Extending the PISCO gridded dew-point record for Peru (2017–2020) from the continued minimum-temperature record, using per-climate-zone seasonal model selection

Piero Palacios
International Potato Center (CIP)
piero.palacios@cgiar.org

Henry Juarez
International Potato Center (CIP)
h.juarez@cgiar.org

Leroy Mwanzia
International Potato Center (CIP)
l.mwanzia@cgiar.org

2026-11-17

Table of contents

Abstract . . . 2
1 Introduction . . . 3
2 Data . . . 3
3 Methods . . . 4
3.1 Anomaly formulation . . . 4
3.2 Local regression with a day-of-year kernel . . . 5
3.3 Selection of predictors and bandwidth . . . 5
3.4 Autoregressive reconstruction . . . 6
3.5 Validation design . . . 6
4 Results . . . 7
4.1 Selected models . . . 7
4.2 Accuracy on withheld years . . . 8
4.3 The reconstructed field . . . 12
4.4 Continuity at the record boundary . . . 14
4.5 Extreme values . . . 15
5 Discussion . . . 17
5.1 Estimating humidity from temperature . . . 17
5.2 Interpretation of the selected bandwidths . . . 18
5.3 Temporal extrapolation and its validation . . . 18
5.4 Stability of long autoregressive reconstruction . . . 18
5.5 Origin of the extreme values . . . 19
6 Limitations . . . 20
7 Supplementary material . . . 21
7.1 Predictor-version sensitivity . . . 21
7.2 Computational characterisation . . . 22
8 Data and code availability . . . 23
9 AI-use disclosure . . . 23
10 SME sign-off . . . 25
References . . . 25

Bibliography 25

> ! Important
>
> **DRAFT: not for circulation.** All quantitative results below are AI-assisted analyses that **require subject-matter-expert (SME) validation before use** (CIP Acceptable Use policy; see Section 9). The statements awaiting expert sign-off are listed in Table 15.

Abstract

The PISCO gridded climate product for Peru provides daily dew-point temperature (T_d) through 2016, while its minimum-temperature ($T_{\min}$) record extends to 2020, leaving a four-year gap in a variable required for reference-evapotranspiration, disease-risk and crop models. We reconstruct T_d for 2017–2020 by learning the $T_{\min} \to T_d$ mapping on the 1981–2016 overlap and applying it forward at every grid cell. The model is a per-(cell, day-of-year) weighted local linear regression on climate anomalies; the predictor subset and the day-of-year smoothing half-window h are selected jointly for each combination of SENAMHI climate zone and day of year by leave-one-year-out cross-validation. Because no observed T_d exists in the gap, reconstruction is autoregressive: each predicted day supplies the T_d lags of subsequent days.

Because the task is extrapolation in time rather than interpolation in space, we validate by withholding the four most recent years (train 1981–2012, reconstruct 2013–2016) and scoring against observed T_d under the same autoregressive conditions the real gap imposes. Within the domain in which the climate classification is defined and the source product is station-constrained, which is 48.9% of the released cells, the model attains $\mathrm{RMSE} = 0.946$ °C at a skill score of 0.251 and a bias of $+0.010$ °C; pooled over the whole released field, half of which lies beyond that domain, $\mathrm{RMSE} = 1.228$ °C at a skill score of 0.200. Accuracy degrades monotonically with distance from the domain, to 1.724 °C beyond 200 km. Against daily climatology, the incumbent choice for gap filling, the pooled reduction is 0.307 °C (95% CI $[0.306, 0.309]$), improving on climatology at 99.99% of grid cells (Cohen's $d = 2.93$, permutation $p \approx 2 \times 10^{-4}$). It also improves on $T_d = T_{\min}$ (FAO-56) by 2.32 °C and on a per-zone ordinary least-squares fit by 1.10 °C. Predictors and bandwidths are selected on the training years alone; allowing selection to see the withheld years would inflate the skill score by 0.009, which bounds the optimism such leakage introduces.

The reconstruction covers 2,178,827 grid cells over 1,461 days (3.18×10^9 values) and introduces no discontinuity at the 2016/2017 boundary: over equal-length four-year windows it differs from the last observed window (mean -0.041 °C, sd 0.176 °C) by no more than two observed windows differ from each other (mean $+0.101$ °C, sd 0.125 °C), with no cell exceeding 1 °C in either comparison. The principal limitation is a small warm tail (0.13% of values above 35 °C), confined to November and to the part of the grid beyond the classified domain, where the source product is itself extrapolated far from any station. Inside that domain the reconstruction produces no value above 35 °C at all; outside it, it attenuates a tail already present in the field it reproduces. We release a per-cell distance to the classified domain so that users may impose their own support threshold.

1 Introduction

Dew-point temperature is a primary input to reference-evapotranspiration, plant-disease-risk and crop-growth models. For Peru the PISCO family of gridded products [1] supplies the national daily fields these applications require. The components are released asynchronously: dew point is available for **1981–2016**, whereas minimum temperature extends to **2020**. Applications needing T_d after 2016 therefore face a four-year gap.

The available substitutes are unsatisfactory. Daily climatology discards all interannual and synoptic variability. The FAO-56 approximation $T_d \approx T_{\min}$ [2] is convenient, but its bias in arid and high-elevation terrain is well documented and has motivated several corrections [3], [4]. A single national regression of T_d on $T_{\min}$ ignores the heterogeneity of Peruvian climate, which spans hyper-arid coast, high Andes and humid Amazon within a few hundred kilometres.

We therefore ask whether the $T_{\min} \to T_d$ relationship learned on the 1981–2016 overlap reconstructs T_d for 2017–2020 more accurately than these incumbents. Two properties distinguish the problem from conventional spatial interpolation and shape the design that follows. First, it predicts **new years at the same grid cells** rather than new locations, so validation must withhold years, not sites. Second, **no observed T_d exists anywhere in the target period**; since T_d lags enter as predictors, reconstruction is necessarily autoregressive, and any honest validation must be run the same way.

Our contributions are: a reconstruction method in which both the predictor set and the temporal smoothing bandwidth adapt to climate zone and season; a validation design appropriate to temporal extrapolation, reported with effect sizes rather than significance alone; the resulting gap-filled field for 2017–2020; and a characterisation of the conditions under which the method fails.

2 Data

Table 1: Data used in this study.

Quantity	Source	Coverage	Role
Dew-point temperature T_d	PISCO v1.1	1981–2016	Target
Minimum temperature $T_{\min}$	PISCO v1.2	1981–2020	Predictor
Grid	PISCO national grid, 0.01°	1,422 × 1,982 = 2,818,404 cells	Domain
Climate zones	SENAMHI national classification [5]	41 mapped units	Unit of model selection

Table 1 lists the inputs. Two of them span different periods, and that asymmetry is the opportunity this paper exploits.

Scope. The grid is a rectangle and the gridded predictors extend across it, so the reconstruction covers 2,667,142 km², about twice the area of Peru. Rather than clip the field to a national boundary, which no atmospheric process respects, we release all of it together with the distance from the classified domain, and we report accuracy against that distance (Section 4.2).

Where the paper speaks of Peru it means the classified domain, which is 48.9% of the released cells; the remainder is territory the source product covers by extension and where, as Section 4.5 shows, it is materially less reliable.

Climate zones. The classification is SENAMHI's national map on the 1981–2010 normal [5], which defines **38 climate types**: 37 Thornthwaite symbols and a glacier class. The published shapefile carries three further map units that are not climate types, two lakes and the coastal *lomas*, giving the 41 units we use. We select per mapped unit rather than per climate type, so the count is 41 throughout; the 37 Thornthwaite symbols in our data match the official list exactly.

The grid has a resolution of 0.01°, roughly 1.1 km, and is a complete rectangle of 1,422 × 1,982 cells spanning 81.40°W–67.19°W and 18.59°S–1.22°N.

The gridded fields are produced by regression kriging within the PISCOeo_pm framework [1]. Dew point is interpolated from **182 stations** on covariates of mean land-surface temperature, elevation, longitude and latitude; T_d is not derived from $T_{\min}$ directly, although the two share covariates. That station count, over a territory of 1.28 million km², is what the support analysis of Section 4.2 is ultimately measuring distance from. Of the 2,818,404 grid cells, **2,178,827** (77.3%) yield reconstructions; the remainder are ocean or lack the climatology or predictor coverage the model requires.

Predictor version. Minimum temperature exists in two PISCO releases, v1.1 (1981–2016) and v1.2 (1981–2020). As predictors of dew point the two are of comparable quality: fitting and scoring an identical configuration on each gives RMSE 1.127 °C against 1.136 °C, a difference of 0.009 °C (Section 7.1). They nevertheless yield different fitted models, the mean contemporaneous $T_{\min}$ coefficient being 0.452 against 0.370 with a paired correlation of 0.685.

Only v1.2 spans the target period, so the reconstruction must use it. We therefore also fit on **v1.2**, so that the coefficients correspond to the release actually supplied at prediction time. Fitting on one release and predicting with the other would apply a sensitivity estimated for one product to another whose fitted sensitivity differs appreciably; we avoid that mismatch by construction rather than quantify it. Since the two releases are of comparable accuracy, this choice costs nothing.

Validation target. The model is scored against the gridded v1.1 T_d, which is the record being extended and therefore the appropriate reference. The reconstruction consequently inherits any bias of that product; station-level evaluation is outside the present scope (Section 6).

3 Methods

3.1 Anomaly formulation

Let $T_i^d(t)$ and $T_i^{\min}(t)$ denote dew-point and minimum temperature at grid cell i on day t, and let $d(t) \in \{1, \ldots, D\}$ with $D = 366$ be the day of year. For each cell and day of year we form a climatological mean over the training years $\mathcal{Y}$,

$$\bar{X}_i(d) = \frac{1}{\mid \{t \in \mathcal{Y} : d(t) = d\} \mid} \sum_{t \in \mathcal{Y},\, d(t)=d} X_i(t),$$

and model anomalies

$$X_i'(t) = X_i(t) - \bar{X}_i\left(\, d(t) \,\right).$$

All fitting and prediction operate on anomalies; the climatology is added back at the end. The climatology is always computed from training years only, so it leaks no information from a withheld period.

3.2 Local regression with a day-of-year kernel

For cell i and day of year d the response is the dew-point anomaly $y_i(t) = T_i^{d'}(t)$ and the candidate predictors are

$$\mathcal{P} = \left\{ 1, T^{\min'}, T_{-1}^{\min'}, T_{-2}^{\min'}, T_{-7}^{\min'}, T_{-30}^{\min'}, T_{-1}^{d'}, T_{-2}^{d'}, T_{-3}^{d'}, T_{-7}^{d'}, T_{-30}^{d'} \right\},$$

where $X_{-\ell}(t) = X(t-\ell)$ and lagged anomalies are formed at the lag day's own day of year. Dew point enters only through lags, never contemporaneously, since it is unobserved in the target period. Writing $\mathbf{x}(t)$ for the vector of predictors retained for this zone and day of year, the model is linear in the anomalies,

$$y_i(t) = \mathbf{x}(t)^\top \boldsymbol{\beta}_{i,d} + \varepsilon_i(t).$$

Observations are weighted by their circular distance in day of year from d. With

$$\delta(a,b) = \min\left(\,|\, a-b \,|, D - |\, a-b \,| \,\right)$$

and a tricube kernel of half-window h,

$$w_h(\delta) = \left[1 - \left(\frac{\delta}{h} \right)^3 \right]^3 \mathbf{1}\{\delta \le h\},$$

the coefficients solve a weighted least-squares problem over all training years simultaneously,

$$\hat{\boldsymbol{\beta}}_{i,d} = \arg\min_{\boldsymbol{\beta}} \sum_t w_h\left(\, \delta(d(t), d) \,\right)\left(\, y_i(t) - \mathbf{x}(t)^\top \boldsymbol{\beta} \right)^2 = \left(\, \mathbf{X}^\top \mathbf{W} \mathbf{X} \right)^{-1} \mathbf{X}^\top \mathbf{W} \mathbf{y},$$

with $\mathbf{W} = \operatorname{diag}\,\{w_h(\delta(d(t), d))\}$. Each (cell, day-of-year) model therefore borrows strength from neighbouring days of year across all training years, with h controlling how much.

3.3 Selection of predictors and bandwidth

Rather than fixing $\mathcal{S} \subseteq \mathcal{P}$ and h globally, we select both for each combination of climate zone z and day of year d, following the multitask design of J. Hwang, P. Orenstein, J. Cohen, K. Pfeiffer, and L. Mackey [6]: predictors are chosen jointly across the grid cells of a zone while coefficients remain per cell. The approach has since been benchmarked against other machine-learning methods for subseasonal forecasting [7] and applied to precipitation [8]; here it is adapted to a reconstruction rather than a forecasting task.

The selection objective is leave-one-year-out cross-validated cosine similarity on anomalies. For a withheld year u, coefficients $\hat{\boldsymbol{\beta}}^{(-u)}$ are obtained by excluding that year, and with $\cos(\mathbf{a}, \mathbf{b}) = \mathbf{a}^\top \mathbf{b} / (\|\mathbf{a}\| \|\mathbf{b}\|)$ the criterion is

$$\mathcal{L}(\mathcal{S}, h) = \frac{1}{|\, y \,|} \sum_{u \in y} \cos\left(\, \mathbf{y}^{(u)}, \mathbf{X}^{(u)} \hat{\boldsymbol{\beta}}_{\mathcal{S},h}^{(-u)} \, \right).$$

Because the weighted normal equations are additive over years, the leave-one-year-out fit is obtained by subtraction, $\mathbf{A}^{(-u)} = \mathbf{A} - \mathbf{A}^{(u)}$, rather than by refitting; the cross-validation is therefore exact at the cost of a single pass.

Predictors are eliminated by backward stepwise search: starting from the full pool, the predictor whose removal most improves $\mathcal{L}$ is dropped, and elimination stops when no removal improves $\mathcal{L}$ by more than a tolerance τ. The search is repeated for each $h \in \{7, 11, 15, 21, 31, 45\}$ and the pair $(\mathcal{S}, h)$ maximising $\mathcal{L}$ is retained, giving $41 \times 366 = 15{,}006$ selected models.

The configuration in production use before this work serves as the reference throughout. It fixes the subset and the bandwidth globally, at

$$\mathcal{S}_0 = \{\, 1, T^{\min\prime}, T^{d\prime}_{-1}, T^{d\prime}_{-2}, T^{\min\prime}_{-1} \,\}, \qquad h = 11,$$

applied identically at every zone and day of year. We call it the fixed five-predictor configuration; it is a subset of $\mathcal{P}$, and the count includes the intercept.

Selection uses a stratified sample of 66,534 cells (up to 2,000 per zone, over all 41 zones). Because the cosine sums are additive across cells, a sample suffices to choose $(\mathcal{S}, h)$; coefficients are subsequently fitted at every cell of the domain. This sample is used only to choose recipes; accuracy is assessed on a different sample of cells (Section 3.5).

3.4 Autoregressive reconstruction

Let t_0 be the last day with observed T_d (31 December 2016). For $t > t_0$ the reconstruction proceeds forward in time,

$$\hat{y}_i(t) = \mathbf{x}(t)^{\top} \hat{\boldsymbol{\beta}}_{i,d(t)}, \qquad \hat{T}^d_i(t) = \hat{y}_i(t) + \bar{T}^d_i\,\big(\,d(t)\,\big),$$

where the dew-point lag entries of $\mathbf{x}(t)$ are taken from observations before the boundary and from the model's own output after it,

$$\big[\,\mathbf{x}(t)\big]_{T^{d\prime}_{-\ell}} = \begin{cases} y_i(t-\ell), & t-\ell \le t_0, \\ \hat{y}_i(t-\ell), & t-\ell > t_0. \end{cases}$$

Minimum-temperature predictors are exogenous and observed throughout, so they never recurse. The recursion is sequential in time but independent across cells.

3.5 Validation design

The quantity of interest is the accuracy of a four-year forward reconstruction, which cannot be observed directly. We therefore reproduce its conditions on years that can be scored: the model is fitted on early years and reconstructs the most recent ones, which are withheld. The primary split withholds four years (fit 1981–2012, reconstruct 2013–2016), matching the length of the real gap; splits withholding one and two years give the dependence on lead time. Reconstruction of the withheld block is fully autoregressive, and the climatology is recomputed from training years only.

Spatial cross-validation, the prevailing strategy for gridded products [9], is not informative here: withholding locations while retaining contemporaneous observations does not reproduce a setting in which the target is unobserved everywhere.

Four methods are scored on the identical withheld block:

1. **Daily climatology**, $\hat{T}_i^d(t) = \bar{T}_i^d(d(t))$, equivalent to predicting zero anomaly;
2. $\boldsymbol{T_d = T_{\min}}$, the FAO-56 approximation [2];
3. **Per-zone ordinary least squares** of T_d on $T_{\min}$;
4. **The selected model** of Section 3.3.

We report root-mean-square error, mean absolute error, bias and correlation, and the skill score relative to climatology. Bias is the mean of prediction minus observation, so a negative value means the reconstruction sits below the observed field. The skill score is

$$\mathrm{SS} = 1 - \frac{\mathrm{RMSE}_{\mathrm{model}}}{\mathrm{RMSE}_{\mathrm{climatology}}}.$$

This is a ratio of root-mean-square errors, not of mean-square errors; a skill score defined on the squared scale would give a larger number for the same forecast. It is also a different quantity from the cosine objective $\mathcal{L}$ of Section 3.3, which selection maximises in sample. The two are reported on different scales and are not comparable.

Differences are summarised by effect size rather than significance alone: the paired per-cell difference in RMSE with 95% confidence intervals from a location-block bootstrap, the fraction of cells improved, Cohen's $d = \bar{\Delta}/s_{\Delta}$ on the paired differences, and a sign-flip permutation null with 5,000 draws. Scoring uses a sample independent of the one used for selection: 64 of the 8,192 spatial partitions of the domain, holding 22,024 cells. Of these, 17,027 have an observed dew-point record over the withheld block and are scored on all 1,461 of its days (n = 24,876,447 cell-days); the remainder fall outside the coverage of the gridded T_d product, which spans 2,178,827 of the 2,818,404 cells of the minimum-temperature grid. The scored cells lie in 33 of the 41 zones. The eight zones with no scored cell hold 2,855 cells, 0.13% of the released field, the two largest being glacier and lake surfaces; accuracy is not validated for them.

Model selection itself can inflate the result, through two distinct routes. If the selection score is computed over all fitting years, recipes chosen on the full 1981–2016 period have already been exposed to the years that are later withheld for scoring. Independently, choosing the best of many candidate recipes favours those whose score is high partly by chance, so the winning score exceeds the model's true skill even when each individual score is unbiased.

Selection is therefore restricted to the training years in every result reported below: predictors and bandwidth are chosen on 1981–2012 alone, by the same procedure applied to a 16-partition stratified sample (up to 200 cells per zone), and the withheld period is never seen at any stage. To measure the first route we additionally run selection over all years on a common candidate grid; the difference between the two runs bounds the inflation that such leakage introduces (Section 4.2).

4 Results

4.1 Selected models

Selection retains most of the candidate pool: the full eleven-predictor set is kept for 70.7% of zone-days and no predictor is retained in fewer than 91% of models (Figure 1 c), indicating that the conventional five-predictor configuration is under-specified. The contemporaneous $T_{\min}$ anomaly is retained almost everywhere, as physical reasoning requires.

The selected bandwidth is broad: $h = 45$ days is modal at 43.1%, with the remainder spread across the smaller values (Figure 1 a). Wide windows are preferred because the $T_{\min} \to T_d$ relationship varies smoothly with day of year, so pooling neighbouring days buys sample size at little cost in bias. The preference is not uniform across the domain (Figure 1 b), and the between-zone variation proves consequential (Section 4.5).

The cross-validated cosine objective $\mathcal{L}$ exceeds that of the fixed five-predictor configuration by $+0.015$ on average (median $+0.008$), with no gain for 16.8% of zone-days (Figure 1 d). These are in-sample quantities computed on the same objective that selection maximises; the out-of-sample assessment follows.

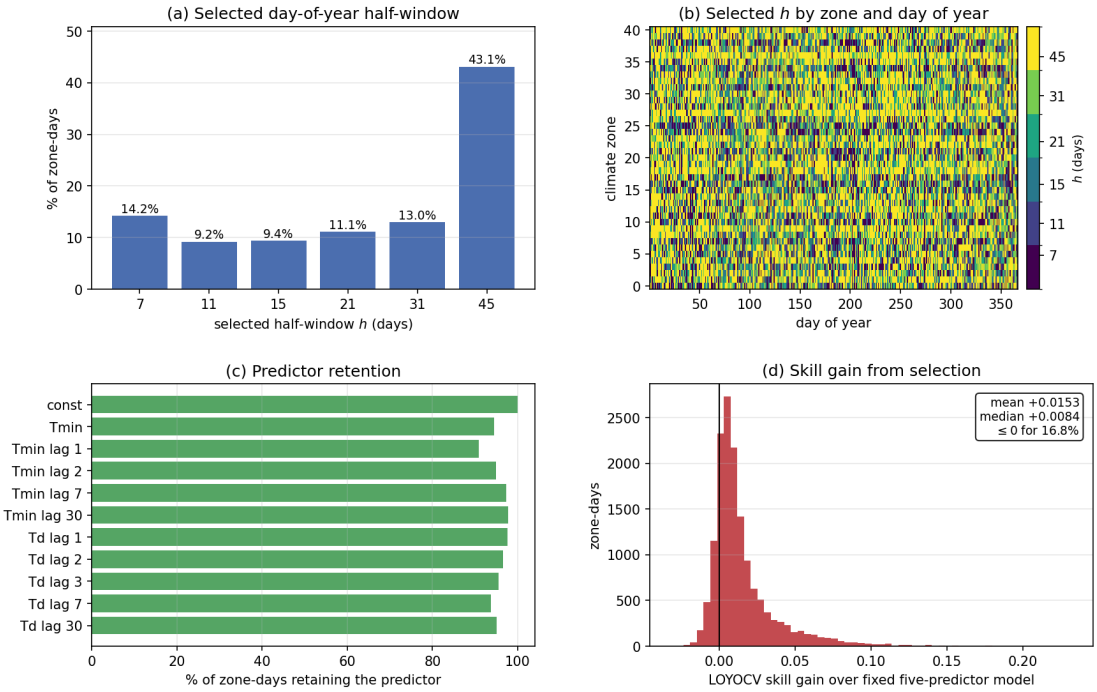


Figure 1: Selected models. (a) distribution of the day-of-year half-window h; (b) h by climate zone and day of year; (c) fraction of models retaining each candidate predictor; (d) gain in the cross-validated cosine objective over the fixed five-predictor configuration.

4.2 Accuracy on withheld years

The selected model outperforms every incumbent at every lead time (Table 2, Figure 2). Against climatology it reduces RMSE by 25% at a one-year lead and by 20% at four years; against the FAO-56 approximation the reduction exceeds 2.3 °C throughout.

Table 2: Root-mean-square error (°C) on withheld years, pooled over all scored cell-days (n = 24,876,447 for the four-year split). The skill score is relative to climatology.

Withheld period	Climatology	$T_d = T_{\min}$	Zone OLS	**Selected model**	Skill score
2016 (1 year)	1.680	3.621	2.378	**1.254**	0.254
2015–2016 (2 years)	1.667	3.676	2.369	**1.276**	0.234
2013–2016 (4 years)	1.535	3.551	2.328	**1.228**	0.200

Whereas Table 2 compares aggregate error, Table 3 asks whether the advantage is large, uniform across space, and distinguishable from chance. It is: the improvement over climatology has a tight confidence interval, holds at essentially every cell, and corresponds to a large effect size.

Table 3: Paired effect sizes on the four-year withheld block (17,027 cells). ΔRMSE is the per-cell reduction in error achieved by the selected model; intervals are from a location-block bootstrap and p from a sign-flip permutation null.

Comparison	ΔRMSE (°C)	95% CI	Skill	Cells improved	Cohen's d	p
vs climatology	$+0.307$	$[0.306, 0.309]$	0.200	99.99%	2.93	2×10^{-4}
vs $T_d = T_{\min}$	$+2.323$	$[2.300, 2.345]$	0.654	99.98%	1.37	2×10^{-4}
vs zone OLS	$+1.101$	$[1.082, 1.118]$	0.473	99.78%	1.05	2×10^{-4}

The selected model has bias -0.033 °C and correlation 0.990. Skill declines monotonically with lead time ($0.254 \to 0.234 \to 0.200$), the expected growth of error under autoregression, without collapsing over four years (Figure 3).

These are pooled figures, and pooling is misleading here. Half the grid lies beyond the domain in which the climate classification is defined and in which the source product is constrained by stations, so the pooled error describes neither the region the reconstruction is used in nor the region it extends into. Splitting on that external criterion, fixed before any error was examined, gives Table 4.

Table 4: Four-year withheld block by observational-support tier: distance bands from the classified domain of, respectively, inside it, 0–50, 50–100, 100–200 and beyond 200 km. Every column degrades monotonically outward.

Support tier	Cells	RMSE	MAE	Bias	r	Clima-tology	Skill
within	8,306	**0.946**	0.697	$+0.010$	0.994	1.264	**0.251**
near	1,779	1.139	0.831	-0.079	0.988	1.464	0.222
mar-ginal	1,433	1.276	0.943	-0.044	0.987	1.582	0.193
extrapo-lated	2,563	1.374	1.017	-0.070	0.980	1.682	0.183
unsup-ported	2,946	1.724	1.283	-0.089	0.953	2.036	0.153

Within the classified domain the reconstruction attains $\mathrm{RMSE} = 0.946$ °C at a skill score of 0.251 and a bias of $+0.010$ °C, essentially unbiased; the pooled 1.228 °C is an average dragged by the unsupported area, where error reaches 1.724 °C. Every column of Table 4 degrades in the same direction, which is what makes the split a description of the data rather than a choice of subset: an arbitrary partition would not order five tiers monotonically in RMSE, MAE, absolute bias, correlation and skill at once. Skill also declines with lead time within every tier separately, so the lead-time result of Figure 3 is not an artefact of mixing regions.

Predictor and bandwidth selection uses the training years only, everywhere in this paper. To measure what that restriction costs, we repeated the four-year experiment on a common candidate grid $h \in \{7, 11, 15, 21\}$, once with selection confined to 1981–2012 and once with selection allowed to see 2013–2016 as well. Skill rises from 0.197 to 0.206 when the withheld years are visible, so leakage of this kind inflates skill by 0.009, about 0.013 °C in RMSE, or 4% of the improvement over climatology. The bias is small but comparable to the differences between design choices reported here, which is why the restriction is enforced rather than assumed harmless. Widening the candidate grid to $h \in \{7, 11, 15, 21, 31, 45\}$ under the same leak-free protocol recovers 0.003 of skill ($0.197 \to 0.200$); that configuration is the one reported throughout.

Corrected forms of the FAO-56 baseline. Established practice corrects $T_d = T_{\min}$ in arid and semi-arid regions, by an aridity-dependent adjustment [3], by a function of temperature and precipitation [4], or by a fixed -2 °C offset [10]. We evaluate the offset, being the form that needs no additional inputs. It does not transfer to this domain (Table 5). The gridded T_d lies systematically *above* $T_{\min}$ here (bias -2.13 °C), which is the opposite of the regime the offset addresses, so subtracting 2 °C increases the error in every Thornthwaite moisture class, including the semi-arid and arid classes for which it was devised: RMSE rises from 2.634 to 3.760 °C in class D and from 2.089 to 2.797 °C in class E. Scanning the offset continuously, the best-fitting single value is $+2.25$ °C, of opposite sign to the convention, and yields RMSE 2.847 °C. Even that optimal member of the offset family remains far behind the selected model.

Table 5: Corrected forms of the FAO-56 baseline, scored on the identical withheld observations (n = 24,876,447). The conventional arid-region offset increases error over this domain.

Baseline variant	RMSE	MAE	Bias
$T_d = T_{\min}$ (uncorrected)	3.551	2.735	−2.126
$T_{\min} - 2$ °C in arid and semi-arid classes	3.633	2.854	−2.367
$T_{\min} - 2$ °C everywhere	5.011	4.263	−4.126
$T_{\min} + 2.25$ °C (best single offset)	2.847		
Selected model, for reference	*1.228*	*0.881*	−0.033

Seasonal bias is also reduced. The fixed configuration carries a December warm bias of +1.45 °C (Section 7.1); under per-zone, per-day-of-year selection the December bias falls to +0.02 °C, with all months within ±0.41 °C.

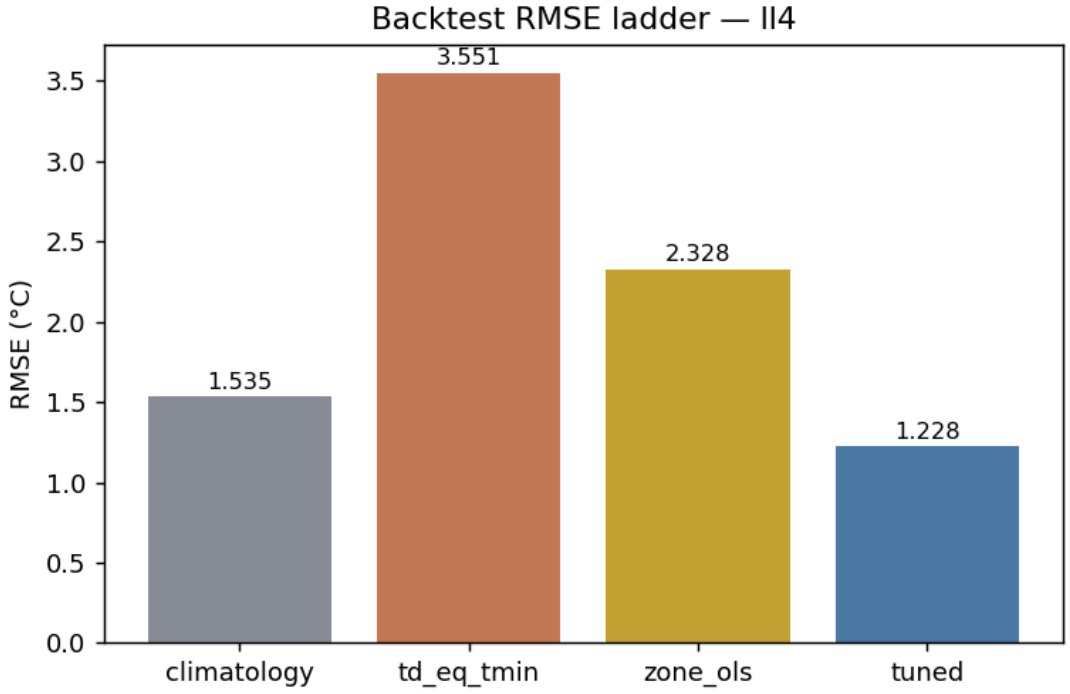


Figure 2: Root-mean-square error on the four-year withheld block for the selected model and the three incumbents.

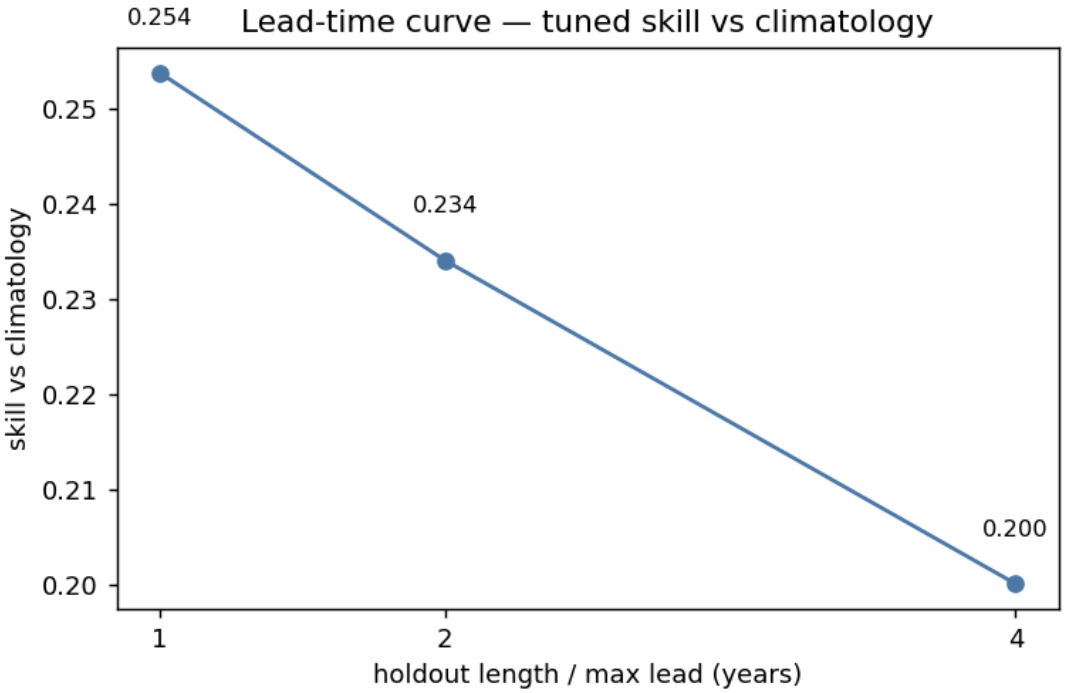


Figure 3: Skill score relative to climatology as a function of lead time.

4.3 The reconstructed field

Applying coefficients fitted on 1981–2016 to the observed 2017–2020 minimum temperature yields 3,183,266,247 values with no missing entries. National mean T_d is 19.39 °C (range −13.29 to 41.16 °C) and annual means are stable across the period (19.36, 19.30, 19.41 and 19.51 °C), showing no drift under sustained autoregression.

The field reproduces the expected national structure, with a humid Amazon near 25 °C, a sharply delineated cold Andean cordillera and the coastal gradient (Figure 4), and continues the observed seasonal cycle across the 2016/2017 boundary with matching amplitude and phase (Figure 5).

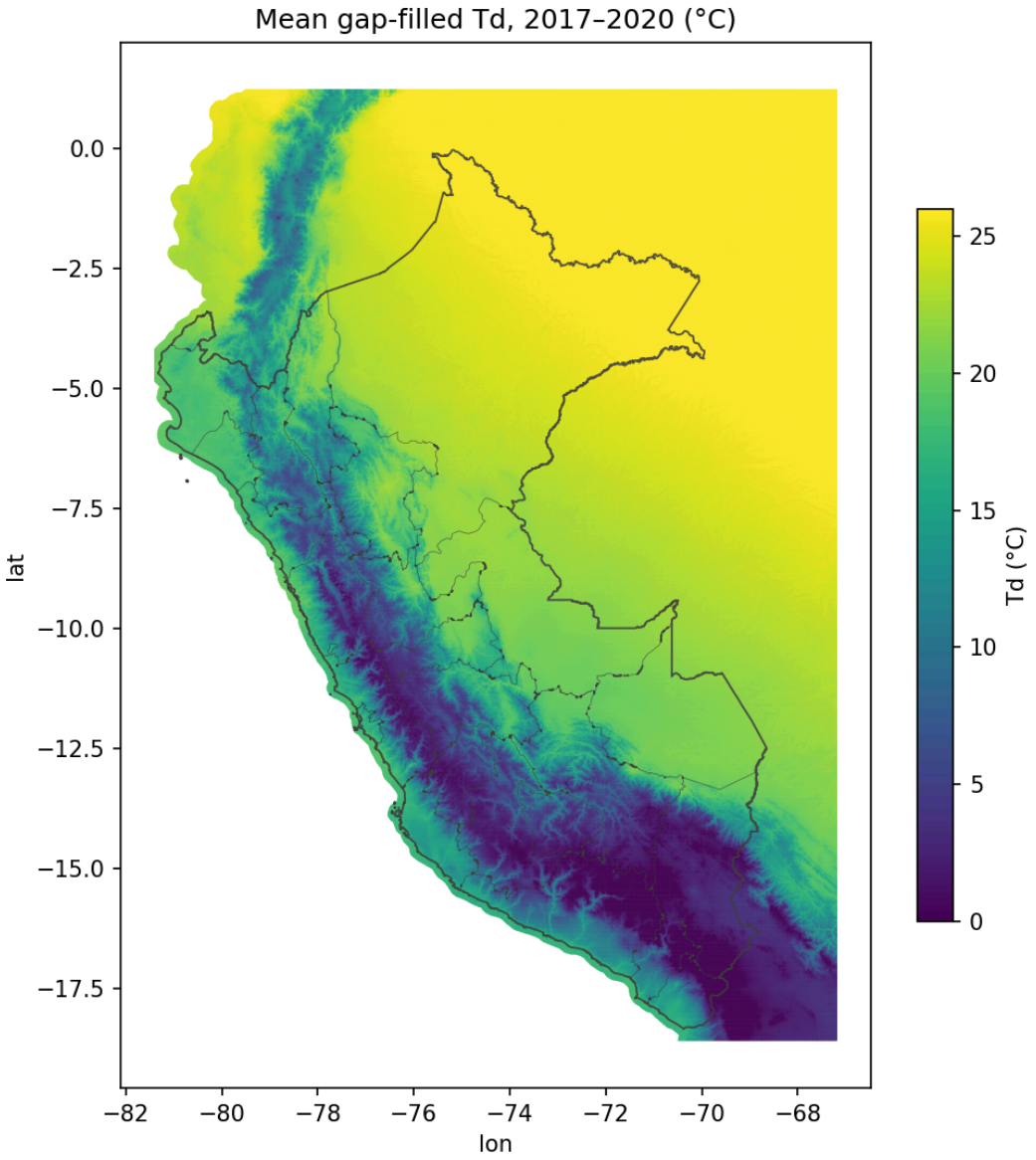


Figure 4: Mean reconstructed dew point, 2017–2020.

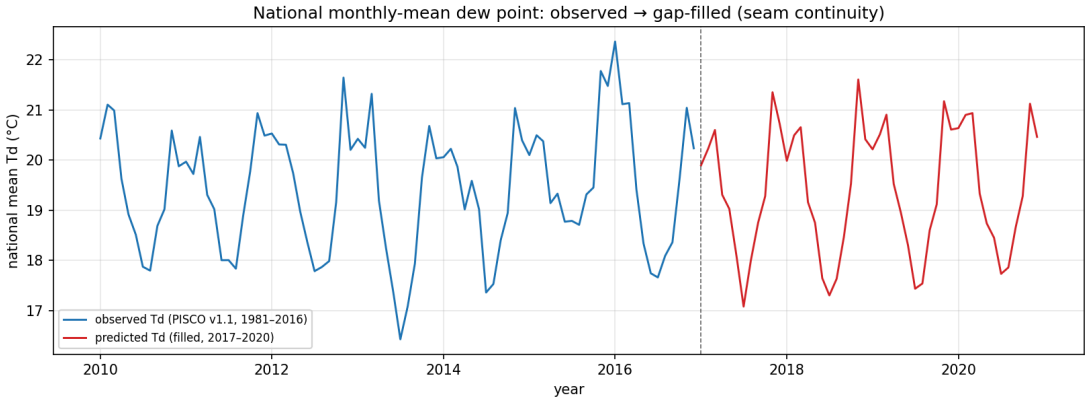


Figure 5: National monthly-mean dew point, 2010–2020: observed through 2016 and reconstructed thereafter.

4.4 Continuity at the record boundary

A reconstruction appended to an observed record must not introduce an artificial step. Comparing adjacent months across the boundary would confound such a step with the seasonal cycle, so we compare equal-length full-year windows, which removes seasonality, and include a model-free control: the difference between two consecutive observed four-year windows, which measures the period-to-period variability of the observed record itself. Table 6 reports both.

Table 6: Difference in annual-mean dew point (°C) between four-year windows, over 2,178,827 cells. The first row involves no model and quantifies natural variability.

Comparison	Mean	sd	1st–99th pct	Cells $\lvert \Delta \rvert > 1$ °C
Observed 2013–16 — observed 2009–12 (control)	+0.101	0.125	−0.18 … +0.42	0.0%
Reconstructed 2017–20 — observed 2013–16	−0.041	0.176	−0.36 … +0.54	0.0%

The reconstruction differs from the last observed window by no more than two observed windows differ from each other: its spread is 1.4 times the control, its mean is closer to zero, and no cell in either comparison differs by more than 1 °C (Figure 6). The difference field is spatially smooth and coherent, without the blockiness or speckle a numerical artefact would produce.

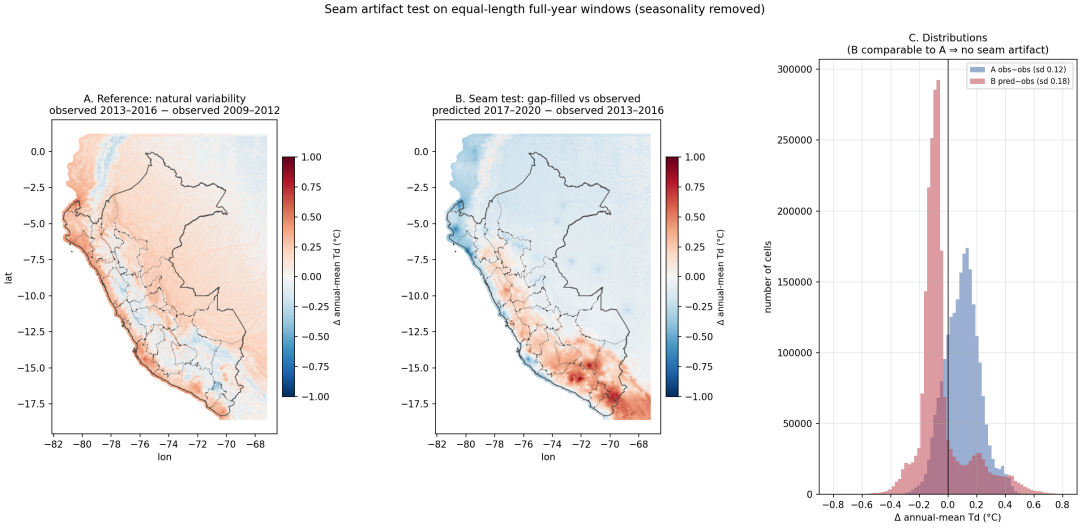


Figure 6: Continuity at the record boundary on equal-length full-year windows. (A) natural variability between two observed windows; (B) reconstructed against observed; (C) distributions.

4.5 Extreme values

A small fraction of reconstructed values is physically implausible: 0.1326% exceed 35 °C, with a maximum of 41.16 °C, against a world record near 35 °C. These values are sharply localised, and their location identifies their origin.

They lie entirely outside the classified domain. Of the 132,920 cells carrying any value above the threshold, **none** falls inside a SENAMHI polygon and all 132,920 fall outside it. That statement invites the objection that a national boundary is a political line no atmospheric process respects, and the objection is correct: the boundary is not the cause. It is a proxy for where the station network that constrains the source fields ends, and the effect it stands for is continuous. Scoring the exceedance rate against distance from the classified domain shows a gradient, not a step (Table 7).

Table 7: Exceedance above 35 °C by distance from the classified domain, over the whole released field (2,178,827 cells, 3.18×10^9 values).

Support tier	km from domain	Cells	% of field	Cells affected	Values above 35 °C
within	inside	1,066,118	48.9%	**0**	**0.0000%**
near	0–50	228,643	10.5%	10	0.0000%
marginal	50–100	179,581	8.2%	391	0.0003%
extrapolated	100–200	328,136	15.1%	18,888	0.0255%
unsupported	beyond 200	376,349	17.3%	113,631	0.7452%

Nearly 60% of the field, everything within 50 km of the classified domain, contains ten affected cells in total out of 1,294,761. The rate then rises by three orders of magnitude across the

remainder. The first fifty kilometres beyond the boundary are as clean as its interior, which is the signature of distance from observational support rather than of a border.

Holding the climate class fixed gives the same gradient, so it is not a difference between classes: within the single class in which every exceedance occurs, the affected fraction runs 0% inside, 0% for the first 20 km beyond, 0.42% at 50–100 km, 10.3% at 100–200 km and 56.3% beyond 300 km.

The values are moreover inherited rather than introduced. Scored on the withheld years, the source product itself carries the larger tail, and the reconstruction attenuates it (Table 8).

Table 8: Extreme dew point in the source product and in the reconstruction, on the four-year withheld block. Inside the classified domain the reconstruction produces no value above 35 °C at all, and its maximum lies below the observed maximum.

	Observed T_d, maximum	Observed, above 35 °C	Reconstructed, maximum	Reconstructed, above 35 °C
Inside the domain	38.34 °C	0.0046%	**34.10 °C**	**0.0000%**
Outside	46.47 °C	0.6724%	40.11 °C	0.2147%

Inside the domain the method therefore produces no implausible value whatever; outside it, it returns roughly a third of the exceedance rate of the field it is reproducing. The implausible values originate in the gridded dew point where that product is itself extrapolated far from any station.

Within the affected region the exceedances concentrate almost entirely in **November**: pooled over 2017–2020, 1.556% of November values exceed the threshold against 0.027% in October, and only 50 exceedances occur in total between January and July (Figure 7). They occur in a single Thornthwaite class, `A (r) A'` (perhumid, no water deficiency, megathermal), which is the class covering the eastern lowlands and the extrapolated area beyond them (Figure 8).

The concentration is not explained by accumulation of autoregressive error. Annual incidence is non-monotonic (0.2015%, 0.1696%, 0.0293% and 0.1299% for 2017–2020) and exceedances occur in short episodes, 40.6% of them isolated single days, with a mean run of 3.42 days. Progressive error growth of the kind reported for free-running autoregressive models [11] would produce lengthening runs and monotonic growth with lead time; neither is present.

Nor is it a failure of average fit. In the affected zone, the November cosine objective (0.7317) is indistinguishable from December (0.7333) and January–July (0.7324), and higher than August–September (0.7030); in-sample R^2 behaves likewise. The failure affects the tail of the predictive distribution rather than its centre, consistent with annual means remaining stable to 0.21 °C.

The seasonal concentration has a separate explanation, in the selected bandwidth. Although $h = 45$ is modal nationally, this class prefers narrower windows (median $h = 21$, 31.7% at $h = 45$), and **November is its narrowest band**: 47% of November days select $h \leq 11$ against 23% in October, with median $h = 15$ in November against 31 in both adjacent months. Selection responds to a rapidly changing time of year by shrinking the window, trading bias for variance. Since the quantity measured is an exceedance rate above a fixed threshold, variance rather

than bias governs it, so a modest increase in predictive spread, invisible in mean accuracy, moves the rate sharply.

That mechanism accounts for *when* the exceedances occur and cannot account for *where*. Predictors and bandwidth are selected per class and day of year, so every cell of this class shares one recipe on a given day; a shared recipe cannot produce a rate of zero at one longitude and a third at another. The spatial pattern is a property of the data each cell is fitted to, and Table 7 identifies which property.

The exceedance peak moves between years, at day of year 313 in 2017, 331 in 2018, 313 in 2019 and 331 in 2020, with the distribution mean shifting between days 314.9 and 322.2.

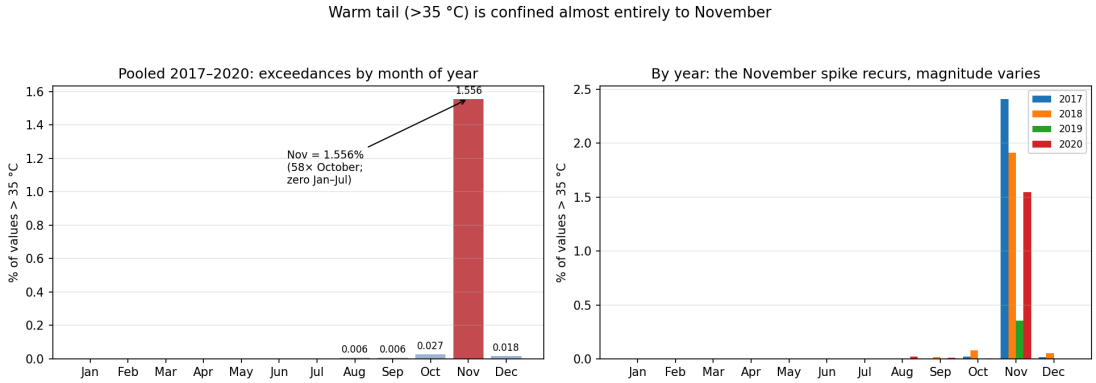


Figure 7: Exceedances above 35 °C by month of year, pooled (left) and by year (right).

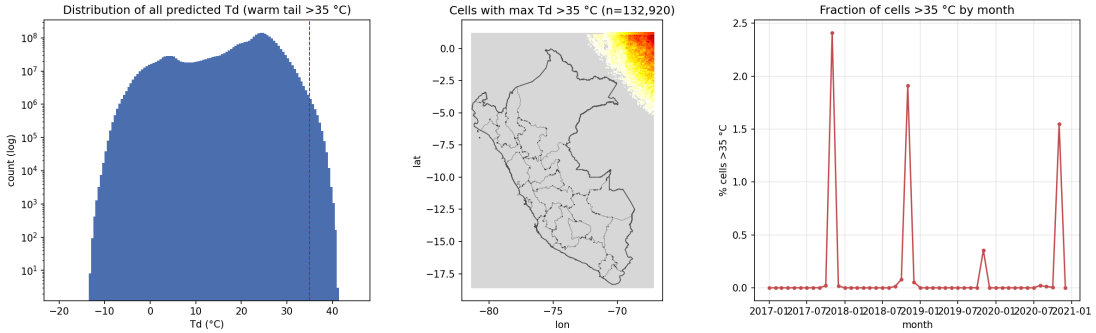


Figure 8: Distribution of reconstructed dew point, spatial location of affected cells, and monthly incidence.

5 Discussion

5.1 Estimating humidity from temperature

The FAO-56 identity $T_d \approx T_{\min}$ [2] is a starting point rather than the state of the art. Its bias in arid, semi-arid and high-elevation terrain is documented and several corrections are in routine use: an aridity-dependent adjustment driven by precipitation and potential evapotranspiration [3], a correction from temperature and precipitation validated at alpine stations [4], and a −2 °C offset widely applied in arid regions [10]. Disaggregation frameworks implement both the identity and fitted alternatives precisely because it fails in a site- and regime-specific way [12]. Those corrections are nonetheless regime-specific themselves: over the present domain the conventional arid-region offset increases error in every moisture class, and the best-fitting

offset has the opposite sign (Table 5). Reported corrections should not be transferred between products and regions without re-estimation.

The most directly comparable result is that of D. Althoff and L. Rodrigues [13], who calibrate a temperature-based approach over the Brazilian Cerrado, a tropical biome with a pronounced dry season, and obtain corrections of **opposite sign in the two seasons**: -0.93 °C in the wet season ($T_d > T_{\min}$) against $+0.57$ °C in the dry season ($T_d < T_{\min}$). The coupling is thus not a fixed offset but a regime-dependent relationship. Selecting a separate model for each climate zone and day of year generalises that observation, and the magnitude of the improvement over both $T_d = T_{\min}$ ($+2.33$ °C) and a per-zone linear fit ($+1.11$ °C) is consistent with it: a single global or even per-zone linear relationship is too rigid for this domain.

5.2 Interpretation of the selected bandwidths

The preference for wide day-of-year windows has an independent precedent: G. Chabot-Couture, K. R. Nigmatulina, and P. Eckhoff [14] interpolate dew point with day-of-year-dependent parameters smoothed over 31-day windows, a comparable scale. Wide windows are appropriate where the relationship varies smoothly with season. Where it does not, selection narrows them, and Section 4.5 shows that this local narrowing, rather than the national preference for broad pooling, governs the method's failure mode. Bandwidth selection is therefore not merely a tuning detail: it is the mechanism by which the method adapts to seasonal structure, and also the mechanism by which it fails when that structure is abrupt.

5.3 Temporal extrapolation and its validation

Work described as gap filling usually reconstructs values missing *inside* an existing record, using surrounding observations in space and time [15]. The present problem differs in kind: the target is unobserved everywhere throughout the period to be filled. The methodological consequence is that spatial cross-validation, which withholds locations while retaining contemporaneous observations [9], cannot measure the quantity of interest, and only a forward temporal hold-out reproduces the operational setting. Systematic reporting of incumbent baselines and effect sizes is likewise uncommon in this literature, and we regard the validation design as part of the contribution rather than as supporting material.

5.4 Stability of long autoregressive reconstruction

Because dew-point lags are predictors and no observation exists in the target period, the reconstruction is a 1,461-step recursion. The deep-learning climate literature documents that such rollouts can degrade progressively or diverge through error accumulation, and that explicit multi-step training is needed to stabilise them [11]. No comparable degradation appears here: skill decays gracefully with lead time, annual means are stable to 0.21 °C over four years, and the incidence of extreme values is non-monotonic in time rather than growing.

Two properties of the formulation account for this. The dominant predictors are anomalies of observed, exogenous $T_{\min}$, supplied from data at every step and never recursed; only the dew-point lag channel is autoregressive. And predictions are formed as anomalies added to a fixed per-cell climatology, which anchors the recursion to a physically sensible level. Emulators that must propagate their entire state forward have neither property. This suggests that hybrid recursions combining observed exogenous forcing with a climatological anchor are substantially more stable over long horizons than fully autoregressive ones, and that stability concerns from that literature do not transfer directly to this class of model.

5.5 Origin of the extreme values

Two separate questions arise, where the exceedances occur and when, and they have different answers. The spatial answer is observational support: the rate is zero inside the classified domain and rises monotonically with distance beyond it (Table 7), and the source product carries a larger tail there than the reconstruction does (Table 8). Regression kriging reverts towards the covariate-driven mean far from data, and for dew point those covariates are mean land-surface temperature, elevation, longitude and latitude [1]; where none of the 182 stations constrains the humidity relationship, the resulting dew point need not respect the physical ceiling that station data would impose. The reconstruction inherits that behaviour through the lagged dew-point predictors while damping it, since a fitted regression on anomalies cannot reproduce the full spread of its target.

The seasonal answer is also a property of the source product, and it is measurable in the training data. Observed dew point in this class carries a November spike at every support tier, isolated between two ordinary months (Table 9). The spike grows with distance from the station network, from $+1.69$ °C inside the classified domain to $+2.29$ °C beyond 200 km, and the November level it is added to grows with distance as well, from 25.7 °C to 29.1 °C. Neither alone reaches the threshold; together, only beyond the domain, the upper tail crosses it. That is why the exceedances are simultaneously a November phenomenon and an outside-the-domain phenomenon without either fact explaining the other.

Table 9: Observed dew point in the affected Thornthwaite class, from the 1981–2016 training record. The rise is against the mean of October and December.

Support tier	November level	Rise above adjacent months
within	25.7 °C	+1.69
near	27.7 °C	+1.88
unsupported	29.1 °C	+2.29

A single month that rises and returns is not the signature of a seasonal regime change, and we do not attribute it to one. The exceedance fraction in the observed record runs 4.97% in October, 22.87% in November and 2.67% in December; a transition between regimes would not reverse within a month. We are unable to test a precipitation explanation in the affected area, because the gridded precipitation product for Peru does not extend there either: 23 of its cells reach the region, with values inconsistent with a rainforest. Where precipitation is available, in the part of this class inside the classified domain, the wet season peaks in April rather than November. The affected area is also equatorial, at a mean latitude of 0.79 °S with 79% of cells north of 2 °S, so the literature on the southern-Amazonian dry-to-wet transition does not apply to it.

The artefact class is not new. G. Chabot-Couture, K. R. Nigmatulina, and P. Eckhoff [14] report interpolated dew point exceeding interpolated air temperature in a day-of-year-parameterised dew-point interpolation, and recommend capping. Nonparametric smoothing is in any case poorly suited to an abrupt feature, since such methods require the underlying function to be smooth and jumps must be estimated separately [16]; a November spike of this kind is exactly such a feature. Amazonian humidity is additionally non-stationary, which compounds the difficulty of estimating the relationship from a 1981–2016 sample: dry-season vapour pressure

deficit has been rising [17] and vertical drying has been observed over deforested Amazonia [18]. Our own test of the training period (Section 6) finds a drift consistent with this.

6 Limitations

Extreme values. The warm tail characterised in Section 4.5 is released unmodified, and accompanied by a per-cell distance to the classified domain so that users may impose a support threshold of their own choosing. We prefer this to editing the values. The exceedances are not an artefact we introduce, being both absent inside the domain and milder than the source product outside it, so suppressing them would misrepresent the field being extended while discarding information about where that field is weakly constrained. A threshold at 50 km removes essentially all of them (Table 7) at the cost of the extrapolated area.

Constraining the minimum bandwidth would address the seasonal concentration, but not the spatial one, which no choice of recipe can reach (Section 4.5); it remains untested.

A post-hoc bound is also available, and the appropriate form is a bound on the *excess* over same-day minimum temperature rather than on $T_{\min}$ itself. Capping at $T_d \le T_{\min}$ is not viable, because 79.6% of reconstructed values exceed same-day $T_{\min}$ by a mean of 3.0 °C, as does the observed gridded T_d; such a cap would alter most of the field by a mean of 2.38 °C. The exceedances are extreme on this measure too, exceeding same-day $T_{\min}$ by 11.4 °C against a typical 3.0 °C, and are therefore separable from the bulk of the distribution (Table 10). A bound of $T_d \le T_{\min} + 8$ °C removes 99.1% of them while altering 2.2% of values by a mean of 0.030 °C over the field, three orders of magnitude less perturbation than a cap at $T_{\min}$. Whether to apply such a bound to the released field is a judgement about whether a small targeted perturbation is preferable to retaining known-implausible values.

Table 10: Effect of bounding the dew-point excess over same-day minimum temperature, evaluated over all 3.18×10^9 reconstructed values.

Bound	Values altered	Mean shift over all values (°C)	Exceedances removed
$T_d \le T_{\min}$	79.6%	2.38	100%
$T_d \le T_{\min} + 5$	13.31%	0.222	100.0%
$T_d \le T_{\min} + 6$	7.67%	0.120	99.98%
$T_d \le T_{\min} + 8$	2.23%	0.030	99.07%
$T_d \le T_{\min} + 10$	0.54%	0.006	89.60%
$T_d \le T_{\min} + 12$	0.09%	0.0007	30.83%

We considered and rejected an attribution of the November concentration to the regional dry-to-wet transition (Section 5). What remains unexplained is the November spike in the source product itself, which is present at every support tier and is not ours to diagnose from the gridded fields alone.

Attribution of the boundary signal. The reconstructed-minus-observed field in Figure 6 carries a coherent regional pattern. Because no observed T_d exists after 2016, that comparison alone cannot separate a genuine 2017–2020 climate anomaly from regional model bias; the bound comes from Section 4.2, where regional errors are measured on withheld years.

Shared covariates. Gridded T_d and $T_{\min}$ share land-surface-temperature and elevation predictors in their construction, so validation against gridded T_d may flatter the mapping relative to an independent reference.

Validation coverage. The released field covers all 41 mapped units, but accuracy is measured on a 17,027-cell sample falling in 33 of them (Section 3.5). The eight unscored zones hold 2,855 cells, 0.13% of the released field, and include glacier and lake surfaces, whose $T_{\min} \to T_d$ relationship need not resemble that of the land zones. The reported error statistics do not extend to them.

No station-level evaluation. The target is the gridded product being extended, which is the correct reference for this task, but the reconstruction inherits that product's biases. Station-level evaluation is the natural next step.

Baseline correction. The FAO-56 baseline is reported in its uncorrected form because, over this domain, the conventional arid-region offset increases its error whether applied regionally or globally (Table 5). We have not evaluated the aridity-dependent formulations of J. S. Kimball, S. W. Running, and R. Nemani [3] or E. Eccel [4], which use precipitation as well as temperature. Gridded precipitation exists for this domain over the identical period, so this is a gap in the present comparison rather than one the data prevent closing; note only that it does not reach the area where the extremes occur, which the precipitation product does not cover either.

Lead time. Skill decays with lead time, so 2020 rests on a longer recursion than 2017. The lead-time curve quantifies this to four years, the length of the actual gap.

Non-stationarity. The training period is not stationary. Splitting it into halves and regressing the dew-point anomaly on the minimum-temperature anomaly against each half's own day-of-year climatology, the slope falls from 0.716 to 0.661 inside the classified domain, a change of -0.055 with Cohen's $d = -0.64$, in the same direction at 72.8% of cells. Level trends run $+0.097$ °C per decade in dew point and $+0.154$ in minimum temperature inside the domain, and rise with distance from the station network to $+0.293$ and $+0.238$ beyond 200 km, the source product drifting fastest exactly where it is least constrained.

The reported accuracy already absorbs this, since the validation fits early years and predicts later ones, so any drift over the withheld block is inside the error it measures. What the test adds is that the drift is slow relative to a four-year horizon and that it is concentrated outside the domain. Extending the reconstruction much further forward on the same coefficients would not be safe on this evidence.

7 Supplementary material

7.1 Predictor-version sensitivity

The two PISCO minimum-temperature releases yield different fitted models of comparable accuracy. Each release was run through the identical pipeline, fitted and scored on itself, over 7.32×10^6 (cell, day-of-year) fits. The fitted coefficients differ, most clearly in the contemporaneous sensitivity:

Table 11: Coefficients fitted on PISCO v1.1 against v1.2 minimum temperature, paired over identical fits. The contemporaneous sensitivity differs most.

Coefficient	v1.1 mean	v1.2 mean	Mean Δ	Mean $\lvert\,\Delta\,\rvert$	Pearson r
$T_{\min}$ anomaly	0.452	0.370	-0.083	0.130	0.685
T_d lag 1	0.782	0.813	$+0.031$	0.047	0.925
T_d lag 2	-0.019	-0.031	-0.012	0.026	0.920
$T_{\min}$ lag 1	-0.236	-0.182	$+0.054$	0.092	0.669
R^2 (anomalies)	0.787	0.778	-0.009	0.019	0.953

The two self-consistent pipelines agree with one another to $\mathrm{RMSE} = 0.610$ °C ($r = 0.995$) with a mean offset of -0.300 °C. Scored against observed dew point they are nevertheless of comparable accuracy, v1.1 giving marginally lower error and v1.2 being better centred:

Table 12: Accuracy against observed dew point for each predictor version ($n = 7.3 \times 10^6$).

Predictor version	RMSE	MAE	Bias	Pearson r
v1.1	1.127	0.844	$+0.443$	0.982
v1.2	1.136	0.867	$+0.143$	0.979

The accuracy difference between releases, 0.009 °C, is negligible against the 0.307 °C improvement the selected model achieves over climatology: as predictors of dew point the two are effectively equivalent, and choosing v1.2 costs nothing. What differs is the fitted relationship, with a paired correlation of only 0.685 on the contemporaneous term, which is why we fit and predict on the same release.

Note that the -0.300 °C offset above is the difference between two **internally consistent** pipelines, each fitted and applied to its own release. It is not a measurement of the penalty for mixing releases, a configuration we did not run; the case for consistency here is that coefficients should correspond to the predictor actually supplied, not an estimated error bound.

This fixed five-predictor configuration also carries a December warm bias of $+1.45$ °C, which per-zone seasonal selection removes (Section 4.2).

7.2 Computational characterisation

Fitting 2.8×10^6 cells $\times$ 366 days of year is dominated by memory traffic rather than arithmetic, which determines how the method scales.

On CPUs, parallel efficiency $E(p) = T(1)/(p\,T(p))$ is essentially independent of problem size, indicating that the limit is memory bandwidth rather than parallel overhead:

Table 13: Strong-scaling efficiency $E(p)$ for three problem sizes (N = grid cells). Efficiency at fixed p varies by less than 0.01 across a fourfold change in N.

Workers p	$N = 512$	$N = 1024$	$N = 2048$
2	0.939	0.917	0.912
4	0.830	0.826	0.824
8	0.709	0.704	0.699
16	0.562	0.560	0.557
32	0.345	0.342	0.337

Weak scaling at constant work per worker gives $E = 1.000$, 0.927 and 0.846 at $p = 1, 2, 4$.

On a single GPU the same conclusion follows from a roofline analysis. With measured bandwidth 633 GB s^{-1} and 4200 GFLOPS double precision, the ridge lies at an arithmetic intensity of approximately 6.6 FLOP B^{-1}, whereas the pipeline operates between 0.08 and 0.61:

Table 14: Roofline position of each pipeline stage at $N = 512$ cells. All stages lie far below the ridge.

Stage	Arithmetic intensity (FLOP B^{-1})	GFLOPS	Regime
Assemble	0.114	44.7	memory-bound
Convolve	0.083	24.8	memory-bound
Solve	0.615	320	memory-bound
Total	0.105	36.1	memory-bound

Solver throughput reaches 2.09×10^9 fits s^{-1} at large batch sizes. The practical consequence is that the implementation is bounded by data movement, so the productive optimisations are those that reduce traffic, and the full-domain fit remains tractable on a single device.

8 Data and code availability

The reconstruction covers 2017–2020 at the PISCO national grid. It is released with a per-cell support layer giving the signed distance from the classified domain, negative inside, so that users may impose a support threshold of their own rather than inherit ours; Table 7 gives the exceedance rate against that distance, and Section 4.5 the reasoning. Values are released unmodified.

Analysis code implements model selection, coefficient fitting, autoregressive reconstruction and the validation harness; the scripts reproducing each figure and reported statistic accompany the manuscript, together with the retrieved literature records supporting Section 5.

[TODO] Deposit the reconstructed field and a frozen code release, and record their identifiers here.

9 AI-use disclosure

In accordance with CIP's Acceptable Use and Privacy policy, two AI systems were used and each is named with what it did.

The analysis pipeline, validation harness, reconstruction driver and figures were developed and executed with AI assistance (Anthropic Claude, via Claude Code) under the direction and review of the author. The assistant wrote and ran analysis code, produced the figures, and drafted this manuscript.

The literature underpinning Section 5 was located with Ai2′s Asta agents [19]: Paper Finder for five topic searches, and the theory-generation agent [20] for literature-grounded hypotheses about the mechanisms discussed. The retrieval and generation records accompany the manuscript. One hypothesis obtained this way, attributing the November extremes to the Amazonian dry-to-wet transition, was **withdrawn** when our own data contradicted it (Section 5); the agents were used as a source of candidate explanations, not as an authority on them.

No confidential, personal or unpublished third-party data was provided to any assistant; all inputs are institutional gridded climate products.

All quantitative results are pending SME validation (Section 10). Values were read directly from the recorded outputs of the runs cited rather than reproduced from memory. The claim attributed to each cited work was checked against the full text of that work; volumes and page numbers remain incomplete on several entries.

10 SME sign-off

Table 15: Sign-offs required before submission.

Item	Reviewer	Date	Status
Validation design and baselines (Section 3.5)			☐
Withheld-year results and effect sizes (Table 2, Table 3)			☐
Boundary continuity (Table 6)			☐
Disposition of extreme values (Section 4.5)			☐
Withdrawal of the dry-to-wet attribution (Section 5)			☐
Attribution of the boundary pattern (Section 6)			☐
Non-stationarity of the fitted relationship (Section 6)			☐
Release of the reconstructed field and metadata			☐

References

Bibliography

[1] A. Huerta, C. Aybar, W. Lavado-Casimiro, and others, "PISCOeo_pm, a reference evapotranspiration gridded database based on FAO Penman–Monteith in Peru," *Scientific Data*, vol. 9, no. 328, 2022, doi: 10.1038/s41597-022-01373-8.

[2] R. G. Allen, L. S. Pereira, D. Raes, and M. Smith, "Crop Evapotranspiration: Guidelines for Computing Crop Water Requirements," Rome, FAO Irrigation and Drainage Paper 56, 1998.

[3] J. S. Kimball, S. W. Running, and R. Nemani, "An improved method for estimating surface humidity from daily minimum temperature," *Agricultural and Forest Meteorology*, vol. 85, pp. 87–98, 1997, doi: 10.1016/S0168-1923(96)02366-0.

[4] E. Eccel, "Estimating air humidity from temperature and precipitation measures for modelling applications," *Meteorological Applications*, vol. 19, pp. 118–128, 2012, doi: 10.1002/met.258.

[5] Servicio Nacional de Meteorología e Hidrología del Perú (SENAMHI), *Climas del Perú: Mapa de Clasificación Climática Nacional*, 1st ed. Lima, Perú: SENAMHI, Proyecto Apoyo a la Gestión del Cambio Climático, 2021.

[6] J. Hwang, P. Orenstein, J. Cohen, K. Pfeiffer, and L. Mackey, "Improving Subseasonal Forecasting in the Western U.S. with Machine Learning," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA: ACM, 2019. doi: 10.1145/3292500.3330674.

[7] S. He, X. Li, T. DelSole, P. Ravikumar, and A. Banerjee, "Sub-Seasonal Climate Forecasting via Machine Learning: Challenges, Analysis, and Advances," in *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)*, 2021. doi: 10.1609/aaai.v35i1.16090.

[8] C. Wang, Z. Jia, Z. Yin, F. Liu, G. Lu, and J. Zheng, "Improving the Accuracy of Subseasonal Forecasting of China Precipitation With a Machine Learning Approach," *Frontiers in Earth Science*, vol. 9, no. 659310, 2021, doi: 10.3389/feart.2021.659310.

[9] M. Kilibarda, T. Hengl, G. Heuvelink, B. Gräler, E. Pebesma, and M. Perčec Tadić, "Spatio-temporal interpolation of daily temperatures for global land areas at 1 km resolution," *Journal of Geophysical Research: Atmospheres*, vol. 119, pp. 2294–2313, 2014, doi: 10.1002/2013JD020803.

[10] L. Huang, Y. Luo, J. M. Chen, Q.-h. Tang, T. Steenhuis, and W. Cheng, "Satellite-based near-real-time global daily terrestrial evapotranspiration estimates," *Earth System Science Data*, vol. 16, pp. 3993–4019, 2024, doi: 10.5194/essd-16-3993-2024.

[11] F. Gallusser, S. Hentschel, A. Krause, and A. Hotho, "Exploring Design Choices for Autoregressive Deep Learning Climate Models," *arXiv preprint*, 2025, doi: 10.48550/arXiv.2505.02506.

[12] K. Förster, F. Hanzer, B. Winter, T. Marke, and U. Strasser, "An open-source MEteoro-LOgical observation time series DISaggregation Tool (MELODIST v0.1.1)," *Geoscientific Model Development*, vol. 9, pp. 2315–2333, 2016, doi: 10.5194/gmd-9-2315-2016.

[13] D. Althoff and L. Rodrigues, "Improvement of reference crop evapotranspiration estimates using limited data for the Brazilian Cerrado," *Scientia Agricola*, vol. 80, no. e20210229, 2023, doi: 10.1590/1678-992X-2021-0229.

[14] G. Chabot-Couture, K. R. Nigmatulina, and P. Eckhoff, "An Environmental Data Set for Vector-Borne Disease Modeling and Epidemiology," *PLoS ONE*, vol. 9, no. 5, p. e94741, 2014, doi: 10.1371/journal.pone.0094741.

[15] R. Serrano-Notivoli, S. Beguería, and M. de Luis, "STEAD: a high-resolution daily gridded temperature dataset for Spain," *Earth System Science Data*, vol. 11, pp. 1171–1188, 2019, doi: 10.5194/essd-11-1171-2019.

[16] T. Zhang, "Testing for jumps in the presence of smooth changes in trends of nonstationary time series," *Electronic Journal of Statistics*, vol. 10, pp. 706–735, 2016, doi: 10.1214/16-EJS1127.

[17] Y. Mu *et al.*, "Increasing dry-season vapor pressure deficit in the Amazon Basin: historical trends, future projections, and links to deforestation," *Environmental Research Communications*, vol. 7, no. 101001, 2025, doi: 10.1088/2515-7620/ae0c5b.

[18] R. Hierro, "Observational Evidence of Vertical Drying Over Deforested Amazonia From GPS-RO Data," *International Journal of Climatology*, vol. 45, no. e70184, 2025, doi: 10.1002/joc.70184.

[19] Allen Institute for Artificial Intelligence (Ai2), "Asta: agentic assistants for scientific research." 2025.

[20] P. Jansen, P. Clark, D. Downey, and D. S. Weld, "Generating Literature-Driven Scientific Theories at Scale," *arXiv preprint*, 2026, doi: 10.48550/arXiv.2601.16282.