World Journal of Environmental Biosciences
Available Online at: **www.environmentaljournals.org**

Volume 15, Issue 3: 102-112

https://doi.org/10.51847/zMqqEu22RP

ISSN 2277- 8047

Multiple Linear Regression Outperforms Machine Learning for Monthly Temperature Forecasting across the Peruvian Altiplano

Leonel Coyla-Idme[1]*, Vidman Ruis Roque-Mamani[1], Smit Alexander Suni-Morales[1], Keysi Salcca-Lagar[1], Alex Arias-Ramírez[1], Antony Jhonatan Flores-Nina[1], Richar Andre Vilca-Solorzano[1]

[1]Professional School of Statistical and Informatics Engineering, Universidad Nacional del Altiplano de Puno, Puno, Peru.

ABSTRACT

Monthly temperature forecasts underpin agricultural planning and frost risk management in the Peruvian Altiplano, yet studies in the region rarely test whether machine learning improves on the standard climatological reference. This study quantifies that improvement at the Puno Principal Climatological Station using 7,279 daily records from 2003 to 2024, and replicates it across eleven independent Altiplano series. Observations were screened, aggregated under the World Meteorological Organization completeness rule, and gaps filled with a calibrated estimator fitted only on training years. Three models were tuned by blocked time-series cross-validation and verified against monthly climatology and persistence. At Puno all three models beat climatology, cutting mean absolute error from 0.597 °C to between 0.415 °C and 0.476 °C, while persistence proved significantly worse than climatology. The neural advantage did not replicate: across 50 random initialisations the perceptron averaged 0.495 °C, and only 4 of 50 beat the deterministic linear model. Linear regression was also the only model whose residuals were normally distributed and serially independent, whereas random forest left a lag-one autocorrelation of 0.427 unexploited. Across 792 test months linear regression was best at 9 of 11 sites, and learning curves extended to 2,475 training months showed the machine learning curves flattening above the linear model without crossing it. Predictive information came almost entirely from combining target-month climatology with recent thermal state; removing humidity, precipitation, and diurnal range changed the error by 0.002 °C. At these sample sizes parsimony and reproducibility therefore favour the linear model, and single-seed reporting of stochastic learners overstates their accuracy.

Keywords: *Monthly temperature forecasting, Machine learning, Climatological baseline, Forecast verification, Peruvian altiplano, Model parsimony*

Corresponding author: Leonel Coyla-Idme
e-mail ✉ lcoyla@unap.edu.pe
Received: 10 July 2026
Accepted: 19 August 2026

INTRODUCTION

Monthly temperature forecasts carry measurable economic weight in high-altitude agricultural regions. In the Peruvian Altiplano, where Puno lies above 3,800 m a.s.l., planting calendars, irrigation scheduling and frost protection all hinge on anticipating thermal conditions weeks ahead, and a frost outside the expected window can eliminate a season of quinoa or potato production for smallholder farmers with no insurance. The operational requirement follows directly: a forecast earns its place only if it beats what a farmer already infers from decades of seasonal experience.

The regional climate makes that requirement harder to satisfy than it first appears. Temperature in the southern Peruvian Andes couples a pronounced annual cycle with modest interannual variability, and station-based analyses report warming trends whose magnitude shifts with elevation and season (Imfeld *et al*., 2021; Anishchenko *et al*., 2026). Bolivian records across the wider Altiplano show comparable behaviour with strong spatial heterogeneity (Seiler *et al*., 2013), and Andean temperature departed from the global mean in its response to large-scale forcing during the early 2000s (Vuille *et al*., 2015; Wong *et al*., 2025). A model built for one Andean station therefore confronts a signal whose seasonal component is large and easy to reproduce, and whose interannual component is small and genuinely difficult to predict.

Machine learning has been applied widely to this problem, and studies in Latin America report encouraging accuracy. Feed-forward neural networks are the reference non-linear approach in forecasting (Zhang *et al*., 1998), although comparative work on air temperature prediction finds their advantage over simpler regression modest (Chevalier *et al*., 2011; Morales *et al*., 2025; Novak & Dvorak, 2025). Within the region, neural networks have modelled climatic factors in Nicaragua (Tirado Picado, 2024) and long short-term memory networks have supported reservoir early-warning systems in Peru (Marín Vilca & Pineda Torres, 2019). Closest to the study site, an agroclimatic prediction system coupling machine learning with Internet of Things sensors was developed for the agricultural sector of Puno (Yana Puma & Quispe Merma, 2020), and a comparison of twelve models for frost prediction across thirteen Altiplano stations found an ensemble to outperform the best statistical

method by 35% (Almeida & Moreira, 2025; Torres-Cruz *et al*., 2025). Together these studies establish that machine learning can represent climatic relationships in the region, and they consistently report low error metrics.

What they do not establish is whether those low error metrics represent predictive information at all. Monthly mean temperature in a strongly seasonal regime is dominated by the annual cycle, so a model predicting the long-term average of the target calendar month already reproduces most of the variance; a coefficient of determination near 0.90 is thus attainable by a model that has learned nothing beyond seasonality. Reported without a climatological comparison, such a value cannot be interpreted, because the reader has no way to separate genuine skill from the trivial reproduction of the seasonal cycle.

A second practice compounds the first. Splitting a temporal series at random, a convention inherited from cross-sectional machine learning, lets information from future months shape the fitted model and inflates apparent accuracy (Bergmeir & Benítez, 2012). An absent baseline and a shuffled split reinforce each other: together they make an uninformative model appear skilful and leave the deficiency invisible. Establishing whether machine learning adds value here requires controlling both at once.

This study therefore evaluates one-month-ahead forecasts of monthly mean temperature at the Puno Principal Climatological Station under a verification protocol built to close both gaps. Multiple linear regression, random forest, and a multilayer perceptron are compared against monthly climatology and persistence on a chronologically held-out test period, with hyperparameters tuned by blocked time-series cross-validation confined to the training years. Differences in accuracy are tested with the Diebold-Mariano statistic (Diebold & Mariano, 1995) rather than inferred from point estimates, and skill is expressed as the mean square error skill score relative to climatology (Murphy, 1988; Abdullah *et al*., 2025; Salem *et al*., 2025).

Five contributions follow: the improvement over climatology is quantified as a skill score rather than an unreferenced coefficient of determination; the apparent advantage of the neural model is shown to be an artefact of a single random initialisation; the protocol is replicated across eleven independent Altiplano series; learning curves are extended to 2,475 training months, testing the sample-size explanation by measurement rather than extrapolation; and a reproducible quality-control and gap-handling procedure for an incomplete Andean record is documented.

MATERIALS AND METHODS

Study area, data source and replication dataset

The Puno Principal Climatological Station records daily observations on the western shore of Lake Titicaca at approximately 3,820 m a.s.l. The dataset comprises 7,279 daily records spanning 1 January 2003 to 31 December 2024, with mean relative humidity (%), total precipitation (mm), and mean, maximum, and minimum temperature (°C). The workbook also contained a sheet labelled as the Los Uros Ordinary Climatological Station, whose 65,538 cells proved identical to those of the Puno sheet apart from the station label, so it was excluded and every analysis rests on the Puno record alone.

Because conclusions drawn from one station cannot be separated from its peculiarities, the protocol was replicated on an independent source. Daily records for thirteen Altiplano locations, the same used in a previous regional frost-prediction study (Torres-Cruz *et al*., 2025), were retrieved from NASA POWER, which distributes MERRA-2 reanalysis output as open data without registration (NASA POWER, 2025). The retrieval covered 1 January 2000 to 31 December 2024 and returned 9,132 daily records per location with no missing values.

Two checks preceded any modelling. First, pairwise comparison of the thirteen series revealed two pairs, Ayaviri with Cojata and Juli with Yunguyo, whose values were numerically identical because each pair falls inside the same MERRA-2 grid cell of roughly 0.5° × 0.625°. The redundant members were removed, leaving eleven independent series with pairwise daily correlations between 0.690 and 0.963. Second, the reanalysis was compared against the station record at Puno over 6,499 matched days. Reanalysis mean temperature is systematically colder by 2.577 °C (paired $t = -178.3$, $p < 0.001$), a bias consistent with a grid cell whose mean terrain elevation differs from that of the station site, yet it tracks variability closely. The reanalysis is therefore unsuitable as ground truth but well suited to replication, because each series is modelled and evaluated against itself and a constant offset cancels within that comparison.

Quality control and monthly aggregation

Daily records were screened before aggregation. Sentinel values below −99 were converted to missing, relative humidity outside 0–100% was rejected, and the ordering constraint Tmin ≤ Tmean ≤ Tmax was verified; no record violated it. Calendar completeness exposed 757 absent dates concentrated in six blocks rather than scattered at random **(Table 1)**. The distinction matters: dispersed losses are absorbed by averaging, whereas block losses remove entire months and sever the lag structure any forecasting model depends on.

Monthly values were computed as the mean of the available daily observations for temperature and humidity, and as the sum for precipitation. A month was admitted only when it satisfied the completeness rule of the World Meteorological Organization (WMO, 2017), which rejects a month with more than 10 missing daily values or 5 or more consecutive ones. Of 264 candidate months, 236 (89.4%) met the rule, and 26 of the 28 rejections corresponded to entire absent blocks (Bello & Musa, 2026).

Reconstruction of missing daily mean temperature

Daily mean temperature was missing on 10.72% of records while maximum and minimum temperature were missing on only 1.85% and 1.26%, an asymmetry that made reconstruction from the daily extremes worth attempting. Evaluated against the 6,499 days on which the observed mean and both extremes were available, the midpoint estimator $(T_{max} + T_{min})/2$ underestimated the observed mean by 0.687 °C (paired $t = -65.21$, $p < 0.001$). The bias is systematic rather than random, because the recorded mean derives from synoptic-hour observations; substituting the raw midpoint would have injected an offset equal to 42% of the standard deviation of the monthly series.

The midpoint was therefore calibrated by ordinary least squares,

$$T_{mean} = a + b \times (T_{max} + T_{min}) / 2 \quad (1)$$

with coefficients estimated exclusively on training-period days (2003–2016, n = 4,541), giving Tmean = 1.5230 + 0.9241 · (Tmax + Tmin)/2. Ten-fold cross-validation on the fitting sample returned a mean absolute error of 0.691 °C, an R^2 of 0.801, and a residual bias of -2×10^{-5} °C; applied to the held-out years it retained a mean absolute error of 0.609 °C and a residual bias of 0.284 °C, confirming that the correction transfers across time without reintroducing the original offset. The procedure recovered 606 daily values across the record, of which 604 fall inside the retained months and represent 8.4% of their days.

Forecasting task and predictors

The task is a one-month-ahead point forecast: given information observable up to and including month t, predict the mean temperature of month t + 1. Twelve predictors evaluable at time t were constructed. Three lagged temperature terms (t, t−1, t−2) and a three-month rolling mean carry thermal memory; the diurnal range, relative humidity and precipitation of month t carry concurrent atmospheric state. The mean temperature of the target calendar month, computed exclusively from training-period observations, supplies the climatological expectation, and the departure of month t from its own climatological mean supplies the current anomaly. Sine and cosine transforms of the target month index encode the annual cycle without a December-to-January discontinuity, and a linear time index represents the secular trend.

Longer lags were deliberately excluded. A twelve-month lag would have been the most autocorrelated predictor, but requiring twelve consecutive valid months cut the usable sample from 186 to 145 observations, because each absent block spans several months; the climatological predictor conveys the same seasonal information at no such cost. After removing rows with any missing predictor, 186 supervised observations remained, covering April 2003 to November 2024.

Models, reference forecasts and evaluation

Three regression models were compared. Multiple linear regression supplies the parametric benchmark and has no free hyperparameters; random forest aggregates decorrelated regression trees and captures non-additive structure (Breiman, 2001); and the multilayer perceptron is a feed-forward neural network with a single hidden layer trained by backpropagation (Rumelhart *et al*., 1986), the standard non-linear forecasting model for three decades (Zhang *et al*., 1998). Predictors were standardised inside the modelling pipeline for the linear and neural models, so that scaling statistics were estimated on training folds alone; all models were implemented in scikit-learn (Pedregosa *et al*., 2011; Sa-Couto *et al*., 2025). Two reference forecasts define the standard the models must clear: climatology predicts the mean of the target calendar month computed from training-period data only, and persistence predicts the value observed in month t. Without these references a reader cannot judge whether any stated accuracy exceeds what the seasonal cycle already delivers (Jolliffe & Stephenson, 2011).

The 186 observations were divided chronologically: target months up to December 2016 formed the training set (n = 137) and those from January 2017 onward the test set (n = 49). Hyperparameters were selected by grid search under blocked time-series cross-validation with five folds, which trains on past folds and validates on the immediately following fold, preserving temporal order (Bergmeir & Benítez, 2012). The search evaluated 54 configurations for random forest and 32 for the perceptron, scored by cross-validated mean absolute error. The test set was consulted once, after hyperparameter selection had closed.

Four accuracy measures were computed on the test set: mean absolute error and root mean square error quantify typical and outlier-sensitive error magnitude, the coefficient of determination quantifies explained variance, and mean bias error exposes systematic over- or under-forecasting:

$$MAE = (1/n) \Sigma |\hat{y}_i - y_i| \quad (2)$$

$$RMSE = \sqrt{[(1/n) \Sigma (\hat{y}_i - y_i)^2]} \quad (3)$$

$$MBE = (1/n) \Sigma (\hat{y}_i - y_i) \quad (4)$$

Skill was expressed as the mean square error skill score (Murphy, 1988),

$$MSSS = 1 - MSE_{model} / MSE_{climatology} \quad (5)$$

which equals zero for a forecast equivalent to climatology, approaches one for a perfect forecast, and turns negative for one worse than climatology. Because point estimates may reflect sampling variation, pairwise differences were tested with the Diebold-Mariano statistic on squared-error loss differentials (Diebold & Mariano, 1995), and 95% confidence intervals came from 2,000 bootstrap resamples of the test set (Mulyani *et al*., 2026).

Complementary analyses

Five further analyses test whether the headline comparison holds under scrutiny. An ablation refitted every model on five nested predictor sets, isolating the contribution of seasonal information, thermal memory, and atmospheric covariates. A seasonal breakdown partitioned the test set into wet (December–March), dry (May–August), and transition (April, September–November) months. Residual diagnostics assessed normality with the Shapiro-Wilk test and serial independence with the Ljung-Box statistic at five lags, since residuals that retain structure indicate predictable signal left unused. Learning curves refitted each model on training subsets of 30 to 137 months, using up to six contiguous windows per size and five initialisations per window.

The replication applied the same pipeline to each of the eleven independent series, with training months up to December 2018 and test months from January 2019 onward, yielding 2,475 pooled training and 792 pooled test observations. Pooled learning curves were measured at sizes from 100 to 2,475 months, extending the range examined at Puno eighteenfold so that the sample-size explanation could be tested against data instead of extrapolated.

The fifth analysis addresses a reporting practice that small samples make hazardous. Random forest and the perceptron

are stochastic, whereas ordinary least squares has a closed-form solution, so reporting a single seed conflates model performance with initialisation luck. Both stochastic models were refitted under 50 seeds, and the resulting distributions of test error are reported alongside the single-seed values.

RESULTS AND DISCUSSION

Characteristics of the temperature record

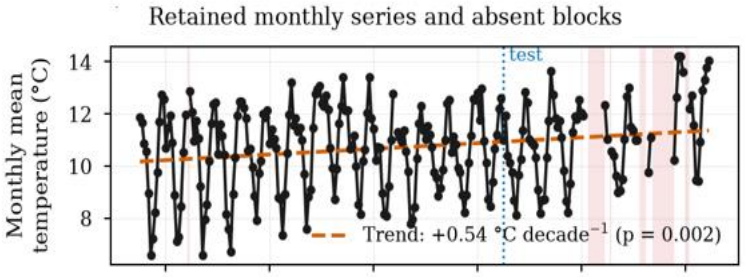
The retained monthly series has a mean of 10.72 °C and a standard deviation of 1.62 °C, ranging from 6.59 °C to 14.21 °C **(Table 1)**. The record warms significantly at +0.54 ± 0.17 °C decade^{-1} (p = 0.0017), which exceeds the long-term rates reported for the tropical Andes across the twentieth century (Vuille *et al*., 2003) and agrees in sign with station-based trends for the southern Peruvian Andes (Imfeld *et al*., 2021). **Figures 1a, 1b** present the series alongside the monthly count of missing days, and show the losses clustering after 2020 rather than spreading evenly across the record.

The seasonal cycle dominates the variance, and this single fact governs the interpretation of every result that follows. Monthly climatological means range from 8.17 °C in July to 12.85 °C in November, an amplitude of 4.68 °C, whereas the within-month standard deviation never exceeds 0.94 °C and averages 0.71 °C **(Figure 1c)**. The autocorrelation function confirms the structure, with coefficients of 0.763 at lag 1, 0.878 at lag 12, and 0.847 at lag 24, and a pronounced −0.521 at lag 6 **(Figure 1d)**. The consequence is unavoidable: a forecast reproducing only the seasonal cycle will already look accurate, which is why the climatological reference is indispensable rather than optional.

Table 1. Characteristics of the daily record, the absent periods, and the retained monthly series at the Puno Principal Climatological Station, 2003–2024.

Characteristic	Value
Record span	2003-01-01 to 2024-12-31
Daily records available	7,279
Calendar days in span	8,036
Absent calendar days	757 (9.4%) in 6 blocks
Longest absent block	October 2022 – July 2023 (304 days)
Missing daily mean temperature	10.72%
Missing daily maximum / minimum temperature	1.85% / 1.26%
Missing daily relative humidity / precipitation	10.36% / 4.60%
Physically inconsistent temperature records	0
Daily values recovered by calibration	606 in the record, 8.4% of the days in retained months
Candidate months	264
Months meeting WMO completeness	236 (89.4%)
Mean of the retained monthly series	10.72 °C
Standard deviation of the retained monthly series	1.62 °C
Range of the retained monthly series	6.59 to 14.21 °C
Coldest / warmest climatological month	July, 8.17 °C / November, 12.85 °C
Seasonal amplitude of the climatology	4.68 °C
Mean within-month standard deviation	0.71 °C
Linear trend	+0.54 ± 0.17 °C decade^{-1} (p = 0.0017)

Note. The World Meteorological Organization completeness rule rejects a month with more than 10 missing daily values or 5 or more consecutive ones. The climatological statistics describe the 236 retained months.



a)

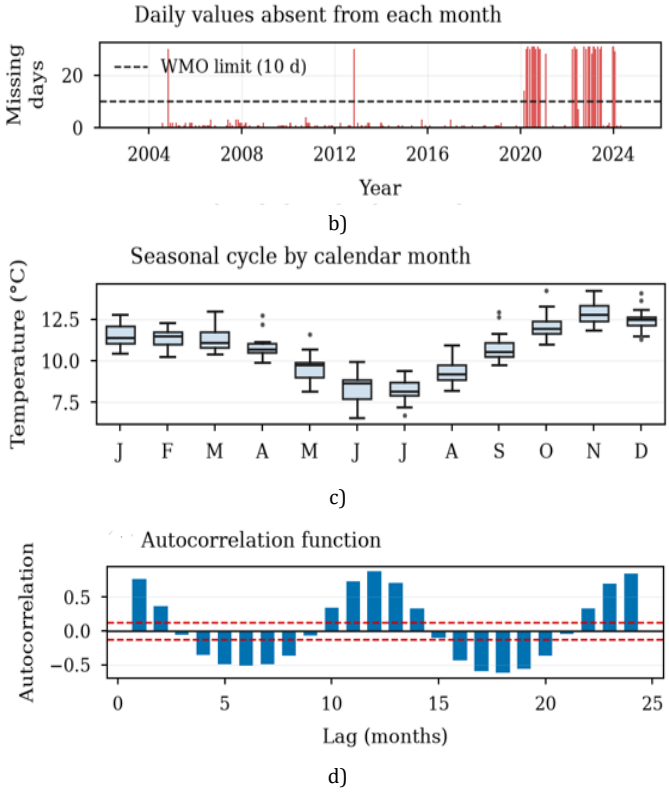


Figure 1. The Puno record and its temporal structure: (a) retained monthly mean temperature with the fitted linear trend, the start of the test period, and shaded periods absent from the record; (b) days missing per month against the World Meteorological Organization completeness limit; (c) seasonal cycle by calendar month, of amplitude 4.68 °C against a mean within-month standard deviation of 0.71 °C; and (d) autocorrelation function up to 24 months, with 95% significance bounds.

Forecast accuracy against the references

All three machine learning models outperformed climatology, though by a moderate rather than a transformative margin **(Table 2, Figure 2a)**. Against the 0.597 °C mean absolute error of climatology, multiple linear regression reached 0.421 °C, the multilayer perceptron 0.415 °C, and random forest 0.476 °C, giving mean square error skill scores of 0.505, 0.489, and 0.310. The best models therefore remove approximately half of the mean square error climatology leaves behind: a real improvement with a clear ceiling.

Persistence performed markedly worse than climatology, at 0.866 °C and a skill score of −0.958, a degradation the Diebold-Mariano test confirms as significant (t = 3.057, p = 0.004). The result is physically coherent: where the seasonal cycle spans 4.68 °C and month-to-month change routinely exceeds the interannual variability of any individual month, carrying the previous month forward is worse than ignoring it. Persistence alone therefore cannot serve as the reference in seasonal climates, and climatology must anchor the comparison.

Statistical testing tempers the ranking the point estimates suggest. Against climatology, linear regression (t = −2.255, p = 0.029) and random forest (t = −2.222, p = 0.031) improved significantly at the 5% level, whereas the perceptron fell marginally short (t = −1.998, p = 0.051), and the bootstrap intervals overlap across the three models **(Table 2)**. With 49 test months the evidence does not support fine distinctions, and reporting them as a strict ordering would overstate what the data can bear.

Table 2. Forecast accuracy, skill and generalisation on the test period (January 2017 – November 2024, n = 49 months).

Measure	Persist.	Clim.	MLR	RF	MLP
MAE (°C)	0.866	0.597	0.421	0.476	0.415
MAE 95% CI	[0.701, 1.041]	[0.478, 0.728]	[0.331, 0.512]	[0.367, 0.586]	[0.323, 0.518]
RMSE (°C)	1.059	0.757	0.532	0.629	0.541
R^2	0.520	0.755	0.879	0.831	0.875
MBE (°C)	-0.088	-0.271	+0.107	-0.160	+0.086
MSSS	-0.958	+0.000	+0.505	+0.310	+0.489
DM p vs climatology	0.004	—	0.029	0.031	0.051

Training MAE (°C)	—	—	0.415	0.226	0.304
Test/train ratio	—	—	1.01	2.11	1.37
Cross-validated MAE (°C)	—	—	0.600	0.498	0.681
Hyperparameters	—	—	none	depth 6, 300 trees	16 units, α = 1.0

Note. MLR = multiple linear regression; RF = random forest; MLP = multilayer perceptron. MAE = mean absolute error; RMSE = root mean square error; MBE = mean bias error; MSSS = mean square error skill score relative to climatology; DM = Diebold-Mariano test against climatology. Confidence intervals use 2,000 bootstrap resamples; the test/train ratio is test MAE divided by training MAE; cross-validated MAE was obtained under blocked time-series cross-validation with five folds on the training period.

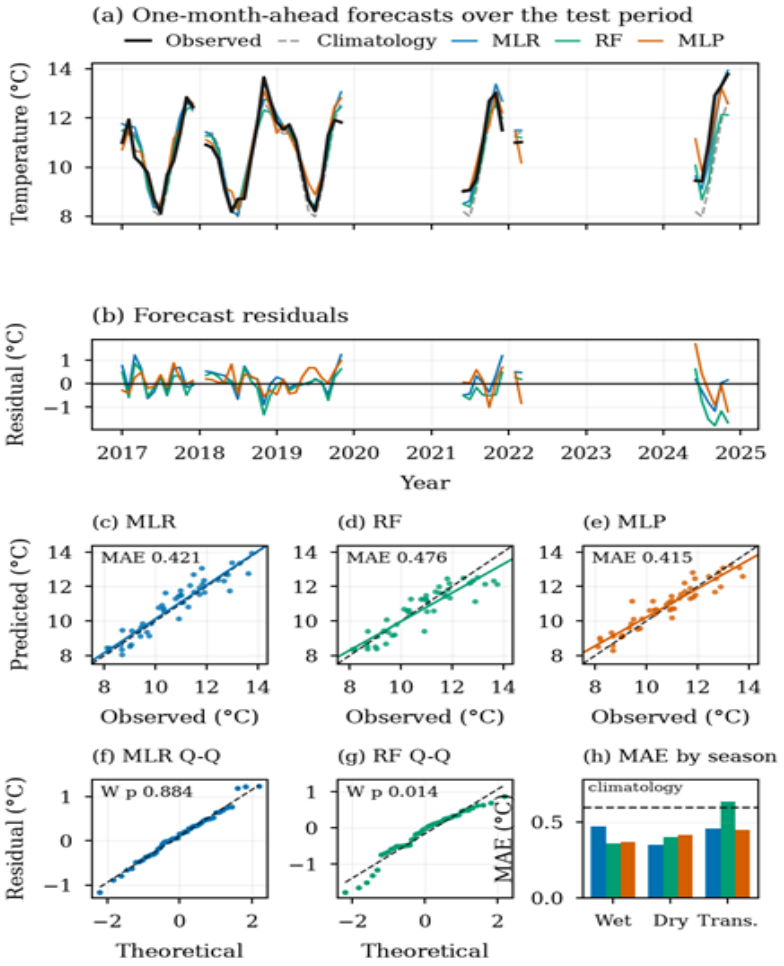


Figure 2. Forecast performance on the test period: (a) observed values against climatology and the three models, with line breaks at months rejected by the completeness rule; (b) forecast residuals; (c–e) observed against predicted values, where the dashed line is the identity and the solid line the least-squares fit; (f, g) normal quantile-quantile plots of the residuals, where departures from the dashed line indicate non-normality; and (h) mean absolute error by season against the climatological reference (dashed). MLR = multiple linear regression; RF = random forest; MLP = multilayer perceptron.

Model complexity brings no measurable benefit

The central result is that the simplest model performs at least as well as the most complex one. Under the single fit reported in **Table 2**, the perceptron achieved the lowest mean absolute error, 0.415 °C against 0.421 °C for linear regression, yet the Diebold-Mariano test detects no significant difference (t = 0.104, p = 0.918), and linear regression returns the lower root mean square error (0.532 °C against 0.541 °C), the higher coefficient of determination, and the higher skill score. The 0.006 °C margin lies far below the resolution that matters operationally, and the two scatters are visually indistinguishable **(Figures 2c, 2e)**.

Generalisation behaviour explains why complexity earns nothing here. Random forest reached a training error of 0.226 °C against a test error of 0.476 °C, a ratio of 2.11 indicating substantial overfitting even after tuning had restricted tree depth to six **(Table 2)**, while the perceptron shows a milder ratio of 1.37 and linear regression essentially none, at 1.01. With 137 training observations and 12 predictors, flexible learners have ample capacity to fit sampling noise, and the strong regularisation selected by cross-validation (α = 1.0) reflects that

pressure directly.

The neural advantage does not survive re-seeding

Refitting the stochastic models under 50 random initialisations dissolves the perceptron's apparent lead **(Table 3)**. Across seeds it averages 0.495 °C with a standard deviation of 0.048 °C and a range from 0.407 °C to 0.647 °C, so the 0.415 °C reported for a single fit sits at the 6th percentile of its own sampling distribution. Only 4 of 50 initialisations beat the deterministic 0.421 °C of linear regression, so 92% of the time an analyst repeating this experiment with a different seed would conclude that linear regression wins. Random forest behaves differently, with a standard deviation of 0.005 °C and no initialisation reaching the linear model, confirming that seed sensitivity is specific to the neural architecture rather than general to stochastic learners.

Averaging the 50 perceptrons into an ensemble does not recover the advantage either: it attains 0.4209 °C against 0.4208 °C for linear regression, with a higher root mean square error (0.564 °C against 0.532 °C). Fifty neural networks combined therefore reproduce, at fifty times the cost and with no interpretable coefficients, what a twelve-parameter closed-form regression delivers in a single fit.

Table 3. Residual diagnostics, sensitivity to random initialisation, predictor ablation and seasonal breakdown on the test period.

Diagnostic	MLR	RF	MLP
Shapiro-Wilk W	0.988	0.940	0.981
W p	0.884	0.014	0.602
Lag-1 autocorrelation	+0.131	+0.427	+0.019
Ljung-Box Q	9.88	16.61	13.60
Q p	0.079	0.005	0.018
Mean MAE over seeds (°C)	—	0.476	0.495
SD over seeds (°C)	—	0.005	0.048
Range over seeds (°C)	—	0.464–0.486	0.407–0.647
Seeds beating MLR	—	0/50	4/50
MAE, seasonal only (k = 3)	0.599	0.597	0.593
MAE, memory only (k = 4)	0.704	0.537	0.610
MAE, seasonal + memory (k = 8)	0.423	0.476	0.476
MAE, without covariates (k = 9)	0.423	0.468	0.506
MAE, all predictors (k = 12)	0.421	0.476	0.415
MAE, wet season (n = 13)	0.473	0.361	0.369
MAE, dry season (n = 18)	0.348	0.401	0.418
MAE, transition (n = 18)	0.456	0.635	0.446

Note. Model abbreviations as in **Table 2**. Shapiro-Wilk tests normality and Ljung-Box Q, computed at five lags, tests serial independence; p-values above 0.05 indicate residuals consistent with normality and white noise respectively. Seed statistics summarise 50 random initialisations; ordinary least squares is deterministic. In the ablation, k = number of predictors, seasonal = target-month climatology and the two cyclical encodings, memory = lagged temperature terms and the three-month rolling mean, and covariates = relative humidity, precipitation and diurnal temperature range. Wet = December–March, dry = May–August, transition = April and September–November. Climatology attains 0.597 °C overall.

Only the linear model satisfies its own assumptions

Residual diagnostics separate the models more sharply than the error metrics do **(Table 3, Figures 2f, 2g)**. Multiple linear regression is the only model whose residuals are simultaneously consistent with normality (Shapiro-Wilk W = 0.988, p = 0.884) and free of serial correlation (Ljung-Box Q = 9.88, p = 0.079). Random forest fails both tests, departing from normality (W = 0.940, p = 0.014) and retaining significant serial dependence (Q = 16.61, p = 0.005) with a lag-one autocorrelation of 0.427. The perceptron is intermediate, with normal residuals (p = 0.602) that nonetheless retain detectable structure (Q = 13.60, p = 0.018).

This diagnostic gap carries more weight than the difference in mean absolute error. A residual autocorrelation of 0.427 means random forest leaves a predictable component of the signal unused, so its errors are systematically patterned, whereas linear regression produces the white noise its inferential framework assumes. Selecting the linear model rests on model adequacy, not on a narrow win in a summary score.

Where the predictive information actually comes from

An ablation over nested predictor sets locates the source of skill precisely **(Table 3)**. Seasonal information alone, comprising target-month climatology and the two cyclical encodings, yields 0.599 °C for linear regression, indistinguishable from the 0.597 °C of climatology itself; three predictors capturing the annual cycle add nothing to simply knowing the calendar month. Thermal memory alone performs better for random forest (0.537 °C) but worse for linear regression (0.704 °C). Only their combination produces the improvement: eight predictors joining seasonality with memory bring linear regression to 0.423 °C, essentially the full-model result.

Atmospheric covariates contribute nothing measurable. Removing relative humidity, precipitation, and diurnal

temperature range changes the linear model from 0.4208 °C to 0.4226 °C **(Table 3)**, and permutation importance likewise ranks humidity and precipitation last for both non-parametric models. The operational implication is favourable: a forecasting service for this station would lose nothing by relying on temperature and calendar information alone, which substantially reduces the data-collection burden.

No model wins in every season

Partitioning the test period by season shows that no single ranking holds throughout the year **(Table 3, Figure 2h)**. During the wet season the non-parametric models clearly lead, with random forest at 0.361 °C and the perceptron at 0.369 °C, compared to 0.473 °C for linear regression. The ordering reverses in the dry season, where linear regression yields 0.348 °C, compared with 0.401 °C and 0.418 °C. In transition months, which carry by far the highest observed variability, with a standard deviation of 1.215 °C, compared with roughly 0.69 °C in the other two regimes, random forest deteriorates sharply to 0.635 °C, above the climatological reference, while linear regression and the perceptron hold near 0.45 °C.

The pattern is coherent with the generalisation evidence: random forest performs best when conditions resemble those it memorised and worst when variability is highest, while linear regression degrades more gracefully in the regime that matters most for frost risk. For an operational service, robustness across seasons should outweigh the lowest average error.

Replication across eleven series confirms the ordering

Applying the identical protocol to eleven independent Altiplano series reproduces the Puno result **(Table 4, Figure 3a)**. Over 792 pooled test months, climatology attains 0.640 °C, multiple linear regression 0.438 °C, the perceptron 0.453 °C, and random forest 0.474 °C. Linear regression is most accurate at 9 of the 11 sites; the perceptron leads only at Macusani (0.353 °C against 0.368 °C) and Mazocruz (0.469 °C against 0.494 °C), and random forest leads nowhere. Every model beats climatology at every site, so the improvement replicates cleanly. Site-to-site variation follows the difficulty of the local climate rather than the algorithm: errors are lowest at Macusani and Ayaviri and highest at Desaguadero, Mazocruz, and Crucero Alto, where climatology itself exceeds 0.72 °C, yet the ranking of models is stable across that range.

Pooling the eleven series permits the sample-size explanation to be tested against data rather than extrapolated **(Figure 3b)**. Linear regression falls to 0.424 °C by 700 pooled training months and then flattens, ending at 0.438 °C with 2,475 months, while random forest declines to 0.477 °C and the perceptron to 0.467 °C; both curves flatten above the linear model without crossing it at any measured size. Eighteen times the training data available at Puno therefore leaves the ordering unchanged, converting a previously untestable limitation into a result.

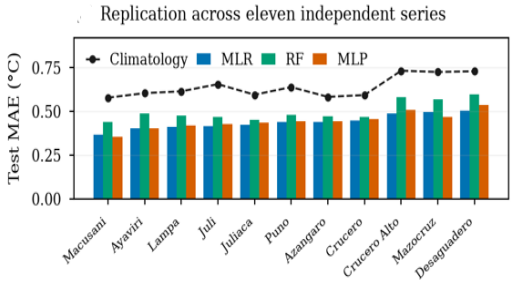
Seed sensitivity persists undiminished at the larger sample. Across 20 initialisations on the pooled training set the perceptron averages 0.510 °C (SD 0.045 °C) and only 2 of 20 beat the deterministic 0.438 °C of linear regression, while random forest remains stable at 0.002 °C with no initialisation reaching the linear model. Because instability does not diminish as the sample grows eighteenfold, it reflects the optimisation landscape of the architecture rather than a shortage of data. Studies reporting large gains from neural architectures on climate data typically work with gridded products offering orders of magnitude more samples still (Rasp *et al*., 2020), and whether deep learning can outperform numerical weather prediction remains contested even in that setting (Schultz *et al*., 2021; Basaad *et al*., 2026; Thach, 2026).

The reanalysis comparison illustrates the study's central point in miniature **(Figures 3c, 3d)**. Reanalysis and station monthly temperatures correlate at r = 0.955, a near-perfect agreement until the shared seasonal cycle is removed, after which the anomaly correlation falls to 0.811. The inflation that makes an uninformative forecast appear skilful also makes two datasets appear more concordant than they are.

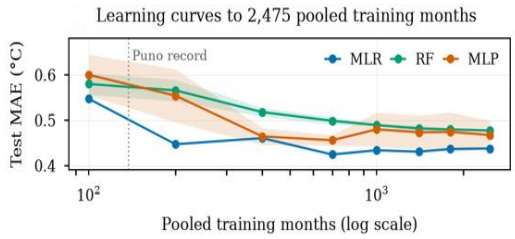
Table 4. Replication of the forecasting protocol across eleven independent Altiplano series, test period January 2019 – December 2024.

Series	n test	Clim.	MLR	RF	MLP
Macusani	72	0.578	0.368	0.439	0.353
Ayaviri	72	0.604	0.402	0.489	0.403
Lampa	72	0.614	0.412	0.476	0.419
Juli	72	0.655	0.416	0.469	0.428
Juliaca	72	0.595	0.423	0.452	0.436
Puno	72	0.638	0.438	0.481	0.445
Azangaro	72	0.583	0.441	0.473	0.444
Crucero	72	0.593	0.448	0.469	0.454
Crucero Alto	72	0.732	0.490	0.583	0.509
Mazocruz	72	0.725	0.494	0.568	0.469
Desaguadero	72	0.729	0.503	0.595	0.537
Pooled	792	0.640	0.438	0.474	0.453

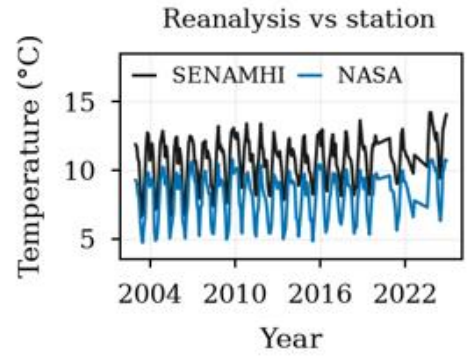
Note. Values are test mean absolute error in °C; model abbreviations as in **Table 2**. Sites are ordered by the error of multiple linear regression, and Pooled combines all eleven series. Ayaviri and Juli each stand for a pair of locations sharing one MERRA-2 grid cell.



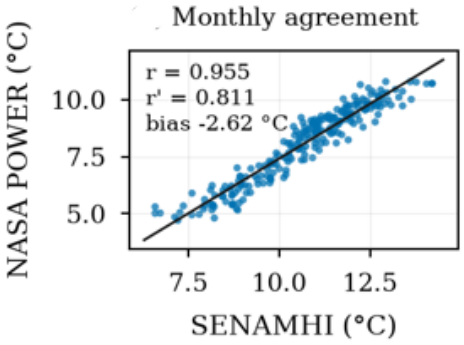
a)



b)



c)



d)

Figure 3. Replication and scaling across eleven independent Altiplano series: (a) test mean absolute error by site, ordered by the error of multiple linear regression, with the climatological reference as a dashed line; (b) pooled learning curves on a logarithmic axis up to 2,475 training months, with shaded bands showing the standard deviation across initialisations and a dotted line marking the 137 months available at Puno; (c) monthly mean temperature from SENAMHI and from NASA POWER at Puno, showing the systematic cold offset; and (d) their monthly agreement, where r is the raw correlation and r' the correlation after removing each series' own monthly climatology. Model abbreviations as in **Figure 2**.

Limitations

Four limitations bound these results. First, only one of the twelve series consists of direct station observations; the replication rests on reanalysis output colder than the station record by 2.577 °C, so it establishes that the ordering of models transfers across sites without establishing the absolute accuracy attainable there. Second, the Puno test set comprises 49 months, sufficient to separate models from climatology but not to resolve differences among them; the pooled replication mitigates but does not eliminate this constraint, since its 792 test months are not fully independent given pairwise correlations up to 0.963 between neighbouring series. Third, 8.4% of the daily values entering the retained months were reconstructed by calibration, which introduces uncertainty absent from a complete record. Fourth, the horizon is one month, and the skill reported here should not be extrapolated to longer horizons, where the informative content of recent thermal state decays.

CONCLUSION

Machine learning improves one-month-ahead monthly temperature forecasts across the Peruvian Altiplano, but the improvement is moderate and does not come from model complexity. At the Puno Principal Climatological Station the models cut mean absolute error from the 0.597 °C achieved by climatology to between 0.415 °C and 0.476 °C, corresponding to skill scores of 0.31 to 0.51, while persistence performs significantly worse than climatology and should not serve as a reference in this seasonal regime. Multiple linear regression is preferable on four independent grounds rather than on the basis of a margin of error. Across 50 initialisations the perceptron's apparent lead proves seed-dependent, with 92% falling short of the deterministic linear model and an ensemble of all fifty merely matching it. Residual diagnostics identify linear regression as the only model whose errors are both normally distributed and serially independent, whereas random forest leaves a lag-one autocorrelation of 0.427 unexploited. Replication across 11 independent Altiplano series and 792 test months places linear regression first at 9 of 11 sites. And learning curves extended to 2,475 pooled training months, eighteen times the record available at Puno, show the machine learning curves flattening above the linear model rather than crossing it, settling by measurement a question that extrapolation could not. Because predictive information arises almost entirely from combining target-month climatology with recent thermal state, while humidity, precipitation, and diurnal range together contribute 0.002 °C, an operational service for this region can be built on an interpretable linear model with a small predictor set. The seed analysis carries a caution beyond this application: reporting a stochastic learner from a single initialisation can overstate its accuracy by more than the difference a study sets out to measure. Three directions follow: obtaining observational records for the remaining Altiplano sites, evaluating longer forecast horizons, and testing whether the seed instability documented here persists in deeper architectures.

ACKNOWLEDGMENTS: The authors thank the Instituto de Investigación and the Vicerrectorado de Investigación of the Universidad Nacional del Altiplano de Puno for support through the Research Seedbeds programme, and the Servicio Nacional de Meteorología e Hidrología del Perú for the climatological records.

CONFLICT OF INTEREST: None

FINANCIAL SUPPORT: This work was carried out within the Research programme in VRI of the Universidad Nacional del Altiplano de Puno.

ETHICS STATEMENT: This study analysed meteorological records only and involved no human or animal participants; ethical approval was not required.

REFERENCES

Abdullah, N. A., Zulkifli, M. I., & Mohamed, A. S. (2025). Refinement of the 8th AJCC staging system for medullary

thyroid cancer: Integrating tumor size and lymph node characteristics with SEER and multicenter validation. *Archives of International Journal of Cancer and Allied Sciences, 5*(2), 34–43. doi:10.51847/R1sIaONOms

Almeida, P. H., & Moreira, R. B. (2025). Adherence to face mask guidelines among Polish healthcare professionals during the COVID-19 pandemic: Are we doing enough? *Journal of Medical Sciences and Interdisciplinary Research, 5*(2), 133–142. doi:10.51847/RGZMFJ0OZl

Anishchenko, M., Pozdniakova-Kyrbiatieva, E., Mosaiev, Y., Pozdniakova, O., & Zhuzha, L. (2026). Use of art therapy, garden therapy, and swimming in the psychological rehabilitation of children in Ukraine. *Journal of Advanced Pharmacy Education and Research, 16*(1), 63–68. doi:10.51847/DeRqLGWsgv

Basaad, S., Budiman, J. A., & Octarina, O. (2026). Shear bond strength of reused orthodontic brackets after multiple bonding cycles: An in vitro study. *Asian Journal of Periodontics and Orthodontics, 6*(1), 25–31. doi:10.51847/LocQtMgOTb

Bello, A., & Musa, Z. (2026). Ethical issues in palliative care for patients with chronic diseases: A systematic review of nursing perspectives. *Journal of Integrative Nursing and Palliative Care, 7*(1), 38–50. doi:10.51847/jbe1AvOMQ3

Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences, 191*, 192–213. doi:10.1016/j.ins.2011.12.028

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. doi:10.1023/A:1010933404324

Chevalier, R. F., Hoogenboom, G., McClendon, R. W., & Paz, J. A. (2011). Support vector regression with reduced training sets for air temperature prediction: A comparison with artificial neural networks. *Neural Computing and Applications, 20*(1), 151–159. doi:10.1007/s00521-010-0363-y

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics, 13*(3), 253–263. doi:10.1080/07350015.1995.10524599

Imfeld, N., Sedlmeier, K., Gubler, S., Correa Marrou, K., Davila, C. P., Huerta, A., Lavado-Casimiro, W., Rohrer, M., Scherrer, S. C., & Schwierz, C. (2021). A combined view on precipitation and temperature climatology and trends in the southern Andes of Peru. *International Journal of Climatology, 41*(1), 679–698. doi:10.1002/joc.6645

Jolliffe, I. T., & Stephenson, D. B. (Eds.). (2011). *Forecast verification: A practitioner's guide in atmospheric science* (2nd ed.). Wiley-Blackwell. doi:10.1002/9781119960003

Marín Vilca, D. G., & Pineda Torres, I. A. (2019). *Modelo predictivo Machine Learning aplicado a análisis de datos hidrometeorológicos para un SAT en represas* [Undergraduate thesis]. Universidad Nacional del Altiplano de Puno.

Morales, D. A., Riquelme, V. S., & Fuentes, T. J. (2025). Factors influencing scholarly output of pharmacy practice chairs: A bibliometric study. *Annals of Pharmacy Education, Safety, and Public Health Advocacy, 5*(1), 173–180. doi:10.51847/6hPPyFnRn5

Mulyani, R., Suharjono, S., Suprapti, B., Hermansyah, A., & Purnama, S. (2026). Diabetes distress and associated factors in type 2 diabetes at an Indonesian hospital: A cross-sectional study. *Journal of Advanced Pharmacy Education and Research, 16*(2), 1–9. doi:10.51847/7y6OKqnuGm

Murphy, A. H. (1988). Skill scores based on the mean square error and their relationships to the correlation coefficient. *Monthly Weather Review, 116*(12), 2417–2424. doi:10.1175/1520-0493(1988)116<2417:SSBOTM>2.0.CO;2

NASA POWER. (2025). *Prediction of Worldwide Energy Resources: Daily point data from MERRA-2* [Data set]. NASA Langley Research Center. https://power.larc.nasa.gov

Novak, T. J., & Dvorak, P. M. (2025). A spatiotemporal neural network framework for EEG-based emotion recognition in depression assessment. *Journal of Medical Sciences and Interdisciplinary Research, 5*(2), 24–38. doi:10.51847/A2pBOYHJW1

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830. https://www.jmlr.org/papers/v12/pedregosa11a.html

Rasp, S., Dueben, P. D., Scher, S., Weyn, J. A., Mouatadid, S., & Thuerey, N. (2020). WeatherBench: A benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems, 12*(11), e2020MS002203. doi:10.1029/2020MS002203

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature, 323*(6088), 533–536. doi:10.1038/323533a0

Sá-Couto, C., Sá-Couto, P., Nicolau, A., Lazarovici, M., Ericsson, C. D., Vieira-Marques, P., & Bispo, I. (2025). Managing safety in teaching clinical skills: Investigating framing errors in cardiopulmonary resuscitation training through a multi-arm randomized controlled equivalence study. *Bulletin of Pioneering Researches of Medical and Clinical Science, 4*(2), 18–29. doi:10.51847/nybbkEFwd7

Salem, H. M., Watanabe, S., & Chang, A. H. (2025). Ethical concerns in managing anorexia nervosa: A content analysis of ethics consultation records. *Asian Journal of Ethics in Health and Medicine, 5*, 25–35. doi:10.51847/oHEI6FgL3V

Schultz, M. G., Betancourt, C., Gong, B., Kleinert, F., Langguth, M., Leufen, L. H., Mozaffari, A., & Stadtler, S. (2021). Can deep learning beat numerical weather prediction? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 379*(2194), 20200097. doi:10.1098/rsta.2020.0097

Seiler, C., Hutjes, R. W. A., & Kabat, P. (2013). Climate variability and trends in Bolivia. *Journal of Applied Meteorology and Climatology, 52*(1), 130–146. doi:10.1175/JAMC-D-12-0105.1

Thach, N. H. (2026). From omnichannel journey design to recommendation intention: Psychological needs and trust in dental care. *Annals of Dental Specialty, 14*(1), 140–150. doi:10.51847/A7b1pICU3f

Tirado Picado, V. R. (2024). Redes neuronales artificiales como modelo de predicción de los factores climáticos en Nicaragua en el periodo 2021–2022. *Ciencia Latina Revista Científica Multidisciplinar, 8*(1), 1458–1474. doi:10.37811/cl_rcm.v8i1.9541

Torres-Cruz, F., Yana-Yucra, D. M., & Vilca-Solorzano, R. A. (2025). Comparative analysis of statistical, machine

learning, and deep learning approaches for frost prediction in the Peruvian Altiplano. *International Journal of Advanced Computer Science and Applications, 16*(9). doi:10.14569/IJACSA.2025.0160992

Vuille, M., Bradley, R. S., Werner, M., & Keimig, F. (2003). 20th century climate change in the tropical Andes: Observations and model results. *Climatic Change, 59*(1–2), 75–99. doi:10.1023/A:1024406427519

Vuille, M., Franquist, E., Garreaud, R., Lavado Casimiro, W. S., & Cáceres, B. (2015). Impact of the global warming hiatus on Andean temperature. *Journal of Geophysical Research: Atmospheres, 120*(9), 3745–3757. doi:10.1002/2015JD023126

Wong, Y.-F., Lin, S.-J., Cheng, H.-C., Hsieh, T.-H., Hsiue, T.-R., Chung, H.-S., Tsai, M.-Y., & Wang, M.-R. (2025). Understanding the impact of medical humanities on internship training and performance. *Annals of Pharmacy Education, Safety, and Public Health Advocacy, 5*(1), 12–21. doi:10.51847/Z1fogzPksy

World Meteorological Organization. (2017). *WMO guidelines on the calculation of climate normals (WMO-No. 1203).* https://library.wmo.int/idurl/4/55797

Yana Puma, J. W., & Quispe Merma, M. D. (2020). *Modelo predictivo agroclimático para detectar cambios climatológicos con machine learning e IoT para el sector agrario de la Región Puno* [Undergraduate thesis]. Universidad Peruana Unión.

Zhang, G., Patuwo, B. E., & Hu, M. Y. (1998). Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting, 14*(1), 35–62. doi:10.1016/S0169-2070(97)00044-7