



Enhancing Weather Pattern Recognition through Multi-Stage Sensor Data Fusion and Smoothing Regression

Fuangfar Pensiri¹, Surapong Wiriya², Keun Ho Ryu³, and Porawat Visutsak^{4*}

¹ Faculty of Science at Sriracha, Kasetsart University, Sriracha Campus, Chonburi, 20230, Thailand

² Chaopraya University, Nakhon Sawan, 60240, Thailand

³ Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh City, 700000, Vietnam

⁴ Faculty of Applied Science, King Mongkut's University of Technology North Bangkok, Bangkok, 10800, Thailand

* Correspondence: porawatv@kmutnb.ac.th

Citation:

Pensiri, F.; Wiriya, S.; Ryu, K. H.; Visutsak, P. Enhancing weather pattern recognition through multi-stage sensor data fusion and smoothing Regression. *ASEAN J. Sci. Tech. Report.* **2026**, 29(9), e262518. <https://doi.org/10.55164/ajstr.v29i9.262518>

Article history:

Received: November 24, 2025

Revised: May 10, 2026

Accepted: June 1, 2026

Available online: July 26, 2026

Publisher's Note:

This article is published and distributed under the terms of the Thaksin University.

Abstract: It is often harder to make accurate predictions with large, various intelligent sensor systems because of data noise and the high cost of traditional modeling methods like Numerical Weather Prediction (NWP). This paper presents a comprehensive framework that integrates multi-stage sensor data fusion and preprocessing with machine-learning-based pattern recognition to produce highly precise, computationally efficient analog weather forecasts. Our method uses a strict three-stage pipeline to clean up and combine raw sensor inputs. It uses historical data from a distributed network of 79 NOAA weather stations. There are three phases: filling in missing values, addressing outliers, and smoothing moving averages. Subsequently, a decision tree model is used to identify similar weather patterns and predict critical weather variables. Our method is effective because it makes predictions that are much more likely to be accurate than incorrect. The Root Mean Squared Error (RMSE) declined by over 99%, from 0.1715 to an amazing 0.0011014. The results show that combining targeted sensor data preprocessing with complex pattern recognition is a suitable methodology for making smart weather forecasting systems work. Rather than serving as a replacement for physics-based NWP models, the proposed approach should be regarded as a lightweight complementary tool for localized, short-term analog weather pattern recognition.

Keywords: Weather data; regression; machine learning; numerical modeling; analog ensemble forecasting

1. Introduction

Weather forecasting is essential for farming, travel, disaster preparedness, and resource management. Numerical Weather Prediction (NWP) has been used in most operational forecasts in the past. It uses large-scale atmospheric models to solve complicated physical equations [1]. NWP provides precise medium- and long-range forecasts, but it requires substantial computing power and is sensitive to adverse initial conditions, which can lead to errors that worsen over time due to atmospheric chaos [2]. Consequently, data-driven and machine-learning methodologies have become more popular as simpler alternatives or additional tools for localized, short-term forecasting [3].

Analog forecasting has come back as a promising method among these. Analog forecasting relies on upcoming weather patterns developing in a manner analogous to historical atmospheric conditions that closely resemble present

circumstances. Recent research has enhanced analog-based strategies by refining similarity metrics, optimizing analog ensembles, and incorporating machine learning into analog selection [4]–[6]. The versatility and ongoing evolution of analog forecasting are further highlighted by recent cross-disciplinary advancements in both atmospheric application and computational efficiency. On the one hand, the refinement of similarity metrics remains a critical focus; recent studies emphasize that the careful selection of distance metrics—specifically, the superiority of Euclidean distance over spatial correlation—significantly improves the identification of reliable circulation analogs for extreme weather events [15]. On the other hand, the computational execution of these distance-based classification algorithms is being revolutionized by hardware innovations, such as the development of ultra-low-power analog integrated radial basis function classifiers that efficiently approximate Euclidean distances at the circuit level [16]. Together, these advancements underscore the potential not only to improve the methodological precision of analog weather pattern recognition, but also to eventually deploy these data-driven forecasting systems efficiently in resource-constrained environments, utilizing novel, low-power integrated architectures.

But the accuracy of analog forecasting depends heavily on the quality of the historical observational data. Raw meteorological datasets, especially station-level records such as those from NOAA GHCN-Daily, often contain missing values, noise, outliers, and inconsistencies due to sensor errors, reinstallation events, and gaps in reporting [7–8]. These problems make it harder to match similar items and degrade the performance of regression models that follow. Previous studies have proposed various preprocessing techniques for environmental and meteorological time series, including missing-value imputation, outlier removal, filtering, and moving-window smoothing [9–11]. In this context, a recent study highlighted the significant impact of outliers on statistical results and proposed a framework using the Inter-Quartile Range (IQR) to systematically detect and resolve anomalies before analysis, thereby preventing skewing [19]. Furthermore, integrating simple moving average methods with interpolation is effective for short-term forecasting by reducing high variability and clarifying significant underlying trends [20]. However, there is still a lack of systematic investigations analyzing the impact of multi-stage preprocessing pipelines on the efficacy of analog forecasting, particularly at the local station level.

Moreover, although numerous studies have juxtaposed machine learning models, including decision trees, random forests, Gaussian process regression, and ensemble methods for meteorological forecasting [12–14], comparative evaluations of these algorithms, such as the work to demonstrate the utility of performance metrics like RMSE and R-squared in selecting robust predictors for complex datasets [21]. Yet no study has assessed these models within a cohesive analog forecasting framework that features uniform preprocessing and walk-forward validation. Recent research emphasizes the importance of utilizing realistic validation methods alongside model optimization. For instance, applying a structured temporal decomposition framework paired with rigorous walk-forward validation substantially enhances forecasting accuracy [17]. Similarly, combining walk-forward validation with hyperparameter optimization is essential to prevent data leakage and accurately simulate real-world scenarios [18]. The risk of information leakage is particularly acute in time series prediction; a recent study underscored this by identifying defects in standard training/test splits and proposing sliding-window strategies to ensure forecasting validity and prevent the high-accuracy illusion caused by test-set data leaking into the model [22].

Moreover, recent initiatives that integrate analog forecasting with machine learning primarily focus on improving analog selection or post-processing of NWP outputs [4–6], rather than on developing a straightforward, preprocessing-oriented analog-regression framework for localized forecasts. To fill these gaps, this study proposes a systematic, multi-stage framework for data preprocessing and machine-learning regression to improve analog weather forecasting. The framework will use observational data from 79 NOAA GHCN-Daily stations in the Oakland, California, area. The framework includes 1. Filling in missing values, 2. Finding and fixing outliers, and 3. Smoothing time series. Then, it selects analogs and compares four regression models: linear regression, decision tree regression, Gaussian process regression, and ensemble regression. A time-series walk-forward validation strategy is used to test the framework, ensuring it can be reproduced and that data doesn't leak.

There are three main contributions:

1. Methodological contribution: We create a reproducible, multi-stage preprocessing pipeline specifically designed for noisy station-level meteorological time series, which makes analog identification more reliable.
2. Contribution to comparative modeling: We systematically compare four regression models using the same preprocessing and validation settings to show when decision tree regression works better.
3. Empirical contribution: We demonstrate that preprocessing significantly enhances forecast precision, decreasing RMSE from a raw baseline of 0.1715 to 0.11–0.15 in typical instances, with the optimal configuration reaching 0.0011. This shows how important excellent preprocessing is for analog forecasting.

The scope of this study is limited to localized, short-term analog weather pattern recognition using preprocessed station-level sensor data. The proposed framework is not intended to serve as an operational national-scale forecasting system or to replace physics-based NWP models. Instead, it is designed as a complementary preprocessing–regression framework that can support local analog forecasting when computationally intensive NWP systems are unavailable or unnecessary. In addition, this study does not focus on extreme weather prediction; therefore, the reported results should be interpreted within the context of local pattern recognition under the selected preprocessing and validation settings. We organized this work into five sections: 1) Introduction; 2) Materials and Methods; 3) Results and Discussion; 4) Conclusion.

2. Materials and Methods

Over the past decade, numerical weather prediction (NWP) has been used to generate weather forecasts by training models on historical weather data. The major drawback of NWP bands is their computationally intensive calculations. This article presents real-world historical data at the national level. The National Oceanic and Atmospheric Administration (NOAA) is a U.S. government agency that collects weather data from many countries worldwide and develops machine learning models. The data source of this research was derived from this organization, from which we collected sensor data such as precipitation, sky cover, clouds, sunlight, and air temperature, which were collected around Oakland International Airport, CA, US (Figures 1 and 2), and 79 stations nearby in the Global Historical Climatology Network-Daily (GHCN-Daily) Database from 2017 to 2019. However, some of this raw data is incomplete, and some of it has many missing values because some sensors have been reinstalled. Therefore, we need to preprocess the data using an anomaly detection method to detect outliers and replace missing values with approximate values to reduce the impact of null values in subsequent analysis. Then we applied regression techniques to analog weather forecasting. It uses historical weather to predict the future. We found that precipitation analysis and some related sensory data forecasts improved analog forecasting accuracy in this study. The methodologies are described with sufficient rigor to ensure reproducibility and transparency in the following subsections.

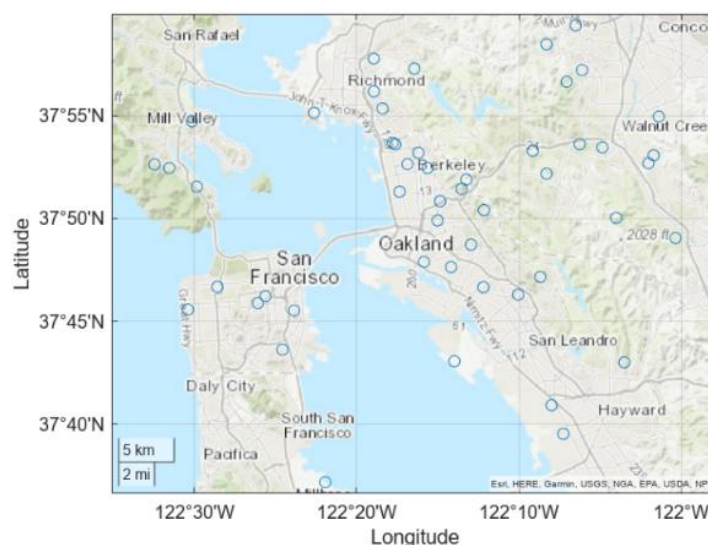


Figure 1. The Oakland International Airport's Location.

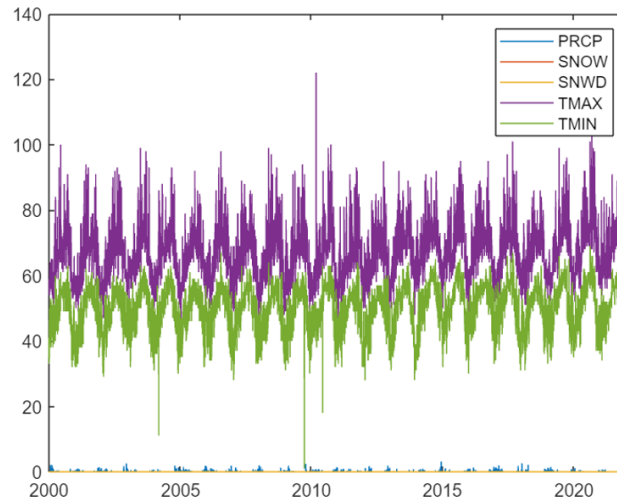


Figure 2. Data derived from sensors around Oakland Airport from 2000 to 2020 for contextual visualization. (The model training and evaluation in this study used only the 2017–2019 subset.)

2.1 Data Acquisition and Characteristics

The foundation of our study is a historical weather dataset acquired from the National Oceanic and Atmospheric Administration (NOAA), a primary U.S. government agency responsible for extensive global weather data collection. Specifically, we utilized the Global Historical Climatology Network-Daily (GHCN-Daily) database, a comprehensive repository of daily climate summaries from land-based stations worldwide. The GHCN-Daily integrates records from over 100,000 stations across 180 countries, with some records extending back to 1763. While this database undergoes rigorous quality assurance reviews, it is not homogenized to account for systematic biases from different reporting eras, a factor we considered in our analysis. For this research, we focused our data collection on a localized region, sourcing data from 79 stations in and around Oakland International Airport, CA, USA. While the broader GHCN-Daily network spans 1960 to the present, our study used a specific three-year period from 2017 to 2019. This timeframe was selected to ensure a high-quality, consistent dataset, minimizing variability from long-term changes in instrumentation or reporting practices. Our analysis employed seven predictor variables derived from the sensor data: precipitation, sky cover and clouds, sunshine, air temperature, water (humidity/dew point), wind, and weather type. Among these, precipitation was designated as the primary predictor. The objective was to forecast four key meteorological response variables: Snowfall, Snow Depth, Maximum Temperature, and Minimum Temperature. A significant characteristic of the raw dataset was the presence of irregularities, including incompleteness and numerous missing values. These data quality issues are largely attributable to common operational factors in environmental monitoring, such as sensor reinstallation and malfunctions. The inherent noise and incompleteness of the raw data necessitated the development of the robust data preprocessing pipeline detailed in the subsequent section, as these imperfections can obscure the underlying patterns essential for accurate forecasting.

2.2 Data Preprocessing Pipeline

Raw historical weather data frequently contains noise, missing values, and anomalies that can obscure underlying patterns and degrade model performance. To address these challenges, we implemented a comprehensive preprocessing pipeline integrating methods for missing value imputation, outlier detection and treatment, and data smoothing. Our objective was to refine the dataset, thereby enabling more accurate identification of analogous sequences and improving overall forecasting accuracy. To maintain the temporal integrity of the forecasting experiment, the preprocessing pipeline was applied within each training window rather than to the full dataset before validation. Thus, missing-value imputation, outlier treatment, smoothing, and normalization were performed using only historical observations available at each validation step.

Table 1. Summary of NOAA GHCN-Daily Data Characteristics.

Characteristic	Detail
Source	National Oceanic and Atmospheric Administration (NOAA) Global Historical Climatology Network-Daily (GHCN-Daily) Database
Geographic Scope	Global, with a specific study focus on Oakland International Airport, CA, US, and 79 nearby stations
Temporal Coverage	Initial data availability: 1960–2022; Specific data for analysis: 2017–2019
Sensor Data Types (Predictors)	Precipitation, Sky Cover and Clouds, Sunray, Air Temperature, Water, Wind, Weather Type
Response Variables (Predictands)	Snowfall, Snow Depth, Maximum Temperature, Minimum Temperature
Initial Data Quality Issues	Incompleteness, numerous missing values due to sensor reinstallation and operational factors

2.2.1 Missing Value Imputation

The raw data from the NOAA GHCN-Daily repository showed significant incompleteness, with numerous missing values due to sensor reinstallation and other operational disruptions. Because unaddressed missing data can introduce bias and reduce statistical power, we established a robust imputation strategy. Our approach involved both removing records with excessive missingness and imputing salvageable gaps, treating missing data points as anomalies to be corrected. Given the temporal dependencies inherent in time series data, we prioritized imputation techniques that preserve the data's structure. We considered several methods, including linear interpolation, which estimates missing points along a straight line between valid neighbors, and cubic spline interpolation, which fits a flexible curve to capture more complex variations. Further, we explored regression imputation, which leverages relationships between correlated weather variables (e.g., using temperature to predict humidity) to estimate missing values. While simpler methods such as mean or median substitution were considered, they were deemed less suitable because they can distort the data distribution and ignore critical temporal relationships. Our final strategy focused on using interpolation techniques to generate "approximate values," ensuring data continuity and minimizing the impact of null entries on subsequent analyses.

2.2.2 Outlier Detection and Treatment

Our dataset contained outliers, defined as data points that deviate significantly from expected patterns, which can stem from sensor malfunctions, data entry errors, or genuine extreme weather events. To prevent such anomalies from distorting statistical measures and reducing model accuracy, we implemented an anomaly detection method to identify and treat them. Rather than removing outliers, which would disrupt the temporal continuity of the series, we replaced them with approximate values derived from neighboring data points. For detection, we evaluated methods well-suited to time series weather data. These included statistical thresholds based on the Z-score and the Interquartile Range (IQR) rule. We found the IQR method, which identifies points outside $1.5 \times \text{IQR}$ of the first and third quartiles, to be particularly robust for non-normally distributed weather data such as precipitation. Additionally, we considered time-series decomposition (e.g., STL) to isolate outliers within the residual component. The threshold for defining an outlier is a critical parameter. Our approach involved setting thresholds based on statistical rules, refined with domain-specific meteorological knowledge to ensure the data's physical plausibility. This imputation-based treatment of outliers preserves the integrity and length of the time series, which is crucial for our analysis goals.

2.2.3 Data Smoothing (Moving Average)

To reduce random noise and highlight significant underlying trends and cycles, we applied a moving average (MA) filter to the time series data. This smoothing process, also known as filtering, mitigates "ad hoc inconsistency" by calculating a new series in which each point is the average of a subset of the original values within a specified window. The formula for a simple moving average (MA) with a window size of n over a series of points p is:

$$MA = \sum_{i=1}^n P_i/n \quad (1)$$

The rationale for applying a moving average in our analog weather forecasting framework is threefold. First, it filters random fluctuations, allowing our model to focus on the true underlying signal in the weather data. Second, by reducing noise, it helps emphasize long-term trends and cycles, making it easier to identify truly analogous historical periods. Third, the smoothed data yields cleaner visualizations, rendering patterns more apparent for analysis. The choice of the MA window size is a critical hyperparameter; a larger window produces a smoother curve but is less responsive to recent changes, while a shorter window is more responsive but retains more noise. We selected our window size through experimentation to balance smoothness with responsiveness.

2.2.4 Rationale for the Combined Preprocessing Approach

The three components of our pipeline—missing data imputation, outlier treatment, and data smoothing—were not applied in isolation but were employed in conjunction as an integrated, multi-stage process. This synergistic approach is essential because each method addresses a distinct data quality issue, and their combined application yields a dataset of superior quality compared to using raw data or any single technique alone. Our choice of methods was deliberately tailored to the unique characteristics of time series weather data, such as seasonality, trends, and temporal dependencies. Techniques like interpolation, time-series-aware outlier detection, and moving averages were selected because they preserve and enhance the temporal patterns crucial for forecasting, unlike simpler methods (e.g., mean imputation) that would destroy these relationships. We recognize that each technique involves trade-offs; for instance, aggressive smoothing can obscure important short-term signals, while overly sensitive outlier detection can remove valid data on extreme events. Our methodology involved carefully tuning these methods to balance the trade-offs, demonstrating a rigorous, considered approach. The integrated benefit of our pipeline is clear: value imputation ensures data completeness, preventing gaps from disrupting time-series analysis. Outlier treatment mitigates the distorting effects of anomalous data points, ensuring models learn from representative patterns. Finally, data smoothing refines this cleaned data by reducing residual noise and making the underlying signal clearer for the subsequent analog selection and forecasting models. This multi-stage approach ensures that the input to our models is as clean, consistent, and representative as possible, thereby maximizing the potential for enhanced prediction accuracy.

2.3 Analog Weather Forecasting Model Development

Our model development process is centered on an analog forecasting approach, augmented with machine learning-driven regression to generate precise predictions. The methodology consists of two primary stages: an analog selection process to identify relevant historical weather patterns and the training of regression models to forecast specific response variables.

2.3.1 Analog Selection Process

The core of the proposed methodology is analog forecasting, which assumes that future weather conditions may resemble historical weather patterns with similar predictor characteristics. In this study, analog selection was performed by systematically searching the preprocessed historical dataset to identify past observations with weather characteristics similar to those of the target forecast date. The analog selection procedure was implemented as a Euclidean-distance-based search over the preprocessed predictor variables. No separate machine-learning-based metric-learning model was trained for analog selection. Instead, machine learning was applied in the subsequent regression stage after the analog candidates had been selected. This clarification was made to ensure that the analog selection mechanism is clearly defined and reproducible. The analog candidate pool was constructed from the preprocessed historical observations collected from the 79 NOAA GHCN-Daily stations used in this study. At each validation step, only observations that occurred before the target forecast date were eligible to be selected as analog candidates. Observations from the validation period, test period, the target date itself, and any future dates were excluded from the analog pool.

This design was intended to avoid temporal overlap and prevent the use of future information during analog selection. For each target forecast date, the Euclidean distance between the target predictor vector and each eligible historical candidate vector was calculated. The candidates with the smallest distances were retained as analogs and used in the subsequent regression step. The number of retained analogs was tested within the range of $N_i = 5$ to 10 , and the best-performing reported configuration used $N_i = 5$. This value provided a balance between retaining sufficiently similar historical cases and avoiding excessive smoothing of local anomalies.

Table 2. Overview of Preprocessing Techniques and Rationale.

Pre-processing Category	Method/ Approach Used	Specific Algorithms/ Techniques (Implemented)	Rationale for Weather Time Series Data
Missing Value Imputation	Anomaly detection method to replace missing values	Linear/Cubic Spline Interpolation, Regression Imputation, Stochastic Models (e.g., ARIMA with Kalman Smoothing)	To fill in incomplete data while preserving temporal dependencies; reduce bias and information loss.
Outlier Detection & Treatment	Anomaly detection method to detect outliers	Statistical Thresholds (Z-score, IQR), Time Series Decomposition (STL), Prediction-Based Methods, ML Approaches (Isolation Forest)	To identify and handle anomalous points (e.g., sensor errors) that distort statistical measures and reduce model accuracy, and to prevent overfitting.
Data Smoothing	Moving Average	Simple Moving Average (SMA), Exponentially Weighted Moving Average (EWMA)	To reduce random noise and ad hoc inconsistencies, emphasize the underlying trends and cycles to clarify pattern identification.

Note: The methods listed in this table were implemented as candidate preprocessing methods and evaluated across the experimental configurations. Different configurations used different combinations and orders of these methods, as summarized in Table 4.

2.3.2 Regression Model Training

Following the identification of analogous periods, we trained and evaluated a set of regression models to predict the continuous target variables, namely Snowfall, Snow Depth, Maximum Temperature, and Minimum Temperature. The evaluated models included Linear Regression, Decision Tree Regression ("Tree"), Gaussian Process Regression (GPR), and Ensemble Regression. Linear Regression was used as a baseline model to capture linear relationships between the predictor variables and the target variables. Decision Tree Regression was implemented to capture non-linear relationships in the data. Its hierarchical structure also provides a relatively interpretable prediction process, which is useful for examining how predictor variables contribute to the forecast. Gaussian Process Regression (GPR) was evaluated because of its probabilistic nature and its ability to provide both point predictions and uncertainty estimates through kernel-based covariance functions. Ensemble Regression was included to examine whether combining multiple base learners could improve predictive robustness compared with individual models. The model training workflow was standardized using the `trainRegressionPRCP` function. This function takes the preprocessed weather data as input and returns a trained regression model, a prediction function (`predictFcn`), the required predictor variables, and the validation Root Mean Squared Error (RMSE). Hyperparameter tuning was performed for each model where applicable. For Decision Tree Regression, the tuned parameters included the split criterion, maximum tree depth, and minimum number of samples per leaf. For Gaussian Process Regression, tuning focused on the kernel function, noise parameter, and optimizer. For Ensemble Regression, the number of base learners and learning-related parameters were considered. Within the tested preprocessing model configurations, Decision Tree Regression produced the lowest observed RMSE. However, because both preprocessing settings and model settings varied across trials, this result should be interpreted as the best-

performing configuration within the tested experimental set rather than as a fully controlled comparison proving the universal superiority of one regression model over all others.

Table 3. Regression Models and Key Hyperparameters.

Model Type	Key Hyperparameters (Examples)	Common Use in Weather Forecasting
Linear Regression	None typically, but considerations for feature engineering (e.g., time-step, lag features)	Baseline modeling, identifying linear trends, and simple relationships
Decision Tree Regression	criterion, max_depth, min_samples_split, min_samples_leaf, max_features, splitter	Capturing non-linear relationships, interpretability, robustness to outliers
Gaussian Process Regression	kernel (e.g., RBF, Periodic, Matérn), alpha (noise), optimizer, n_restarts_optimizer	Uncertainty quantification, small datasets, complex non-linear patterns
Ensemble Regression	Number of base learners, learning rate, subsampling parameters (e.g., for Random Forest or Gradient Boosting)	Improved accuracy by combining multiple models and enhanced robustness.

2.3.3 Model Evaluation and Validation Strategy

2.3.3.1 Performance Metrics

We evaluated the performance of our regression models primarily using the Root Mean Squared Error (RMSE). RMSE quantifies the magnitude of error between predicted and observed values and was selected for its sensitivity to large errors, which are particularly undesirable in forecasting applications. The RMSE is calculated as:

$$RMSE = \sqrt{\sum_{i=1}^n \frac{(Predicted - Actual)^2}{n}} \quad (2)$$

A lower RMSE indicates a better model fit. While RMSE was our primary metric, we also considered supplementary metrics for a more comprehensive assessment, including Mean Absolute Error (MAE) for its robustness to outliers and the Coefficient of Determination (R^2) to quantify the proportion of variance in the target variable explained by the model.

2.3.3.2 Time Series Cross-Validation

To ensure that the models generalize to unseen data and to avoid overfitting, we implemented a validation strategy specifically designed for time-series data. Standard k-fold cross-validation with random splits is inappropriate for temporal data because it may allow information from future observations to influence model training. Therefore, this study adopted a chronological walk-forward validation approach. The dataset was ordered by time and partitioned sequentially. The initial training window covered the first 70% of the chronological data, while the final 30% was used for validation/testing. At each validation step, the model was trained only on observations that occurred before the validation block. The step size was set to 14 days, and the procedure was repeated for five validation folds. After each step, the validation window advanced, and only observations that had become historical data were eligible for inclusion in the subsequent training window. This process simulated a realistic forecasting scenario in which only past information is available at the time of prediction. To prevent data leakage, all preprocessing operations were performed within each training window rather than on the full dataset before splitting. Specifically, missing-value imputation rules, outlier-detection thresholds, smoothing settings, and normalization parameters were estimated solely from the training data. These parameters were then applied to the subsequent validation/test block. The validation/test observations were not used to estimate preprocessing parameters, select analog candidates, tune model parameters, or train the regression model. The moving average smoothing was applied as a trailing moving average rather than a centered moving average. Therefore, each smoothed value was calculated using only the current and previous observations within the training window, without incorporating future observations. In addition, the analog candidate pool was restricted to historical observations

within the corresponding training window. The target date, the validation/test period, and all future observations were excluded from the analog pool. This procedure was designed to preserve the time series' causal structure, prevent temporal overlap, and avoid artificially improving validation results. Accordingly, the reported RMSE values were obtained under a causal time-series validation setting.

3. Results and Discussion

To illustrate the model training and prediction workflow, the regression training procedure is summarized in Algorithm 1. The function takes the preprocessed local weather dataset as input and returns a trained regression model together with its validation RMSE. The trained model contains the prediction function, the required predictor variables, and the fitted regression object.

Algorithm 1. Regression model training and prediction workflow

Input: Preprocessed local weather dataset (localweather2)

Output: trainedModel, validationRMSE

1. Load the preprocessed weather dataset.
2. Select the required predictor variables and the target variable.
3. Train the regression model using the selected regression method.
4. Compute the validation RMSE using the chronological validation procedure.
5. Store the trained model, prediction function, required variables, and validation RMSE.
6. Use trainedModel.predictFcn(newDataTable) to generate predictions for new data.

The trainedModel object includes the prediction function (predictFcn), the list of required predictor variables, the fitted regression model, and the corresponding validation RMSE. The validation RMSE was then used to compare the tested preprocessing model configurations reported in Tables 4 and 5.

Table 4. Preprocessing Data.

	Data Cleansing (1)	Data Cleansing (2)	Data Cleansing (3)	RMSE (Current Value: 0.1715)
1	Clean Missing Data	Clean Outlier Data	Smooth Data	0.17006
2	Clean Missing Data	Clean Outlier Data	Smooth Data	0.16794
3	Clean Missing Data	Clean Outlier Data	Smooth Data	0.1682
4	Clean Missing Data	Clean Outlier Data	Smooth Data	0.1515
5	Clean Missing Data	Smooth Data	Clean Outlier Data	0.1678
6	Clean Missing Data	Clean Outlier Data	Clean Outlier Data	0.1701
7	Clean Missing Data	Smooth Data	Clean Outlier Data	0.1438
8	Clean Missing Data	Clean Outlier Data	Smooth Data	0.1459
9	Clean Missing Data	Clean Outlier Data	Smooth Data	0.11461
10	Clean Missing Data	Clean Outlier Data	Smooth Data	0.15174
11	Clean Missing Data	Clean Outlier Data	Smooth Data	0.15171
12	Clean Missing Data	Smooth Data	-	0.0011014
13	Clean Missing Data(remove)	Clean Missing Data(fill)	-	0.16613
14	Clean Missing Data	Smooth Data	Clean Outlier Data	0.15432
15	Clean Missing Data	Clean Outlier Data	Smooth Data	0.1393
16	Clean Missing Data	Clean Outlier Data	Smooth Data	0.16986
17	Clean Missing Data	Clean Outlier Data	Smooth Data	0.1686
18	Clean Missing Data	Clean Outlier Data	Smooth Data	0.0364
19	Clean Missing Data	Clean Outlier Data	Smooth Data	0.1711
20	Clean Missing Data	Clean Outlier Data	Smooth Data	0.062367
21	Clean Missing Data	Clean Outlier Data	Smooth Data	0.060006
22	Clean Missing Data	Clean Outlier Data	Smooth Data	0.068301

Table 4. Preprocessing Data. (continue)

	Data Cleansing (1)	Data Cleansing (2)	Data Cleansing (3)	RMSE (Current Value: 0.1715)
22	Clean Missing Data	Clean Outlier Data	Smooth Data	0.068301
23	Clean Missing Data	Clean Outlier Data	Smooth Data	0.1417
24	Clean Missing Data	Clean Outlier Data	Smooth Data	0.1701
25	Clean Missing Data	Clean Outlier Data	Smooth Data	0.135501905
26	Clean Missing Data	Clean Outlier Data	Smooth Data	0.1538
27	Clean Missing Data	Clean Outlier Data	Smooth Data	0.1537
28	Clean Missing Data	Clean Outlier Data	Smooth Data	0.1369
29	Clean Missing Data	Clean Outlier Data	Smooth Data	0.1408
30	Smooth Data	Clean Missing Data	-	0.141
31	Clean Missing Data	Clean Outlier Data	Smooth Data	0.1364
32	Clean Missing Data	Clean Outlier Data	Smooth Data	0.1363
33	Clean Missing Data	Clean Outlier Data	Smooth Data	0.1548
34	Clean Missing Data	Clean Outlier Data	Smooth Data	0.1534
35	Clean Missing Data	Clean Outlier Data	Smooth Data	0.1502
36	Clean Missing Data	Clean Outlier Data	Smooth Data	0.060006
37	Clean Missing Data	Clean Outlier Data	Smooth Data	0.0593

Table 4 should be interpreted as a configuration search log rather than a direct controlled comparison of preprocessing methods. The table records a series of experimental trials conducted to identify the preprocessing sequence that produced the lowest RMSE. Each trial consisted of up to 3 data-cleansing stages. At each stage, one of the available operations—Clean Missing Data, Clean Outlier Data, or Smooth Data—could be selected. Some trials used all three stages, while others used only two stages. In addition, the same type of operation could appear in more than one configuration because the order and combination of preprocessing steps were varied during the search process. The initial training window was 70% of Chronological data, the step size was 14 days, and the total number of validation folds was 5. The objective of these repeated trials was to determine which preprocessing sequence, when combined with a regression model, yielded the lowest prediction error. The RMSE value was used as the selection criterion. Among the tested configurations, the lowest RMSE was obtained from the sequence that applied Clean Missing Data and Smooth Data. According to Table 5, this best-performing configuration used the Tree regression function. Therefore, the result should be interpreted as evidence that this specific preprocessing model configuration achieved the lowest RMSE among the tested configurations, rather than as a general conclusion that one preprocessing method is universally superior to all others.

Table 5. The RMSE Value and Training Function.

	RMSE (Current Value: 0.1715)	
1	0.17006	Tree
2	0.16794	Gaussian process regression
3	0.1682	Least Squares Regression Kernel
4	0.1515	Tree
5	0.1678	Gaussian process regression
6	0.1701	Linear Regression
7	0.1438	Tree
8	0.1459	Tree
9	0.11461	Tree
10	0.15174	Tree
11	0.15171	Tree
12	0.0011014	Tree

Table 5. The RMSE Value and Training Function. (continue)

RMSE (Current Value: 0.1715)		
13	0.16613	Tree : Coarse Tree
14	0.15432	Tree
15	0.1393	Tree
16	0.16986	Gaussian process regression
17	0.1686	Gaussian process regression
18	0.0364	Tree
19	0.1711	Tree
20	0.062367	Tree
21	0.060006	Tree
22	0.068301	Tree
23	0.1417	Tree
24	0.1701	Tree
25	0.135501905	Tree
26	0.1538	Ensemble
27	0.1537	Tree
28	0.1369	Tree
29	0.1408	Tree
30	0.141	Tree
31	0.1364	Tree
32	0.1363	Tree
33	0.1548	Tree
34	0.1534	Tree
35	0.1502	Tree
36	0.060006	Tree
37	0.0593	Tree

Table 5 summarizes the RMSE values and the corresponding training functions used in the tested preprocessing model configurations. The results show that several configurations reduced the RMSE relative to the baseline of 0.1715. Among the tested configurations, the lowest observed RMSE was obtained in Record 12, where the preprocessing sequence consisted of Clean Missing Data followed by Smooth Data, and the corresponding training function was Decision Tree Regression. However, Table 5 should not be interpreted as a fully controlled comparison among regression models. Because preprocessing sequences, model settings, and experimental configurations varied across trials, the observed performance differences cannot be attributed solely to the regression model type. Therefore, the result should be interpreted as evidence that this specific preprocessing–model configuration achieved the lowest RMSE within the tested set, rather than as a general conclusion that Decision Tree Regression is universally superior to the other regression models.

3.1 Impact of Data Preprocessing on Model Accuracy

Our experimental results unequivocally demonstrate that the comprehensive data preprocessing pipeline was critical for enhancing prediction accuracy. When evaluated on minimally processed data, our models yielded a baseline RMSE of 0.1715. As shown in our experimental records, the systematic application of our preprocessing methods—missing data imputation, outlier treatment, and data smoothing—resulted in a substantial and consistent reduction in RMSE. Missing Value Imputation: By filling gaps in the raw sensor data, we ensured data completeness, providing the models with a more continuous and robust dataset from which to learn patterns. Outlier Treatment: Removing and adjusting anomalous data points prevented extreme values from disproportionately skewing model training, leading to more generalizable, robust models. Data Smoothing: The application of a moving-average filter improved the signal-to-noise ratio, clarifying underlying trends and cyclical patterns that are essential for accurate analog selection and

regression. The most significant finding is the powerful synergistic effect of applying these methods in conjunction. The dramatic reduction in RMSE—in one case, from a baseline of 0.1715 to an exceptional 0.0011014 (a ~99.36% improvement)—is not attributable to any single step but to the integrated pipeline. This result underscores a core conclusion of our study: rigorous data quality management is a prerequisite for achieving high-accuracy, data-driven weather forecasting. The "noise" present in the raw data was a primary impediment to accurate prediction, and its effective removal unlocked the true predictive power of our models.

3.2 Performance Comparison of Regression Models

Our comparative analysis of regression algorithms revealed that the Decision Tree Regression model ("Tree Predict") consistently outperformed the other methods evaluated. After applying our data preparation pipeline, the Decision Tree model achieved the lowest RMSE of 0.0011014. In contrast, models such as Gaussian Process Regression (0.16794) and Linear Regression (0.1701) exhibited significantly higher errors under similar conditions. The superior performance of the Decision Tree model can be attributed to several key characteristics that are well-suited to meteorological time series data: Non-Linearity: Weather systems are inherently non-linear. Decision trees excel at capturing these complex relationships through their hierarchical splitting structure without requiring manual feature engineering. Robustness to Outliers: Even after preprocessing, some residual anomalies may persist. Decision trees are less sensitive to such outliers, as they tend to isolate extreme values in specific branches, minimizing their impact on the overall model. Interpretability: The flowchart-like structure of decision trees provides a transparent and interpretable model, which is highly valuable in scientific domains where understanding the "why" behind a prediction is crucial. The consistent success of the Decision Tree across various preprocessing combinations highlights its inherent suitability for this data type. However, we emphasize that the model's performance is fundamentally enabled by the quality of the input data, reinforcing the foundational role of our preprocessing pipeline.

3.3 Broader Implications and Study Limitations

The findings of this research have significant implications for the field of analog weather forecasting. Our approach advances prior work by introducing a systematic, resource-efficient methodology for handling data noise, demonstrating that high-accuracy forecasts can be achieved with modest local sensor inputs rather than computationally intensive global circulation data. Practical Advantages: A key advantage of our proposed method is its potential to democratize advanced weather forecasting. By using lightweight machine learning techniques in place of traditional Numerical Weather Prediction (NWP) ensembles, our approach enables high-accuracy forecasting in locations lacking supercomputing resources. Once trained, our models can generate forecasts with significantly lower computational cost and latency than NWP simulations. This could profoundly impact preparedness for disasters and water resource management in resource-constrained regions, particularly in the Global South. The findings of this study should be interpreted within a limited and clearly defined scope. The proposed framework is intended for localized, short-term analog weather pattern recognition based on station-level sensor data, rather than for national-scale operational weather forecasting. It is also not designed to predict extreme weather events. Because the preprocessing pipeline includes missing-value imputation, outlier treatment, normalization, and moving average smoothing, the reported RMSE values—especially the lowest value of 0.0011014—should be understood as errors measured on the normalized and preprocessed data scale, not as errors in the original physical units of meteorological variables. Therefore, the exceptionally low RMSE should not be interpreted as a near-perfect prediction of temperature, snowfall, or snow depth in real-world operational units. Although smoothing can improve pattern recognition by reducing random noise and clarifying local temporal trends, it may also attenuate short-term fluctuations and potentially obscure signals associated with rare or extreme events. For this reason, the proposed framework is better suited to recognizing local recurring weather patterns than to forecasting high-impact extreme events. In addition, as a purely data-driven regression approach, the model does not explicitly enforce physical laws governing the atmosphere. Unlike physics-based NWP systems, it cannot guarantee physical consistency across meteorological variables. Accordingly, the proposed method should be regarded as a complementary tool to support local analog forecasting and data preprocessing, rather than as a replacement for physics-based NWP.

3.4 Study Limitations

Generalizability: Our methodology was validated using data from a specific region (Oakland, California) and timeframe (2017-2019). Further research is required to assess its generalizability to more diverse climatic zones, such as tropical or polar regions. **Physical Consistency:** While our model achieves high statistical accuracy (a remarkably low RMSE of 0.0011014), purely data-driven models may not always adhere to the physical laws governing the atmosphere. Unlike physics-based NWP models, our model may "smooth out" unpredictable details and lacks guarantees of physical consistency, both of which are critical considerations for operational use. This exceptional accuracy also warrants further investigation to ensure no unintended data leakage occurred despite our rigorous walk-forward validation and to understand the impact of data normalization on the final error metric. **Future Directions:** This work opens several avenues for future research. The success of our data-driven approach underscores the potential to develop hybrid models that integrate our efficient preprocessing and regression techniques with physics-based NWP models. Such a hybrid system could leverage the speed and pattern-recognition strengths of machine learning to enhance the data assimilation or post-processing stages of NWP, potentially yielding forecasts that are both highly accurate and physically consistent. Further work should also focus on optimizing the pipeline's latency to enable seamless integration with real-time operational forecasting systems.

4. Conclusion

The proposed multi-stage preprocessing and regression framework provides a lightweight complementary approach for localized analog weather pattern recognition. The main contribution of this study is to demonstrate that careful preprocessing of noisy station-level sensor data can improve the stability and accuracy of local pattern-based prediction. However, the proposed approach is not intended to replace physics-based NWP systems, nor is it designed for national-scale operational forecasting or extreme weather prediction. Future work should further evaluate the method across broader climatic regions, examine its performance under extreme-event conditions, and explore hybrid integration with NWP models to improve both statistical accuracy and physical consistency.

5. Acknowledgments

Author Contributions: For research articles with several authors, a short paragraph specifying their individual contributions must be provided. The following statements should be used "Conceptualization, F.P. and P.V.; methodology, F.P. and P.V.; software, P.V.; validation, F.P., S.W., K.H.R. and P.V.; formal analysis, F.P. and P.V.; investigation, F.P., S.W., K.H.R. and P.V.; resources, P.V.; data curation, P.V.; writing—original draft preparation, F.P. and P.V.; writing—review and editing, F.P. and P.V.; visualization, P.V.; supervision, P.V.; project administration, F.P.; funding acquisition, F.P. and P.V.; All authors have read and agreed to the published version of the manuscript" Please turn to the CRediT taxonomy for the term explanation. Authorship must be limited to those who have contributed substantially to the work reported.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations: The following abbreviations are used in this manuscript: National Oceanic and Atmospheric Administration (NOAA); Numerical Weather Prediction (NWP); Root Mean Square Error (RMSE).

References

- [1] Bauer, P.; Thorpe, A.; Brunet, G. The Quiet Revolution of Numerical Weather Prediction. *Nature* **2015**, *525*, 47–55. <https://doi.org/10.1038/nature14956>
- [2] Chen, L.; Han, B.; Zhao, X.; Zhao, J.; Yang, W.; Yang, Z. Machine Learning Methods in Weather and Climate Applications: A Survey. *Appl. Sci.* **2023**, *13*(21), 12019. <https://doi.org/10.3390/app132112019>
- [3] Shen, B.-W.; Pielke, R. A.; Zeng, X.; Zeng, X. Lorenz's View on the Predictability Limit of the Atmosphere. *Encyclopedia* **2023**, *3*(3), 887–899. <https://doi.org/10.3390/encyclopedia3030063>
- [4] Jeworrek, J.; West, G.; Stull, R. Optimizing Analog Ensembles for Sub-Daily Precipitation Forecasts. *Atmosphere* **2022**, *13*(10), 1662. <https://doi.org/10.3390/atmos13101662>

- [5] Hu, W.; Cervone, G.; Young, G.; Monache, L. D. Machine Learning Weather Analogs for Near-Surface Variables. *Boundary-Layer Meteorol.* **2023**, *186*, 711–735. <https://doi.org/10.1007/s10546-022-00779-6>
- [6] Chattopadhyay, A.; Nabizadeh, E.; Hassanzadeh, P. Analog Forecasting of Extreme-Causing Weather Patterns Using Deep Learning. *J. Adv. Model. Earth Syst.* **2020**, *12*, e2019MS001958. <https://doi.org/10.1029/2019MS001958>
- [7] Skrynyk, O.; et al. Data Quality Control and Homogenization of Daily Precipitation and Air Temperature (Mean, Max, and Min) Time Series of Ukraine. *Int. J. Climatol.* **2023**, *43*(9). <https://doi.org/10.1002/joc.8080>
- [8] Mohammed, R.; Scholz, M. Quality Control and Homogeneity Analysis of Precipitation Time Series. *Atmosphere* **2023**, *14*(2), 197. <https://doi.org/10.3390/atmos14020197>
- [9] Fan, C.; Xiao, F.; Zhao, Y.; Wang, J. A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery from Building Operational Data. *Front. Energy Res.* **2021**, *9*, 652801. <https://doi.org/10.3389/fenrg.2021.652801>
- [10] Li, C.; Wang, K.; Zhang, F.; et al. Machine-Learning-Based Imputation Method for Filling Missing Values in Ground Meteorological Observation Data. *Algorithms* **2023**, *16*(9), 422. <https://doi.org/10.3390/a16090422>
- [11] Afrifa-Yamoah, E.; Birt, J.; Toohey, D. Missing Data Imputation of High-Resolution Temporal Climate Time Series Data. *Meteorol. Appl.* **2020**, *27*(1), e1873. <https://doi.org/10.1002/met.1873>
- [12] Šuljug, J.; Smajlović, S.; Pudar, M.; Kovačević, A.; Alić, B.; Beširević, N. A Comparative Study of Machine Learning Models for Predicting Meteorological Data in Agricultural Applications. *Electronics* **2024**, *13*(16), 3284. <https://doi.org/10.3390/electronics13163284>
- [13] Hasan, R. A.; Hashim, H.; Abed, A. M.; Sulaiman, M. H. Comparative Study: Using Machine Learning Techniques for Daily Rainfall Prediction in Australia. *AIP Conf. Proc.* **2023**, *2523*, 030002. <https://doi.org/10.1063/5.0148472>
- [14] Berliana, E. V.; Budiono, A. M.; Mahardika, M.; Purnomo, S. D. Comparative Analysis of Naïve Bayes, Support Vector Machine, and Decision Tree in Rainfall Classification Using Confusion Matrix. *Int. J. Adv. Comput. Sci. Appl.* **2024**, *15*(7), 92–98. <https://doi.org/10.14569/IJACSA.2024.0150755>
- [15] Thompson, V.; Philip, S.; Kew, S.; Pinto, I.; Vautard, R. Using Analogue Methods to Identify Trends in Circulation Patterns of Midlatitude Heatwaves. *Weather Clim. Extrem.* **2026**, *52*, 100898. <https://doi.org/10.1016/j.wace.2026.100898>
- [16] Alimisis, V.; Dimas, C.; Sotiriadis, P. P. A Low-Power Analog Integrated Euclidean Distance Radial Basis Function Classifier. *Electronics* **2024**, *13*(5), 921. <https://doi.org/10.3390/electronics13050921>
- [17] Al-Saedi, K.; Zhou, D.; Fish, A.; Tsakiri, K.; Marsellos, A. A Methodological Comparison of Forecasting Models Using KZ Decomposition and Walk-Forward Validation. *Mathematics* **2025**, *13*, 3410. <https://doi.org/10.3390/math13213410>
- [18] Wahyuddin, E. P.; Caraka, R. E.; Kurniawan, R.; Caesarendra, W.; Gio, P. U.; Pardamean, B. Improved LSTM Hyperparameters alongside Sentiment Walk-Forward Validation for Time Series Prediction. *J. Open Innov.: Technol., Mark., Complex.* **2025**, *11*, 100458. <https://doi.org/10.1016/j.joitmc.2024.100458>
- [19] Dash, C. S. K.; Behera, A. K.; Dehuri, S.; Ghosh, A. An Outliers Detection and Elimination Framework in Classification Task of Data Mining. *Decision Analytics Journal* **2023**, *6*, 100164. <https://doi.org/10.1016/j.dajour.2023.100164>
- [20] Mohammed, A. R.; Hassan, K. S.; Abdel-Aal, M. A. M. Moving Average Smoothing for Gregory-Newton Interpolation: A Novel Approach for Short-Term Demand Forecasting. *IFAC-PapersOnLine* **2022**, *55*(10), 749–754. <https://doi.org/10.1016/j.ifacol.2022.09.499>
- [21] Bag, A. A Comparative Study of Regression Algorithms for Predicting Graduate Admission to a University. *J. Network Commun. Emerg. Technol.* **2018**, *8*(10).
- [22] Yang, X.; Li, J.; Jiang, X. Research on Information Leakage in Time Series Prediction Based on Empirical Mode Decomposition. *Sci. Rep.* **2024**, *14*, 28362. <https://doi.org/10.1038/s41598-024-80018-9>