

SPATIAL CONSISTENCY QUALITY CONTROL OF DAILY SURFACE AIR TEMPERATURE OBSERVATIONS IN VIET NAM USING INVERSE DISTANCE WEIGHTING AND RANDOM FOREST

Bui Thi Khanh Hoa, Do Thi Anh Huyen, Nguyen Vinh Thu, Nguyen Viet Hung

⁽¹⁾Viet Nam Meteorological and Hydrological Administration

Received: 18 August 2026; Accepted: 17 September 2026

Abstract: This study developed and evaluated a spatial consistency quality control (QC) method for daily surface air temperature observations from 186 SYNOP stations across Viet Nam during 1961-2023. The variables examined included daily mean (T_{avg}), minimum (T_{min}), and maximum (T_{max}) temperatures. The proposed framework combines inverse distance weighting (IDW) with a machine learning (ML) model. IDW serves as the baseline reference, whereas ML is employed to better capture nonlinear relationships associated with spatial location, elevation, seasonality, and climatic region. Final QC flags were derived from the combined outputs of both models, allowing alerts to be stratified into method-specific detections, concurrent detections by IDW and ML, and high-suspicion cases. The ML model consistently outperformed IDW for all three variables. Its mean absolute error was 0.47°C for T_{avg} , 0.59°C for T_{min} , and 0.73°C for T_{max} , compared with 0.58°C , 0.74°C , and 1.00°C , respectively, for IDW. Under the IDW-ML framework, 98.87-99.10% of the evaluated observations were classified as good quality, while alert rates ranged from 0.90% to 1.13% and high-suspicion cases accounted for only 0.05-0.13%. Approximately one-third of all alerts were detected concurrently by both methods, thereby narrowing the subset of stations requiring priority manual review. These findings demonstrate that the combined IDW-ML approach provides an effective framework for spatial QC of daily temperature observations, particularly in networks characterized by complex terrain, uneven station distribution, and island stations, as in Viet Nam.

Keywords: Quality control (QC), spatial consistency checking (SCC), Viet Nam, inverse distance weighting (IDW), machine learning (ML).

1. Introduction

Surface air temperature is a fundamental meteorological variable and plays an important role in weather observation, forecasting, disaster risk warning, and climate-change research. Temperature observations from hydro-meteorological stations are widely used in scientific and operational applications; consequently, requirements for data accuracy, homogeneity, and reliability have become increasingly stringent. Therefore, their accuracy, consistency, and reliability are essential. Although daily temperature usually varies more

smoothly than precipitation, errors can still arise from sensor malfunction, changes in observing conditions, communication failures, data transmission problems, or processing mistakes. Such errors can affect climate databases, operational products, and subsequent climate analyses. For this reason, quality control (QC) is a necessary step before station temperature data are used for research or operational services [1].

Spatial consistency checking (SCC) is an important component of meteorological QC. It evaluates whether an observation is consistent with the surrounding temperature field estimated from neighbouring stations. Compared with simple physical-limit or temporal checks, SCC can identify observations that are physically plausible in isolation but

Corresponding author: Bui Thi Khanh Hoa
E-mail: khanhhoa303@gmail.com

inconsistent with the regional spatial pattern. This is particularly useful for daily temperature because temperature fields usually show meaningful spatial continuity and inter-station correlation. Traditional SCC methods include direct neighbour comparison, spatial interpolation, and spatial regression. Among these methods, IDW is widely used because it is simple, transparent, reproducible, and easy to audit in operational QC [2], [3].

However, distance-based methods have important limitations. In areas with complex terrain, strong elevation gradients, coastal effects, sparse station density, or climate-transition zones, geographical proximity does not always imply climatic similarity. Two nearby stations may have different thermal regimes because of elevation, exposure, land-sea influence, or local topography. Therefore, spatial QC for temperature should not rely only on distance. Additional information such as elevation difference, temperature-anomaly correlation, climate region, and station metadata is needed to construct a more reliable reference network. ML methods, particularly Random Forest (RF), can complement traditional approaches by incorporating spatial, topographic, seasonal, and neighbour-based predictors in a nonlinear framework, while IDW remains useful as a physically interpretable baseline.

In Viet Nam, station temperature data have been widely used in studies of hot days, heatwaves, climate variability, and temperature extremes. Previous studies have analysed daily maximum temperature in the North Central region, the relationship between hot days and large-scale circulation, and heatwave variability across the country [4-6]. Regional and global datasets such as SA-OBS and C-LSAT also include station data and apply QC and interpolation procedures to develop gridded climate products [7-9]. However, these studies mainly focus on climate analysis or gridded dataset construction, rather than on the development and evaluation of a station-day spatial QC algorithm for the Viet Nam SYNOP network. Existing QC studies in Viet Nam have also focused mainly on other variables

or applications, such as rainfall-gauge QC for quantitative precipitation estimation [10].

A dedicated SCC framework for daily SYNOP temperature observations in Viet Nam is therefore still needed. The Viet Nam station network includes complex mountains, long coastal zones, climate-transition areas, uneven station density, and island or offshore stations. These characteristics make neighbour selection and spatial interpretation more challenging than in a relatively homogeneous network. Inappropriate reference-station selection may increase false alerts, especially for highland, coastal, island, and network-edge stations. This study addresses this gap by developing and evaluating an IDW-ML spatial QC framework for daily Tavg, Tmin, and Tmax observations from the Viet Nam SYNOP network during 1961-2023. The novelty of the study lies in combining elevation-adjusted IDW, RF-based spatial prediction, robust residual analysis, and evidence-based combined QC flags. The framework not only compares the predictive performance of IDW and RF, but also separates alerts into IDW-only, ML-only, and concurrent IDW-ML evidence classes to support expert review rather than automatic data deletion.

2. Data and Methods

2.1. Data

2.1.1. Data Collection and Standardization

The input data for SCC comprise two main components: station observations and station metadata. The temperature variables comprise daily maximum (Tmax), minimum (Tmin), and average (Tavg) temperature from 186 surface meteorological stations in Viet Nam during 1961-2023. Metadata include geographic coordinates (longitude and latitude), station elevation, and station type; these variables are used to define spatial relationships and to support neighbour selection, elevation adjustment, and temperature interpolation [2]. Before applying SCC, the data are standardized by station identifier, observation date, unit, and numeric type. Duplicate station-date keys, physical limits, logical relationships among Tmin, Tmax, and Tavg, metadata validity, and isolated temporal

jumps are checked. Only valid observations with complete coordinates and elevation and without serious violations in previous QC stages are used as reference observations. Records that fail these conditions may still be evaluated but are not used as reference data for other stations.

2.1.2. Construction of the Spatial Network

The spatial reference network was constructed separately for Tavg, Tmin, and Tmax. For each target station, distances to all candidate stations were calculated using the Haversine formula. Candidate neighbours were then screened using four criteria: distance, elevation difference, metadata completeness, and data availability. In the operational configuration, candidate stations had to be located within a 200 km search radius; up to seven neighbouring stations were retained, and at least three valid neighbours were required to evaluate a target observation. These predefined settings were used to balance two requirements: obtaining sufficient neighbouring stations in a network with uneven station density, and avoiding reference estimates dominated by distant or physically non-representative stations. For island or offshore stations, where island/land metadata were available, stations from the same spatial class were preferred to avoid direct island-mainland comparisons. If fewer than three reliable neighbours were available, the observation was classified as not evaluable rather than being forced into a suspect class.

Neighbour selection was not based solely on geographic distance. Among the candidates, the algorithm prioritized stations that were closer to the target station, had smaller elevation differences, showed higher historical temperature-anomaly correlation, and belonged to the same climatic region where applicable. The anomaly correlation was calculated from deseasonalized temperature anomalies. Stations with $\rho \geq 0.6$ and sufficient overlapping samples were preferred. When this correlation criterion could not be fully satisfied, a fallback neighbour set based on distance and elevation was allowed, but the minimum-neighbour requirement was still retained. This design

was used to reduce physically inappropriate comparisons in mountainous, coastal, island, and network-edge areas.

2.2. Elevation correction and spatial consistency indicators

Before spatial interpolation, the temperature at each neighbouring station was adjusted to the elevation of the target station:

$$X_{j \rightarrow i, t, v}^* = X_{j, t, v} + r_v \times \frac{z_i - z_j}{1000} \quad (1)$$

Where, $X_{j, t, v}$ is the observed temperature of variable v at neighbouring station j and time t , $X_{j \rightarrow i, t, v}^*$ is the elevation-adjusted value, z_i and z_j are the elevations of the target and neighbouring stations in metres, and r_v is the empirical temperature-elevation gradient in $^{\circ}\text{C km}^{-1}$.

The gradients were determined separately for Tavg, Tmin, and Tmax from relationships in the quality-controlled historical SYNOP network during the model-development stage. The coefficients were then kept fixed in the SCC experiment and were not re-estimated from the target observation being evaluated. In the operational run, the empirical gradients were $-5.710^{\circ}\text{C km}^{-1}$ for Tavg, $+5.683^{\circ}\text{C km}^{-1}$ for Tmin, and $-6.106^{\circ}\text{C km}^{-1}$ for Tmax. These values were used only to normalize neighbouring-station temperatures to the elevation of the target station before spatial estimation, not as a comprehensive climatological analysis of lapse-rate variability. Variable-specific gradients were used because mean, minimum, and maximum temperatures do not exhibit identical spatial-topographic behaviour.

The IDW reference value was calculated from the selected elevation-adjusted neighbours as:

$$\hat{X}_{i, t, v}^{IDW} = \frac{\sum_{j \in N_i(v)} w_{ij} X_{j \rightarrow i, t, v}^*}{\sum_{j \in N_i(v)} w_{ij}}, \text{ with } w_{ij} = d_{ij}^{-p} \quad (2)$$

Where $N_i(v)$ is the neighbour set for station i and variable v , d_{ij}^{-p} is the distance between stations i and j , p is the distance-weighting exponent, and w_{ij} is the IDW weight. The IDW spatial residual was defined as:

$$r_{i,t,v}^{IDW} = X_{i,t,v} - \hat{X}_{i,t,v}^{IDW} \quad (3)$$

Residual abnormality was evaluated using a robust Z-score based on the median and median absolute deviation:

$$Z_{i,t,v}^{IDW} = \frac{r_{i,t,v}^{IDW} - \text{median}(r_{g,v}^{IDW})}{1.4826 \times \text{MAD}_{g,v}^{IDW}} \quad (4)$$

Where $r_{i,t,v}^{IDW}$ is the IDW residual for station i , time t , and variable v ; median $r_{g,v}^{IDW}$ is the median IDW residual of the corresponding reference group g ; and $\text{MAD}_{g,v}^{IDW}$ is the median absolute deviation of IDW residuals in that group. The factor 1.4826 scales MAD to be comparable with the standard deviation under an approximately normal distribution.

The reference group g was defined hierarchically, with priority given to station-month statistics, followed by station-season, climate-region-month, and network-month statistics when the sample size was insufficient. A QC flag was assigned using both absolute residual thresholds and robust Z-score thresholds. Four states were used: -1, not evaluable; 0, pass; 1, suspect; and 2, high suspicion.

The three temperature variables were processed separately, but their physical dependence was retained through internal consistency screening. Records violating the condition $T_{\min} \leq T_{\text{avg}} \leq T_{\max}$ were marked for review and excluded from the reference pool. After QC was completed for each variable, the final station-day flag was defined as the highest warning level among T_{avg} , $T_{\min}$, and $T_{\max}$, while the variable responsible for the alert was retained for interpretation.

2.3. SCC algorithm, Random Forest validation, and leakage prevention

The complete workflow of the SCC framework is summarized in Figure 1. The workflow includes data screening, neighbour selection, elevation correction, IDW estimation, RF prediction, residual calculation, combined flagging, and validation.

The machine learning component of the SCC framework was based on an RF model. Separate

RF models were developed for T_{avg} , $T_{\min}$, and $T_{\max}$, and their performance was evaluated using an expanding-window temporal validation design. For the 1961–2023 dataset, the first fold used 1961–1970 for training and 1971–1975 for testing. The training window was then expanded by five years at each step, with each subsequent test block occurring after the corresponding training period. The final fold used 1961–2020 for training and 2021–2023 for testing. In all folds, the test period was later than the training period, so future observations were not used to predict records.

For each fold, the RF model was fitted only with records satisfying the training-quality criteria, including non-missing target values, valid station metadata, sufficient neighbouring-station information, and valid preliminary screening status. The contemporaneous value at the target station was excluded from the predictor set. Predictor variables included station location, elevation, seasonal indicators, lagged target-station values, and summary statistics from valid neighbouring stations, such as neighbour mean, median, spread, distance-weighted temperature, number of neighbours, distance metrics, and anomaly-correlation information. Contemporaneous observations from neighbouring stations were retained because they represent independent spatial information, but the value being checked was not used as its own predictor.

For each validation fold, the RF prediction was compared with the observed value to obtain the out-of-sample residual:

$$r_{i,t,v}^{RF} = X_{i,t,v} - \hat{X}_{i,t,v}^{RF} \quad (5)$$

Where, $\hat{X}_{i,t,v}^{RF}$ is the RF-estimated temperature. Model performance was evaluated using MAE, RMSE, Bias, and R^2 . RF residuals were standardized using the same robust framework as the IDW residuals. Conservative absolute thresholds were also applied: 3.0°C and 4.5°C for suspect and high-suspicion T_{avg} , and 4.0°C and 6.0°C for suspect and high-suspicion $T_{\min}$ and $T_{\max}$, respectively. After out-of-sample validation, final RF models were retrained using

the clean reference data and stored separately for the three variables together with predictors, neighbour sets, metadata, configuration, and QC thresholds.

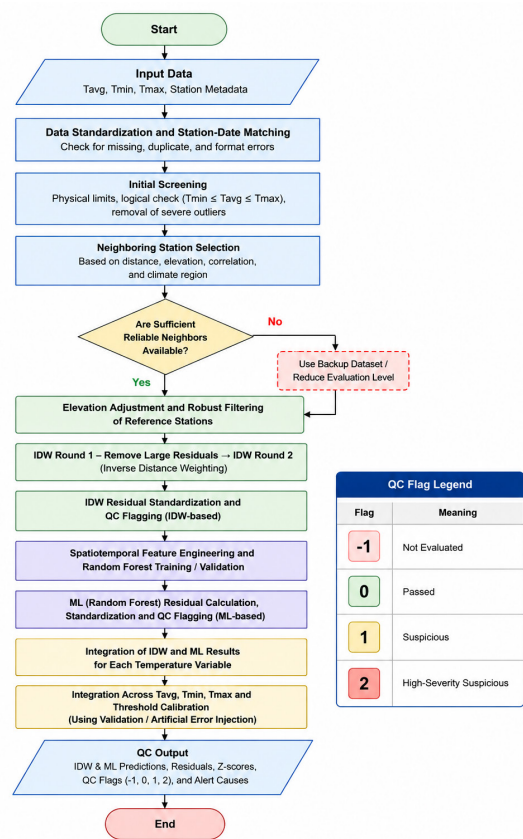


Figure 1. Workflow of the proposed SCC framework for daily SYNOP temperature observations

2.4. Combined flagging, evaluation, and model stability

The IDW and ML outputs were merged by station and date. If neither method produced an alert, the combined flag was set to 0. If one method produced a suspect flag, the combined flag was set to 1. If either method produced a high-suspicion flag, the combined flag was set to 2. When both IDW and ML produced suspect flags and their residuals had the same sign, the observation was upgraded to high suspicion because two independent spatial references indicated departure in the same direction. A high-suspicion flag indicates stronger evidence of spatial inconsistency, but it does not by itself confirm that the observation is erroneous.

Performance was assessed using coverage, prediction accuracy, alert characteristics, artificial-error detection, and temporal stability. Coverage describes the proportion of station-day records for which a valid spatial reference and QC flag could be produced. Prediction accuracy was summarized by MAE, RMSE, Bias, and R². Because records classified as not evaluable do not provide a defensible spatial QC decision, alert statistics were calculated on the evaluable subset:

$$Alert\ rate = \frac{n_{suspect} + n_{high}}{e_{evaluable}} \times 100\% \tag{6}$$

Where, $n_{suspect}$ and n_{high} are the numbers of records flagged as suspect and high suspicion, respectively, and $n_{evaluable}$ is the number of records successfully evaluated by the SCC framework. The not evaluable rate was reported separately to describe limitations caused by sparse neighbours, missing metadata, insufficient data availability, or limited model coverage.

Because complete expert-confirmed labels were not available for all historical observations, controlled artificial errors with different magnitudes were inserted into copies of the data to evaluate detection capability. Recall, Precision, FPR, and F1-score were used to summarize the response of the QC system to these known perturbations. Where station-history information, equipment logs, expert-confirmed cases, or original observation documents were available, flagged records were compared with those independent sources. In the absence of complete independent labels, QC flags were interpreted as evidence for expert review rather than automatic confirmation of erroneous observations.

Model stability over 1961-2023 was examined by comparing coverage, prediction errors, residual distributions, alert rates, and high-suspicion rates across variables, years, seasons, climatic regions, and special station groups. This evaluation was used to identify periods or stations where model behaviour may be affected by changes in station density, data coverage, metadata quality, or local representativeness.

Special stations, including island or offshore stations, coastal stations, highland and mountainous stations, network-edge stations, and stations with unusually high alert rates, were therefore interpreted conservatively. For these stations, final assessment required inspection of time series, appropriate neighbouring stations, metadata, weather context, station history, and original observation records.

3. Results

The standardized dataset contains daily Tavg, Tmin, and Tmax observations from 186 surface meteorological stations in Viet Nam during 1961-

2023, comprising 4,279,860 station-day records.

3.1. Results of IDW QC

Observed and IDW-estimated values are generally close to the 1:1 line for Tavg, Tmin, and Tmax (Figure 2a-c), although dispersion increases from Tavg to Tmax. MAE is 0.58, 0.74, and 1.00°C, while RMSE is 0.87, 1.10, and 1.44°C for Tavg, Tmin, and Tmax, respectively. Bias is 0.03, 0.08, and -0.04°C, indicating negligible systematic bias. These results show that IDW provides a reasonable spatial reference for daily SYNOP temperature, with lower accuracy for Tmax.

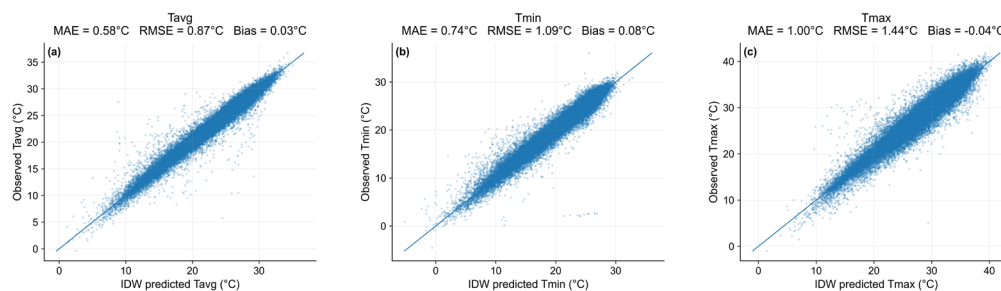


Figure 2. Scatterplots of observed and IDW-estimated Daily SYNOP temperature

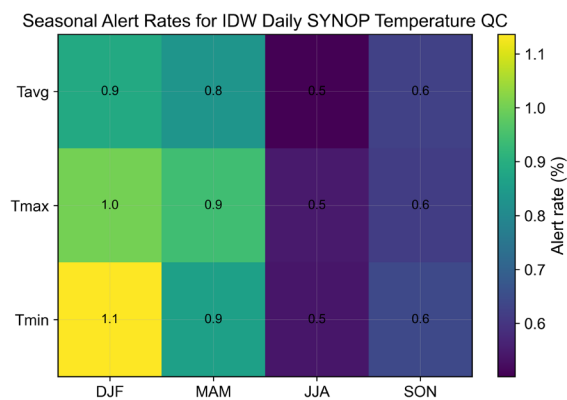


Figure 3. Seasonal alert rates for IDW Daily SYNOP Temperature QC

Figure 3 presents seasonal IDW alert rates for Tavg, Tmin, and Tmax. DJF, MAM, JJA, and SON denote December-January-February, March-April-May, June-July-August, and September-October-November, respectively. IDW alerts are generally low but show clear seasonality, with higher rates in DJF and MAM and lower rates in JJA. The highest alert rates occur in DJF, particularly for Tmin (1.1%) and Tmax (1.0%), while JJA has the lowest rate, approximately 0.5% for all three variables.

3.2. Results of ML QC

The ML-based SCC results are evaluated using scatterplots (Figure 4a-c). ML produces lower errors than IDW for all three variables. MAE is 0.47, 0.59, and 0.73°C for Tavg, Tmin, and Tmax, respectively, while RMSE is 0.69, 0.83, and 1.02°C. Bias values are close to zero, indicating no notable systematic bias. The improvement is most evident for Tmax, suggesting that the ML component provides useful additional information for variables with stronger spatial variability.

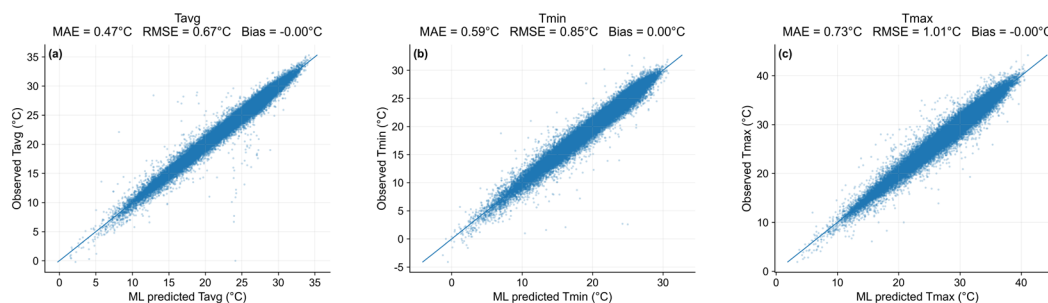


Figure 4. Scatterplots of observed and ML-estimated Daily SYNOP Temperature

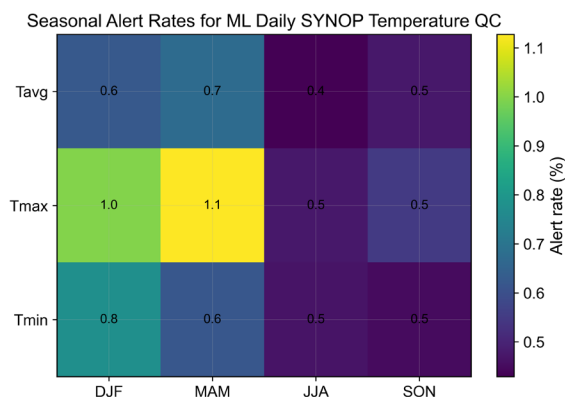


Figure 5. Seasonal alert rates for ML Daily SYNOP Temperature QC

Figure 5 shows the seasonal alert rates from ML. ML alert rates are lower than IDW for Tavg and Tmin in most seasons. For Tmax, alert rates remain higher in DJF and MAM and decrease in JJA and SON. Overall, the seasonal pattern is consistent with the IDW results, although the magnitude of alerts differs between the two methods.

3.3. Combined IDW-ML QC spatial consistency results

The residual plot comparing IDW and ML shows a clear positive relationship for Tavg, Tmin, and Tmax (Figure 6a-c). Most points are

concentrated near small residuals, while records far from the main cluster represent cases requiring higher review priority. The combined framework therefore provides additional information by distinguishing single-method alerts from concurrent IDW-ML alerts.

Figure 7 summarizes the evidence classes of the combined QC. Records flagged by only one method are more common than records flagged simultaneously by both methods, indicating complementary behavior. The group flagged by both methods with residuals in the same direction is small but has stronger QC evidence.

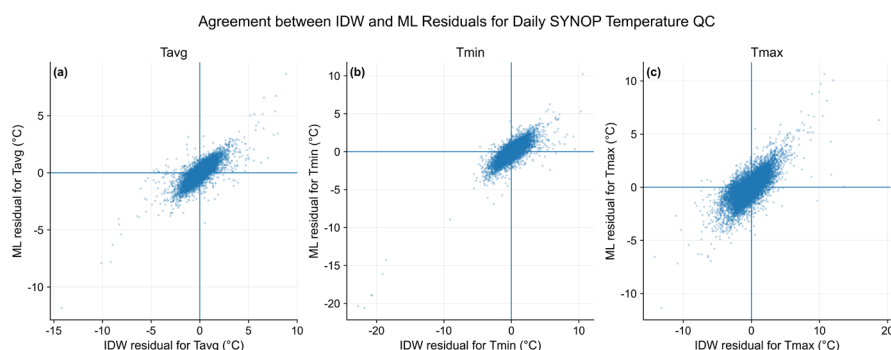


Figure 6. Agreement between IDW and ML Residuals for Daily SYNOP Temperature QC

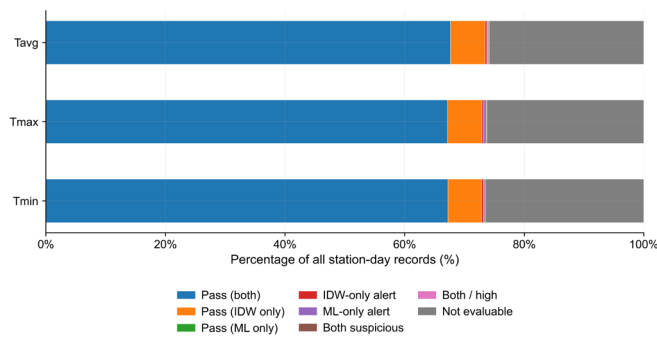


Figure 7. Distribution of evidence classes for the combined QC method

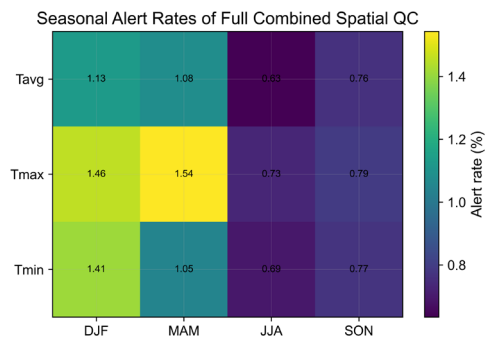


Figure 8. Seasonal Alert Rates of Full Combined Spatial QC

Figure 8 presents seasonal alert rates of the combined QC. Combined alert rates are higher than those of each method because records detected by either IDW or ML are retained. The highest rates occur in DJF and MAM, particularly for Tmax, while JJA shows the lowest rates. Among evaluable records, the combined pass rate reaches 99.10% for Tavg, 99.02% for Tmin, and 98.87% for Tmax. The combined alert rate ranges from 0.90% to 1.13%, and the high-suspicion rate remains low, at 0.05–0.13%.

Figure 9 shows the spatial distribution of alert rates from the combined QC. Most stations have low alert rates, whereas a limited number of stations show higher values. These stations are generally located in areas with complex terrain, coastal influence, network-edge conditions, or distinctive local climatic characteristics. Therefore, spatial QC flags are interpreted as indicators for prioritized review rather than automatic evidence of erroneous observations.

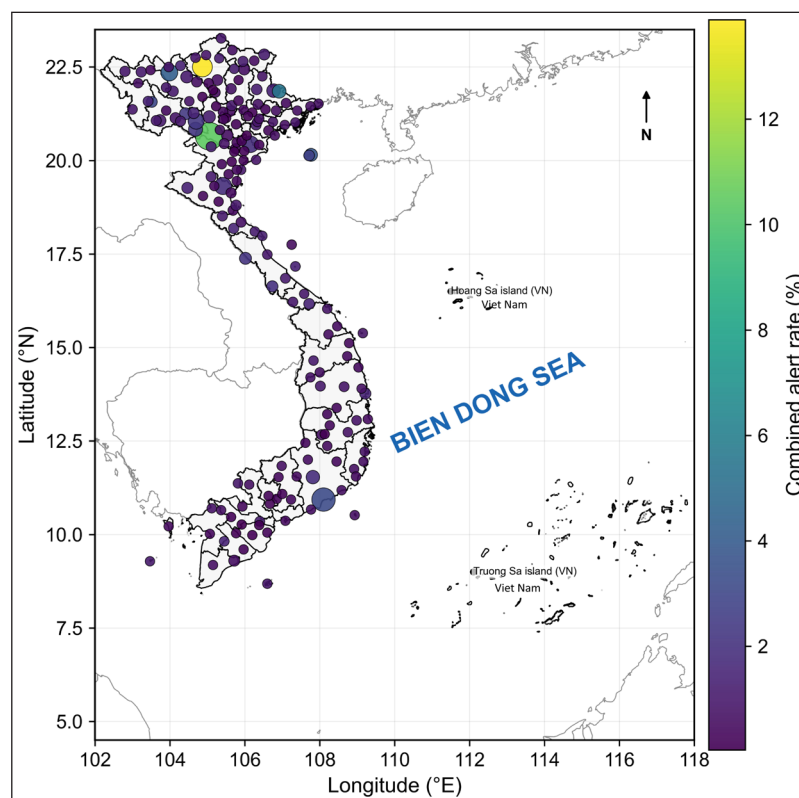


Figure 9. Spatial distribution of alert rates from the combined QC

A detailed diagnostic procedure was applied to stations with notable alert patterns. Table 1 summarizes representative stations and their diagnostic classes. Priority group A includes Phan Thiet, Tay Hieu, Co To, which show relatively high alert rates and agreement between IDW and ML for key variables. Group B includes stations such as Bac Quang, and Sa Pa,

where alert rates are notable but interpretation requires caution because of complex terrain or spatial representativeness of neighbouring stations. Group C includes stations with low coverage or no ML prediction, such as Hon Ngu and Truong Sa. These cases should be treated as insufficiently evaluated rather than pass or anomaly-free.

Table 1. Detailed diagnosis of temperature stations with notable characteristics

Station	Variable	ML Coverage	Alert Rate (%)	Season	Dominant diagnostic	Combined QC Relationship	Diagnostic interpretation	Evidence strength	Review priority
Phan Thiet	Tavg	71.3	6.52	SON	High ML residual	Both alert, same direction for Tavg	Tavg-specific anomaly	Medium-high	A
Tay Hieu	Tmin, Tmax	≈83.6	Tmin 2.83; Tmax 2.18	DJF	High ML residual	Both alert, same direction	Temperature extreme anomaly	High	A
Co To	Tavg, Tmin, Tmax	≈84.1	1.99 - 2.83	SON	Z-score exceedance	Both alert, same direction for Tavg, Tmax	Multi-variable anomaly	High	A
Bac Quang	All variable	≈84	2.95 - 4. 21	DJF	Absolute threshold, Z-score	Evidence from multiple methods	Multi-variable anomaly	Medium	B
Sa Pa	Tavg, Tmin	≈83.4	1.71 - 2.72	DJF	Absolute threshold exceedance	Neighbouring stations require review	Distinctive highland station	Medium	B
Hon Ngu	Not determinable	1.6	Not representative	MAM	Absolute threshold exceedance	Insufficient coverage	Low coverage	Low	C
Truong Sa	Not evaluated	0	-	-	No ML prediction	Not evaluable	No ML coverage	No conclusion	C

3.4. Evaluating algorithm performance by injecting errors

Artificial error injection was used as a controlled validation experiment to assess the detection capability of the SCC framework. Known temperature perturbations were inserted into copies of the observations and evaluated using the same QC procedure. The experiment covered both small perturbations, which may overlap with natural spatial variability, and larger departures that are more likely to indicate observation or processing errors.

Table 2 shows the detection performance of the combined SCC framework across all injected error magnitudes. Detection generally increased with error magnitude. Perturbations of $\pm 1^{\circ}\text{C}$ and $\pm 2^{\circ}\text{C}$ were more difficult to detect, whereas moderate to large perturbations were identified more consistently. For Tavg, detection reached 98.0% at $\pm 3^{\circ}\text{C}$ and nearly complete detection at larger magnitudes. For Tmin and Tmax, detection was lower for small perturbations but increased substantially for $\pm 5^{\circ}\text{C}$ to $\pm 10^{\circ}\text{C}$.

Table 2. Detection rate of the combined SCC framework under different artificial-error magnitudes

Variable	$\pm 1^{\circ}\text{C}$	$\pm 2^{\circ}\text{C}$	$\pm 3^{\circ}\text{C}$	$\pm 5^{\circ}\text{C}$	$\pm 7^{\circ}\text{C}$	$\pm 10^{\circ}\text{C}$
Tavg	9.0%	66.5%	98.0%	99.5%	100%	100%
Tmin	6.0%	31.0%	78.0%	96.5%	98.0%	99.0%
Tmax	2.5%	18.5%	65.5%	94.0%	97.5%	98.0%

Table 3. Comparison of IDW, ML, and combined QC performance at selected artificial-error levels

Variable	Error level	IDW	ML	IDW-ML
Tavg	$\pm 3^{\circ}\text{C}$	96.5%	93.0%	98.0%
Tmin	$\pm 5^{\circ}\text{C}$	93.0%	93.5%	96.5%
Tmax	$\pm 5^{\circ}\text{C}$	88.0%	90.0%	94.0%
Tmax	$\pm 7^{\circ}\text{C}$	96.0%	96.0%	97.5%

Table 3 compares IDW, ML, and the combined method at selected perturbation levels. The combined method generally achieved higher detection than either individual method, especially for moderate departures. Missed cases were mainly associated with records lacking sufficient neighbours or ML predictions, and with small perturbations that did not exceed the residual-based thresholds. Overall, the experiment indicates that the proposed framework is most effective for detecting moderate to large spatial inconsistencies while maintaining a conservative response to small deviations.

4. Discussion

The results show that spatial consistency checking is a useful additional QC layer for daily SYNOP temperature observations in Viet Nam. The main value of the proposed framework is not only its detection capability, but also its ability to provide interpretable QC evidence. This is important for long-term climate data, where suspicious observations should be reviewed in relation to neighbouring stations, metadata, station history, and weather conditions, rather than being automatically removed. This approach is consistent with modern climate-data QC systems that combine multiple checks to detect errors while preserving genuine meteorological extremes [1], [13].

A key strength of the study is the complementary use of IDW and Random

Forest. IDW provides a transparent and reproducible spatial baseline that can be easily audited in operational practice. However, distance alone is not always sufficient in Viet Nam, where the station network includes mountains, coastal zones, climate-transition areas, islands, and regions with uneven station density. In such conditions, two nearby stations may not necessarily represent the same thermal regime. This limitation of distance-based spatial estimation has also been noted in previous studies on temperature QC and spatial interpolation [4], [11]. The RF model helps address part of this problem by incorporating elevation, location, season, climate region, and neighbouring station statistics. Therefore, ML is used here as an additional nonlinear reference, not as a replacement for physically interpretable QC.

The combined IDW-ML flagging strategy is the main methodological contribution of this study. A single model may produce alerts that reflect either real observation problems or limitations of the reference field. By separating IDW-only, ML-only, and concurrent IDW-ML alerts, the proposed framework provides a clearer hierarchy of evidence. Records flagged by both methods, especially when residuals have the same sign, provide stronger support for expert review. In contrast, single-method alerts are retained as lower-confidence evidence because they may reflect local representativeness, neighbour selection, or model sensitivity. This

graded structure is more suitable for operational QC than a simple pass/fail decision.

The explicit treatment of not evaluable records is also important. In spatial QC, an observation should only be judged when a defensible reference can be constructed. Records without enough reliable neighbours, valid metadata, or ML prediction should not be treated as pass records. This distinction is particularly relevant for early historical periods, sparse regions, island stations, and network-edge stations. Separating “not evaluable” from “pass” helps avoid overconfidence in parts of the dataset where the spatial reference is weak.

The separate processing of Tavg, Tmin and Tmax is justified by their different physical behaviours. Tavg generally has a smoother spatial structure, whereas Tmax is more sensitive to radiation, cloudiness, surface conditions and station exposure. Tmin can be influenced by nocturnal cooling, terrain sheltering, cold-air accumulation and local inversion conditions. These differences support the use of variable-specific elevation gradients, residual distributions and thresholds. At the same time, the internal consistency check among Tmin, Tavg and Tmax is necessary to prevent physically inconsistent records from being used as references for neighbouring stations.

The framework has practical value for both operational QC and historical data consolidation. It can help identify records requiring review before data release, prioritize stations with persistent or high-confidence alerts, and support metadata checking when unusual residual patterns appear. For example, repeated same-direction residuals may indicate issues related to station relocation, elevation metadata, sensor exposure, observation practice, or the suitability of neighbouring stations. However, the QC flag itself should be considered as evidence for investigation, not as a final error label. Final decisions still require comparison with original observation records, station metadata, equipment logs, neighbouring stations, and actual weather conditions, consistent with the broader view that network QA depends on both statistical tests and station-

level knowledge [12].

Several limitations should be acknowledged. First, complete expert-confirmed error labels are not available for the full 1961–2023 period. The artificial error experiments provide controlled validation, but they cannot fully represent all real-world observation problems, such as gradual sensor drift, undocumented station relocation, changes in observation practice, or changes in the local environment around a station. Second, the reliability of spatial QC depends on the representativeness of neighbouring stations. This remains challenging in mountainous, coastal, island, and climate-transition environments, even after screening by distance, elevation difference, anomaly correlation, and climate region. Third, the RF component depends on the quality and temporal stability of the training data. Although the expanding-window validation was designed to reduce temporal data leakage, the Viet Nam SYNOP network is not temporally uniform. Early training folds may be less representative of regions where station density increased later, particularly in parts of central and Southern Viet Nam, highland areas, coastal zones, and island stations. Therefore, RF performance should be interpreted together with prediction coverage, not evaluable records, regional error behaviour, and station-level diagnostics, rather than as a uniform measure of model skill across all regions and periods.

Future work should therefore focus on strengthening independent validation and improving the treatment of special station environments. Priority should be given to linking QC flags with station-history documents, equipment maintenance records, observer notes and expert-confirmed error cases. Additional supporting data, such as nearby AWS observations, reanalysis products or high-resolution gridded temperature fields, may help interpret difficult cases, but they should not replace verification from station records. Methodologically, future studies should compare the present IDW-RF framework with spatial regression, kriging, graph-based approaches, and explainable ML methods. Adaptive thresholds by

season, climate region, and station type should also be explored, especially for highland, coastal, and island stations.

Overall, the proposed framework provides a practical and interpretable approach for spatial QC of daily SYNOP temperature observations in Viet Nam. Its novelty lies in combining elevation-adjusted IDW, RF prediction, robust residual analysis, and evidence-based combined flagging in a single workflow. The framework improves the detection of moderate to large spatial inconsistencies while maintaining the conservative interpretation required for long-term climate datasets.

5. Conclusions

This study developed and evaluated a spatial QC framework for daily SYNOP station temperature data in Viet Nam, covering T_{avg} , T_{min} , and T_{max} at 186 stations during 1961–2023. The framework combines elevation-adjusted IDW interpolation with a machine learning model and incorporates metadata, neighbour selection, robust residual standardization, and variable-specific thresholds. The approach evaluates each observation relative to its surrounding spatial field and is designed for a station network characterized by complex terrain, long coastlines, island stations, and uneven station density.

Both IDW and ML reproduce the spatial structure of daily temperature, with ML achieving lower errors for all three variables. ML MAE is 0.47°C for T_{avg} , 0.59°C for T_{min} , and

0.73°C for T_{max} , compared with 0.58°C , 0.74°C , and 1.00°C for IDW. For evaluable records, the combined QC pass rate is 98.87–99.10%, the alert rate is 0.90–1.13%, and the high-suspicion rate is 0.05–0.13%. These results indicate selective rather than widespread flagging.

The main contribution of the study is an evidence-stratified QC structure rather than a binary pass/fail decision. IDW provides a transparent baseline, while ML adds capacity to represent nonlinear relationships associated with terrain, season, and climate region. Separating IDW-only, ML-only, concurrent IDW-ML, and high-suspicion cases helps data controllers prioritize review, particularly for highland, coastal, island, and network-edge stations.

Overall, the Combined QC framework can effectively support spatial quality control of daily SYNOP temperature observations in Viet Nam. Its value lies not in automatically declaring observations correct or incorrect, but in stratifying suspicion, providing complementary evidence, and directing expert review. QC flags should therefore be used as review-support indicators rather than automatic grounds for correction or deletion. Future work should verify flagged records using station records, equipment logs, original observations, weather event information, and independent data sources, and should extend comparisons with other spatial QC methods such as spatial regression, kriging, and more interpretable ML models.

Author contribution: Conceptualization, Bui Thi Khanh Hoa and Nguyen Vinh Thu; methodology, Bui Thi Khanh Hoa, Do Thi Anh Huyen, and Nguyen Vinh Thu; software, Bui Thi Khanh Hoa and Nguyen Viet Hung; validation, Bui Thi Khanh Hoa, Do Thi Anh Huyen, and Nguyen Vinh Thu; visualization, Bui Thi Khanh Hoa and Nguyen Viet Hung; writing - review and editing, Bui Thi Khanh Hoa, Do Thi Anh Huyen, Nguyen Vinh Thu, Nguyen Viet Hung.

Acknowledgement: This study was supported by the Ministry-level Science and Technology Research Project "Research on the Development of Tools for Quality Assessment and Processing Procedures of Temperature and Rainfall Observation Data from Meteorological and Hydrological Stations" (Grant No. TNMT.2024.06.05), funded by the Ministry of Agriculture and Environment of Viet Nam.

Declarations: All authors declare that they have no conflicts of interest. The authors confirm that our manuscript is original research and has not been published and is not under consideration for publication elsewhere.

References

- [1] I. Durre, M. J. Menne, B. E. Gleason, T. G. Houston, and R. S. Vose, "Comprehensive automated quality assurance of daily surface observations," *Journal of Applied Meteorology and Climatology*, vol. 49, no. 8, pp. 1615-1633, 2010, doi: 10.1175/2010JAMC2375.1.
- [2] K. G. Hubbard and J. You, "Sensitivity analysis of quality assurance using the spatial regression approach-A case study of the maximum/minimum air temperature," *Journal of Atmospheric and Oceanic Technology*, vol. 22, no. 10, pp. 1520-1530, 2005, doi: 10.1175/JTECH1790.1.
- [3] P. Goovaerts, "Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall," *Journal of Hydrology*, vol. 228, nos. 1-2, pp. 113-129, 2000, doi: 10.1016/S0022-1694(00)00144-X.
- [4] P. T. Ly and H. L. T. Thuy, "Spatial distribution of hot days in north central region, Vietnam in the period of 1980-2013," *Vietnam Journal of Earth Sciences*, vol. 41, no. 1, pp. 36-45, 2019, doi: 10.15625/0866-7187/41/1/13544.
- [5] T. H. Chinh, P. V. Tan, M. V. Khiem, V. V. Thang, H. P. Lam, W. Lohpaisankrit, et al., "Spatiotemporal variability of hot days in association with the large-scale atmospheric drivers over Vietnam," *Advances in Meteorology*, vol. 2023, Art. no. 6612199, 2023, doi: 10.1155/2023/6612199.
- [6] P. T. Ly, P. T. Ha, P. V. Tan, and V. T. Yen, "Variability of heatwaves across Vietnam in recent decades," *Vietnam Journal of Earth Sciences*, vol. 45, no. 4, pp. 517-530, 2023, doi: 10.15625/2615-9783/19057.
- [7] E. J. M. van den Besselaar, G. van der Schrier, R. C. Cornes, A. S. Iqbal, and A. M. G. Klein Tank, "SA-OBS: A daily gridded surface temperature and precipitation dataset for Southeast Asia," *Journal of Climate*, vol. 30, no. 14, pp. 5151-5165, 2017, doi: 10.1175/JCLI-D-16-0575.1.
- [8] X. Yun, B. Huang, J. Cheng, W. Xu, S. Qiao, and Q. Li, "A new merge of global surface temperature datasets since the start of the 20th century," *Earth System Science Data*, vol. 11, no. 4, pp. 1629-1643, 2019, doi: 10.5194/essd-11-1629-2019.
- [9] S. Wei, Q. Li, Q. Xu, Z. Li, H. Zhang, and J. Lin, "Updates to C-LSAT 2.1 and the development of high-resolution land surface air temperature and diurnal temperature range datasets," *Earth System Science Data*, vol. 17, no. 9, pp. 4985-5005, 2025, doi: 10.5194/essd-17-4985-2025.
- [10] M. Tonouchi, T. K. H. Bui, V. H. Nguyen, and M. C. Nguyen, "Quality check of rain gauge data for quantitative precipitation estimate," *Vietnam Journal of Hydrometeorology*, vol. 6, no. 15, pp. 21-27, 2023, doi: 10.36335/VNJHM.2023(15).21-27.
- [11] You et al., "Comparison of methods for spatially estimating station temperatures in a quality control system," *International Journal of Climatology*, vol. 28, no. 6, pp. 777-787, 2008, doi: 10.1002/joc.1571.
- [12] M. A. Shafer, C. A. Fiebrich, D. S. Arndt, S. E. Fredrickson, T. W. Hughes, et al., "Quality assurance procedures in the Oklahoma Mesonet," *Journal of Atmospheric and Oceanic Technology*, vol. 17, no. 4, pp. 474-494, 2000, doi: 10.1175/1520-0426(2000)017<0474>2.0.CO;2.
- [13] R. J. H. Dunn, K. M. Willett, P. W. Thorne, E. V. Woolley, I. Durre, A. Dai, et al., "HadISD: A quality-controlled global synoptic report database for selected variables at long-term stations from 1973-2011," *Climate of the Past*, vol. 8, no. 5, pp. 1649-1679, 2012, doi: 10.5194/cp-8-1649-2012.