E. Orduna-Malea, M. Thelwall, and E. Delgado López-Cózar, "Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories," *J Informetr*, vol. 12, no. 4, pp. 1160–1177, Nov. 2018, doi: 10.1016/j.joi.2018.09.002.

[11] A.-W. Harzing and S. Alakangas, "Google Scholar, Scopus and the Web of Science: a longitudinal and cross-disciplinary comparison," *Scientometrics*, vol. 106, no. 2, pp. 787–804, Feb. 2016, doi: 10.1007/s11192-015-1798-9.

[12] A. Martín-Martín, E. Orduna-Malea, M. Thelwall, and E. Delgado López-Cózar, "Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories," *J Informetr*, vol. 12, no. 4, pp. 1160–1177, Nov. 2018, doi: 10.1016/j.joi.2018.09.002.

[13] K. G. M. Moons *et al.*, "PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods," *BMJ*, p. e082505, Mar. 2025, doi: 10.1136/bmj-2024-082505.

[14] R. F. Wolff *et al.*, "PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies," *Ann Intern Med*, vol. 170, no. 1, pp. 51–58, Jan. 2019, doi: 10.7326/M18-1376.

[15] K. G. M. Moons *et al.*, "PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration," *Ann Intern Med*, vol. 170, no. 1, pp. W1–W33, Jan. 2019, doi: 10.7326/M18-1377.

[16] I. Kaiser *et al.*, "Using the Prediction Model Risk of Bias Assessment Tool (PROBAST) to Evaluate Melanoma Prediction Studies," *Cancers (Basel)*, vol. 14, no. 12, p. 3033, Jun. 2022, doi: 10.3390/cancers14123033.

[17] L. Alonso and F. Renard, "A New Approach for Understanding Urban Microclimate by Integrating Complementary Predictors at Different Scales in Regression and Machine Learning Models," *Remote Sens (Basel)*, vol. 12, no. 15, p. 2434, Jul. 2020, doi: 10.3390/rs12152434.

[18] P. Noi, J. Degener, and M. Kappas, "Comparison of Multiple Linear Regression, Cubist Regression, and Random Forest Algorithms to Estimate Daily Air Surface Temperature from Dynamic Combinations of MODIS LST Data," *Remote Sens (Basel)*, vol. 9, no. 5, p. 398, Apr. 2017, doi: 10.3390/rs9050398.

[19] L. Li and Y. Zha, "Estimating monthly average temperature by remote sensing in China," *Advances in*

Space Research, vol. 63, no. 8, pp. 2345–2357, Apr. 2019, doi: 10.1016/j.asr.2018.12.039.

[20] M. K. Alomar *et al.*, "Data-driven models for atmospheric air temperature forecasting at a continental climate region," *PLoS One*, vol. 17, no. 11, p. e0277079, Nov. 2022, doi: 10.1371/journal.pone.0277079.

[21] S. Xu *et al.*, "Spatial Downscaling of Land Surface Temperature Based on a Multi-Factor Geographically Weighted Machine Learning Model," *Remote Sens (Basel)*, vol. 13, no. 6, p. 1186, Mar. 2021, doi: 10.3390/rs13061186.

[22] S. Hong, C. Park, and S. Cho, "A Rail-Temperature-Prediction Model Based on Machine Learning: Warning of Train-Speed Restrictions Using Weather Forecasting," *Sensors*, vol. 21, no. 13, p. 4606, Jul. 2021, doi: 10.3390/s21134606.

[23] A. Derdouri, Y. Murayama, and T. Morimoto, "Spatiotemporal Thermal Variations in Moroccan Cities: A Comparative Analysis," *Sensors*, vol. 23, no. 13, p. 6229, Jul. 2023, doi: 10.3390/s23136229.

[24] D. McCarty, J. Lee, and H. W. Kim, "Machine Learning Simulation of Land Cover Impact on Surface Urban Heat Island Surrounding Park Areas," *Sustainability*, vol. 13, no. 22, p. 12678, Nov. 2021, doi: 10.3390/su132212678.

[25] M. A. Guillén-Navarro, R. Martínez-España, A. Llanes, A. Bueno-Crespo, and J. M. Cecilia, "A deep learning model to predict lower temperatures in agriculture," *J Ambient Intell Smart Environ*, vol. 12, no. 1, pp. 21–34, Jan. 2020, doi: 10.3233/AIS-200546.

[26] T. C. M. Martin, H. R. Rocha, and G. M. P. Perez, "Fine scale surface climate in complex terrain using machine learning," *International Journal of Climatology*, vol. 41, no. 1, pp. 233–250, Jan. 2021, doi: 10.1002/joc.6617.

[27] L. Madaus, P. McDermott, J. Hacker, and J. Pullen, "Hyper-local, efficient extreme heat projection and analysis using machine learning to augment a hybrid dynamical-statistical downscaling technique," *Urban Clim*, vol. 32, p. 100606, Jun. 2020, doi: 10.1016/j.uclim.2020.100606.

[28] H. Meyer, C. Reudenbach, T. Hengl, M. Katurji, and T. Nauss, "Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation," *Environmental Modelling & Software*, vol. 101, pp. 1–9, Mar. 2018, doi: 10.1016/j.envsoft.2017.12.001.

[29] M. Oliveira, L. Torgo, and V. Santos Costa, "Evaluation Procedures for Forecasting with Spatiotemporal Data," *Mathematics*, vol. 9, no. 6, p. 691, Mar. 2021, doi: 10.3390/math9060691.

[30] C. Kumar, G. Walton, P. Santi, and C. Luza, "Random Cross-Validation Produces Biased Assessment of Machine Learning Performance in Regional Landslide Susceptibility Prediction," *Remote Sens (Basel)*, vol. 17, no. 2, p. 213, Jan. 2025, doi: 10.3390/rs17020213.

[31] Y. Wang, M. Khodadadzadeh, and R. Zurita-Milla, "Spatial+: A new cross-validation method to evaluate geospatial machine learning models," *International Journal of Applied Earth Observation and Geoinformation*, vol. 121, p. 103364, Jul. 2023, doi: 10.1016/j.jag.2023.103364.

[32] J. Che *et al.*, "Reconstruction of Near-Surface Air Temperature over the Greenland Ice Sheet Based on MODIS Data and Machine Learning Approaches," *Remote Sens (Basel)*, vol. 14, no. 22, p. 5775, Nov. 2022, doi: 10.3390/rs14225775.

[33] A. Hassani, G. S. Santos, P. Schneider, and N. Castell, "Interpolation, Satellite-Based Machine Learning, or Meteorological Simulation? A Comparison Analysis for Spatio-temporal Mapping of Mesoscale Urban Air Temperature," *Environmental Modeling & Assessment*, vol. 29, no. 2, pp. 291–306, Apr. 2024, doi: 10.1007/s10666-023-09943-9.

[34] I. P. Pradhan, K. K. Mahanta, Y.-A. Liou, A. Chauhan, and D. P. Shukla, "Machine learning based high-resolution air temperature modelling from landsat-8, MODIS, and In-Situ measurements with ERA-5 inter-comparison in the data sparse regions of Himachal Pradesh," *Bulletin of Atmospheric Science and Technology*, vol. 5, no. 1, p. 22, Dec. 2024, doi: 10.1007/s42865-024-00085-8.

[35] A. Joy, K. Satheesan, and A. Paul, "High-resolution maximum air temperature estimation over India from MODIS data using machine learning," *Remote Sensing Applications: Society and …*, 2025, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352938525000163

[36] M. Zheng, J. Zhang, J. Wang, S. Yang, J. Han, and T. Hassan, "Reconstruction of 0.05° all-sky daily maximum air temperature across Eurasia for 2003–2018 with multi-source satellite data and machine learning models," *Atmos Res*, vol. 279, p. 106398, Dec. 2022, doi: 10.1016/j.atmosres.2022.106398.

[37] A. Karagiannidis, G. Kyros, K. Lagouvardos, and V. Kotroni, "Real-Time Estimation of Near-Surface Air Temperature over Greece Using Machine Learning Methods and LSA SAF Satellite Products," *Remote Sens (Basel)*, vol. 17, no. 7, p. 1112, Mar. 2025, doi: 10.3390/rs17071112.

[38] C. Xu, M. Lin, Q. Fang, J. Chen, Q. Yue, and J. Xia, "Air temperature estimation over winter wheat fields by integrating machine learning and remote sensing techniques," *International Journal of Applied Earth Observation and Geoinformation*, vol. 122, p. 103416, Aug. 2023, doi: 10.1016/j.jag.2023.103416.

[39] I. Buo, V. Sagris, and J. Jaagus, "Gap-filling satellite land surface temperature over heatwave periods with machine learning," *IEEE Geoscience and Remote …*, 2021, [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9391991/

[40] R. R. Andriambololonaharisoamalala *et al.*, "Downscaling of Urban Land Surface Temperatures Using Geospatial Machine Learning with Landsat 8/9 and Sentinel-2 Imagery," *Remote Sens (Basel)*, vol. 17, no. 14, p. 2392, Jul. 2025, doi: 10.3390/rs17142392.

[41] J. Lee, "Estimating Near-Surface Air Temperature From Satellite-Derived Land Surface Temperature Using Temporal Deep Learning: A Comparative Analysis," *IEEE Access*, vol. 13, pp. 28935–28945, 2025, doi: 10.1109/ACCESS.2025.3539581.

[42] M. Babaei, fatemeh rajaei, H. Siroosi, and N. Alamdari, "Unmasking the Thermal Signature: Using Remote Sensing and Machine Learning to Study Lake Urmia's Role in Regulating Land Surface Temperature," 2024. doi: 10.2139/ssrn.5035577.

APPENDIX A1. RISK OF BIAS ASSESSMENT OF THE REVIEWED STUDIES USING PROBAST

Study	Region	Sensor(s)	In-situ stations (#)	Spatial resolution	Algorithms tested	Validation method	RMSE	MAE	R^2	Remarks
Xu et al., 2023	China (Wheat fields)	MODIS, ERA5	118	1 km	TVX+MLR, RF, DBN	10-fold CV (random)	1.62 °C	1.20 °C	0.97	Integrating TVX with RF provided the best performance in agricultural regions.
Andriambololonaharisoamalala et al., 2025	Perth, Australia (Urban)	Landsat 8/9, Sentinel-2	n/a	Downscaling 30m–10m	GeoML (RF+Kriging)	Ground-based measurements	2.7 °C	<2.2 °C	r=0.85	GeoML outperformed HUTS in urban LST downscaling.
Karagiannidis et al., 2025	Greece	LSA SAF (Meteosat)	126	5 km	RF, XGBoost, ANN	Train-test split (random)	1.34 °C	0.96 °C	0.976	RF was the best model for "all-sky" temperature estimation.
Jangho Lee, 2025	Illinois, USA	GOES, Landsat	63	2–10 km	LSTM, TCN, N-BEATS	Temporal holdout (2021)	<1.8 °C	–	–	Temporal models significantly outperformed non-temporal ones.
Wang et al., 2022	Jingjinji, China	MODIS, STRM	1527	1 km	RF, DT, FNN, GLM, SVM	10-fold CV (random)	1.29 °C	0.94 °C	0.99	RF was the best algorithm; DSR was the second most important variable after LST.
Joy et al., 2025	India	MODIS, SRTM	182	5.6 km	XGBoost, ANN, GAM, MLR	5-fold CV (random)	1.79 °C	1.37 °C	0.90	XGBoost achieved the best performance; LST was the most influential predictor.
Che et al., 2022	Greenland Ice Sheet	MODIS, RACMO	25	780 m	NN, GPR, SVM, RF	Leave-location-out	2.67 °C	–	r=0.96	NN achieved the highest accuracy in extreme conditions.
Zheng et al., 2022	Eurasia	MODIS, GLASS, GLDAS	4476	5 km	HGB, ET, RF, DBN	10-fold CV (random)	1.74 °C	1.30 °C	0.984	HGB was the most accurate and robust model with missing values.
Buo et al., 2021	Estonia	MODIS	25	1 km	RF, OLS	5-fold CV (random)	1.37 °C	–	0.85	RF was effective for filling LST gaps during heat waves.
Babaei et al., 2023	Lake Urmia, Iran	Landsat 5–8	n/a	30 m	Multiple ML	Temporal split	–	–	–	Predicted future LST, showing drastic decline in cooling effect of the lake.
Hassani et al., 2024	Warsaw, Poland (Urban)	Landsat, Sentinel, MODIS, ERA5	5	1 km	XGBoost, OK, WRF	Ground-based measurements	1.06 °C	–	0.94	XGBoost outperformed Kriging and WRF in urban temperature mapping.
Pradhan et al., 2024	Himachal Pradesh, India	Landsat 8, MODIS	4	30 m	MLP, GAM, MLR	Train-test split (70/30)	1.13 °C	0.83 °C	0.94	MLP achieved the best performance with sparse data in mountainous areas.

APPENDIX A2. RISK OF BIAS AND QUALITY ASSESSMENT OF THE REVIEWED STUDIES (ADAPTED FROM PROBAST)

Article	Data (Quality & Size)	Modeling	Validation (CV)	Reported Metrics	Brief Notes
Xu et al., 2023	In-situ from 118 CMDC stations (2013–2021), MODIS, ERA5. Good quality, size not specified.	TVX + MLR, RF, DBN. RF and DBN hyperparameters detailed, tuning via grid search.	10-fold cross-validation.	R^2, MAE, RMSE.	Integrated TVX method improved accuracy in agricultural areas.
Andriambololonaharisoamalala et al., 2025	LST from Landsat 8/9 (30m), Sentinel-2 (10m). Validated with ground-based and ECOSTRESS data. Size not specified.	Hybrid GeoML (RF + Kriging). RF hyperparameters reported in detail.	Validation with independent reference data (not CV).	Pearson correlation, RMSE, MAE.	Focused on urban LST downscaling using ML–geostatistical integration.
Karagiannidis et al., 2025	LSA SAF (2020–2022), 126 in-situ stations in Greece. Size not specified.	RF, XGBoost, ANN. Hyperparameter tuning with GridSearchCV.	Train/validation split (80/20).	MAE, MBE, RMSE, R^2.	Identified RF as the most efficient model for all-sky temperature estimation.
Jangho Lee, 2025	GOES LST, Landsat NDVI/NDBI, 63 in-situ stations (2019–2023). Large dataset (1,471,239 training / 365,778 testing).	LSTM, TCN, N-BEATS. Hyperparameters (dropout, architecture) described.	Temporal holdout (2021 for testing).	RMSE, MAE, R^2, precision, recall, F1.	Comparative evaluation of 3 temporal DL models with multiple look-back windows.
Wang et al., 2022	1527 in-situ stations, MODIS LST, DSR, NDVI, LC, DEM (2018–2019). Large dataset (166,008 daily, 992,705 instantaneous).	RF, DT, FNN, GLM, SVM. Hyperparameter tuning not specified.	10-fold cross-validation.	MAE, RMSE, R^2.	Highlighted DSR importance in Ta estimation.
Joy et al., 2025	182 in-situ stations, MODIS LST/NDVI (2010–2022). Large dataset (190,080 training / 63,456 testing).	XGBoost, ANN, GAM, MLR. Tuning via grid search and 5-fold CV.	5-fold cross-validation.	RMSE, MAE, R^2.	XGBoost outperformed other models with high accuracy.
Che et al., 2022	25 AWS stations in Greenland (2007–2019), MODIS LST/albedo, RACMO2.3p2. 20,874 samples. Data split: 10 stations training, 15 validation.	NN, GPR, SVM, RF. Hyperparameter tuning not detailed.	Station-based train/validation split.	R, RMSE, Bias.	NN achieved highest accuracy in extreme conditions.
Zheng et al., 2022	4476 in-situ stations in Eurasia (2003–2018), MODIS, GLASS, GLDAS. Large dataset (>2 million samples).	HGB, ET, RF, DBN. Hyperparameter tuning via grid search and learning curves.	10-fold cross-validation.	R^2, MAE, RMSE, Bias.	HGB chosen for accuracy and robustness with missing values.
Buo et al., 2021	25 EWS stations, MODIS LST/NDVI, Estonia landscape DB. Large dataset (983,443 training / 655,629 testing).	RF. OLS as baseline. Hyperparameter tuning with grid search.	5-fold CV (grid search).	RMSE, R^2.	Focused on LST gap-filling during heat waves.
Babaei et al., 2023	Landsat 5, 7, 8 (1986–2023). Size not specified.	Multiple ML algorithms. Hyperparameter tuning with grid search.	Train (2010–2016), test (2023).	R^2, MSE.	Predicted future LST; documented declining

					cooling effect of Urmia Lake.
Hassani et al., 2024	85 Netatmo stations, Landsat/Sentinel/MODIS/ERA5. Crowdsourced data, QA applied. Size not specified.	XGBoost, OK, WRF. Hyperparameter optimization with 10-fold CV.	10-fold cross-validation.	MAE, RMSE, R^2.	Compared ML models with Kriging and numerical simulations.
Pradhan et al., 2024	4 IMD stations in Himachal Pradesh (2013–2020), Landsat-8, MODIS. Dataset split: 70% training / 30% testing.	MLP, GAM, MLR. MLP hyperparameters explained in detail.	Train/test split.	R^2, RMSE, MAE, NSE, KGE, etc.	Conducted in data-scarce mountainous region.