REVISTA INVESTIGACION OPERACIONAL
https://doi.org/10.5281/zenodo.23302376 **VOL. 48, NO. 1, 28-45, 2027**

ROBUST SELECTION OF GENERATIVE AGROCLIMATIC MODELS UNDER INCONCLUSIVE EVIDENCE: MULTI-SEED VAE-LSTM VALIDATION AND NACH SYNTHESIS

Johnatan Corrales-Bonilla[1], Juan Salazar-Arias, Paco Vásquez-Carrera, Yoandrys Morales-Tamayo, Christian Corrales-Otáñez, Sebastián Borja-Urbano
Universidad Técnica de Cotopaxi, La Maná, Ecuador.

ABSTRACT
Choosing among generative models for agroclimatic data is rarely settled by a single comparison: contrasts point in different directions, some are inconclusive, and the choice must still be made. This paper treats that situation as the problem itself. A VAE-LSTM generator for the canton of La Maná, Ecuador, is evaluated against three references — a deterministic LSTM autoencoder, a conditioned bootstrap and a convolutional TimeVAE — over fifteen independent seeds under a strictly chronological design (train 2020–2022, validation 2023, test 2024), with the test year evaluated only once, after every modelling decision had been fixed on validation. Four distributional metrics are reported for every model and variable. The evidence does not converge: the conditioned bootstrap attains the smallest Wasserstein distance on three of four variables, the VAE-LSTM separates typified scenarios with large effect sizes, and a purpose-built within-scenario realism test shows that this separation is not accompanied by faithful within-scenario dispersion. A Neutrosophic Analysis of Competing Hypotheses turns that mixture into an auditable verdict: every entry of the truth–indeterminacy–falsity matrix is derived from the reported contrasts by a two-stage calibration written out in full, and the ranking is examined separately under four aggregation rules. The reading that the model's potential utility is capability-dependent leads under all four rules and in 93–100 % of resampled block weightings, but within profiles dominated by indeterminacy and with the realism block counting against it; restricted to the one evidence block all three readings share, it is statistically indistinguishable from the no-detectable-difference reading. The contribution is a reproducible procedure for deciding between generative models when the evidence stays partly inconclusive.
KEYWORDS: model selection under uncertainty; deep generative models; multi-seed validation; distributional metrics; neutrosophic analysis of competing hypotheses; agroclimatic data.
MSC: 62P12, 68T07, 62F03.

RESUMEN
Elegir entre modelos generativos de datos agroclimáticos rara vez se resuelve con una sola comparación: los contrastes apuntan en direcciones distintas, algunos resultan inconcluyentes y aun así hay que decidir. Este trabajo toma esa situación como el problema mismo. Un generador VAE-LSTM para el cantón La Maná, Ecuador, se evalúa frente a tres referencias —un autocodificador LSTM determinista, un bootstrap condicionado y un TimeVAE convolucional— sobre quince semillas independientes bajo un diseño estrictamente cronológico, con el año de prueba evaluado una sola vez y después de fijar toda decisión de modelado en validación. Se reportan cuatro métricas distribucionales para cada modelo y variable. La evidencia no converge: el bootstrap condicionado logra la menor distancia de Wasserstein en tres de las cuatro variables, el VAE-LSTM separa los escenarios tipificados con tamaños de efecto grandes, y una prueba de realismo intra-escenario construida a tal fin muestra que esa separación no viene acompañada de una dispersión interna fiel. Un Análisis Neutrosófico de Hipótesis en Competencia convierte esa mezcla en un veredicto auditable: cada entrada de la matriz verdad–indeterminación–falsedad se deriva de los contrastes reportados mediante una calibración en dos etapas escrita íntegramente —una estadística y otra interpretativa, separadas—, y el orden se examina por separado bajo cuatro reglas de agregación. La lectura según la cual la utilidad potencial del modelo depende de la capacidad encabeza el caso base bajo las cuatro reglas, pero dentro de perfiles dominados por la indeterminación, con el bloque de realismo en contra, y resulta indistinguible de la lectura de ausencia de diferencia detectable si la comparación se restringe al único bloque de evidencia que las tres lecturas comparten. La contribución es un procedimiento reproducible para decidir entre modelos generativos cuando la evidencia permanece en parte inconcluyente.
PALABRAS CLAVE: selección de modelos bajo incertidumbre; modelos generativos profundos; validación multisemilla; métricas distribucionales; análisis neutrosófico de hipótesis en competencia; datos agroclimáticos.

[1]Corresponding author: johnatan.corrales5518@utc.edu.ec.

1. INTRODUCTION

Continuous agricultural monitoring in the humid Andean tropics depends on climate series that are rarely continuous. In the canton of La Maná, Ecuador, cloud cover and rainfall interrupt the flights of the multispectral platform that supplies NDVI, NDRE and GNDVI, and the agroclimatic variables that condition those indices — air temperature, specific humidity, corrected precipitation and wind speed — are themselves recorded with gaps. Deep generative models are an obvious candidate for filling that record, and variational autoencoders over recurrent architectures have produced encouraging results on climate series [15, 20, 30]. Synthetic generation is by now an established line in agriculture, from image augmentation [2, 18] to tabular and time-series synthesis [3, 17, 33], with general-purpose frameworks available for the latter [28]; probabilistic machine-learning forecasts of weather [32] and of ENSO [10] have followed the same trajectory.

The difficulty this paper addresses is not building such a generator. It is deciding whether to use it. Generative models are usually compared on a single distributional distance, over a single training run, against a single reference. When the comparison is repeated over independent seeds, against several references, on several variables and on more than one criterion, the result is almost never a clean ordering: some contrasts favour the model, some favour a much simpler baseline, and many are inconclusive at the resolution the experiment can reach. The decision still has to be taken, and taking it by picking the favourable subset is not a method.

We therefore treat robust model selection under partly inconclusive evidence as the object of the study, and use the agroclimatic generator as the empirical case. Three readings of the evidence are set in competition. H1: the VAE-LSTM is distributionally superior to the references. H2: no distributional difference between the VAE-LSTM and the references is detectable at the resolution of this experiment. H3: the potential utility of the VAE-LSTM is capability-dependent — its distinctive value, if any, lies in latent-space scenario generation rather than in marginal distributional fidelity. H2 is deliberately a statement about detectability and not about equivalence: an equivalence claim requires an equivalence test, and where one is performed it is reported as such.

The contributions are three. First, a multi-seed evaluation of four generators under a strictly chronological design in which every modelling decision is fixed on a validation year that precedes the test year, with four distributional metrics reported in full rather than announced and omitted. Second, a within-scenario realism test that distinguishes a generator whose typified scenarios are merely far apart from one whose scenarios also reproduce the dispersion of the real conditioned data — a distinction that separation statistics alone cannot make. Third, a Neutrosophic Analysis of Competing Hypotheses [24], which extends Heuer's Analysis of Competing Hypotheses [12] by admitting degrees of truth, indeterminacy and falsity in place of a forced consistent/inconsistent verdict [25], whose calibration is written out in full so that every entry of the resulting matrix is derivable from the reported contrasts, and whose ranking is examined separately under each aggregation rule.

All results reported below were computed for this article from the public NASA POWER record [34] under a strictly chronological design, in which the validation period precedes the test period.

2. DATA AND EXPERIMENTAL DESIGN

2.1 Study area and data

The canton of La Maná (lat. −0.95°, lon. −79.24°, 220 m a.s.l.) lies in Cotopaxi province, Ecuador. Its humid Andean tropical climate is characterized by a mean temperature of 22–26 °C, annual precipitation above 1800 mm with a rainy season from November to May, and relative humidity of 78–92 % [14]. Daily records were obtained from NASA POWER [34] for two windows with distinct purposes. The modelling window, 1 January 2020 to 31 December 2024, supplies T2M (°C), QV2M (g/kg), PRECTOTCORR (mm/d) and WS2M (m/s), together with T2M_MAX and T2M_MIN, which the reference evapotranspiration calculation requires. The climatological window, 1 January 1991 to 31 December 2020, supplies the thirty-year precipitation record against which the standardized precipitation index is calibrated. Both windows passed a physical-range check with no values rejected and no interpolation needed. NASA POWER has been evaluated against station records for tropical precipitation and temperature extremes [36]; that

evaluation is external to the present study, which compares generated series against NASA POWER and not NASA POWER against independent observations.

2.2 Chronological design and control of temporal leakage

Model selection on time series requires that the periods used to fit, to select and to evaluate follow one another in time. The design set out in Table 1 satisfies train < validation < test, and the test year was read once, after every modelling decision had been fixed.

Role	Period	Days	14-day windows	What it controls
Train	2020-01-01 – 2022-12-31	1096	1083	min–max scaler fitting; weight updates
Validation	2023-01-01 – 2023-12-31	365	352	early stopping; latent dimension; architecture
Test	2024-01-01 – 2024-12-31	366	353	reported metrics only, evaluated once

Table 1. Chronological design. Every decision is taken on a period strictly earlier than the one on which the result is reported.

The scaler is fitted on the training period alone; validation and test windows are transformed with the training minima and maxima and never contribute to them. Early stopping monitors the validation loss with a patience of 15 epochs and restores the best validation state. The latent dimension and the architecture are chosen by validation loss. No quantity computed on the test year enters any of these decisions. The held-out test year was accessed only after all modelling decisions had been fixed; the fifteen pre-specified trained replicas were then evaluated on that same held-out period, so the fifteen seeds are replicas of one evaluation and not fifteen independent looks at the test data.

2.3 Models

Four generators are compared. The VAE-LSTM is the architecture under study: a recurrent encoder of two LSTM layers (64 then 32 units) over a 14-day window of the four variables, two dense heads for the mean and log-variance of a d-dimensional latent, and a mirrored decoder (32 then 64 units) with a sigmoid output, trained on reconstruction error with a Kullback–Leibler term weighted by 10^{-3}. The LSTM autoencoder is the same network without the KL term and without sampling, and therefore isolates the contribution of the variational component. The conditioned bootstrap resamples 14-day training windows with replacement and carries no learned parameters; it is the reference a generative model must beat to justify its cost. The TimeVAE baseline [7] is a convolutional variational autoencoder over the same window, included so that the comparison is not confined to recurrent architectures. Under the chosen latent dimension $d = 8$, training stopped after 105 ± 34 epochs for the VAE-LSTM, 29 ± 9 for the autoencoder and 75 ± 12 for TimeVAE. All four were implemented in PyTorch [31] and trained on CPU.

2.4 Distributional metrics

Four metrics are computed for every model, variable and seed on the test year, and all four are reported. The 1-Wasserstein distance [38] is the primary metric and orders the models throughout; it measures the optimal transport cost between the synthetic and the real empirical distributions and is expressed in the units of each variable. The maximum mean discrepancy [9] uses a Gaussian kernel with the median heuristic for the bandwidth. The continuous ranked probability score [8] evaluates the synthetic ensemble against each real observation and is averaged over them; its behaviour on extremes has been examined separately [35]. The Kolmogorov–Smirnov statistic reports the largest gap between the two empirical distribution functions. The three secondary metrics are reported in full in Table 6 rather than described as following the same pattern.

2.5 Statistical protocol

Common random numbers make the seed a matched block. For seed k the same value initializes the network weights, the order in which mini-batches are drawn, the latent draws used at generation and the indices of the evaluation subsample. The four models therefore see, under seed k, the identical temporal dataset and the identical evaluation sample, and differ only in the mechanism under study. That is what licenses the paired Wilcoxon signed-rank test across seeds; a seed index shared by unrelated stochastic streams would not. Effect size is reported as the paired rank-

biserial correlation. The twelve contrasts of the model comparison are corrected together by the Holm procedure [13], and the three latent-dimension contrasts likewise.

Non-significance is not read as equivalence anywhere in this paper. Where an equivalence claim is at stake, two one-sided tests are run [22] against a region of practical equivalence declared in advance: ± 10 % of the mean validation loss for the latent dimension, and a factor of two in either direction — that is, $|log_2 \text{ ratio}| \leq 1$ — for the within-scenario variance ratio. A non-significant contrast with a large effect estimate is reported as compatible with insufficient precision to discriminate the hypotheses, which is weaker than a claim about power.

2.6 Environmental indices

Reference evapotranspiration is computed by Hargreaves–Samani [11] from the daily record, as in Eq. (1):

$$ET_0 = 0.0023 \cdot Ra \cdot (T_{\text{mean}} + 17.8) \cdot \sqrt{T_{\max} - T_{\min}} \tag{1}$$

with extraterrestrial radiation Ra obtained per day of year for the site latitude by the standard FAO-56 formulation [5]. The observed daily thermal amplitude at La Maná averages 9.43 ± 2.56 °C. The resulting ET_0 averages 4.177 ± 0.667 mm d^{-1} over the modelling window, and the water balance $P - ET_0$ averages −0.360 mm d^{-1}. Machine-learning estimators of reference evapotranspiration have been proposed for settings where more meteorological inputs are available [1, 16]; the temperature-based formulation is used here because it needs only the variables the available series supply.

The standardized precipitation index is calibrated on a separate thirty-year climatological reference dataset used exclusively for SPI calibration: the years 1991–2020 at the same site and for the same variable. It is not a second test set, it is never used to fit, select or evaluate any generative model, and no quantity derived from it enters the model comparison. Monthly precipitation totals are fitted by a gamma distribution separately for each of the twelve calendar months, because the annual cycle at this site is pronounced and a single fit would ignore it. Zero-precipitation months are handled by the mixed distribution of Eq. (2), with q the observed proportion of dry months [37]:

$$H(x) = q + (1 - q)G(x) \tag{2}$$

At this site $q = 0$ in all twelve months. The gamma shape and scale are estimated by maximum likelihood on the positive values, and the index is obtained by the normal quantile transformation of Eq. (3) [26]:

$$SPI = \Phi^{-1}\big(H(x)\big) \tag{3}$$

Goodness of fit is reported in Table 11.

2.7 Calibration of the neutrosophic synthesis

The synthesis assigns to each pair (evidence block, hypothesis) a triple $\langle T, I, F \rangle$ of degrees of truth, indeterminacy and falsity. It is built in two stages that are kept strictly apart, because they answer different questions. Stage A asks what each statistical contrast showed, and is carried out without reference to any hypothesis. Stage B asks what that outcome means for each competing hypothesis, and is carried out by a mapping that is tabulated in full. Separating the two is what makes every entry of the final matrix mechanically derivable from the reported statistics, and what makes visible which part of an entry comes from the data and which from an interpretive assumption. The rules of both stages were fixed before the final ranking was computed and applied uniformly to the three hypotheses; no preregistration is claimed.

The framework itself is not introduced here. The Analysis of Competing Hypotheses is due to Heuer [12]; its neutrosophic extension, which replaces the forced consistent/inconsistent verdict by independent degrees of truth, indeterminacy and falsity, is due to Leyva Vázquez and Smarandache [24], resting on the treatment of evidence degrees and ontological indeterminacy of [25]. What is new here is the empirical calibration of that framework: converting repeated statistical results into explicit evidence states, keeping demonstrated equivalence distinct from inconclusive evidence, mapping those states to $\langle T, I, F \rangle$ under explicitly competing hypotheses, making every matrix entry reproducible from the reported tables, and examining rank robustness across alternative aggregation rules instead of a single one.

Stage A: statistical evidence state.

Each block E_i is a set of n_i elementary contrasts, and each contrast is classified into exactly one of five states by a single rule applied identically to all five blocks. A contrast is conclusive only if it is statistically significant after multiplicity correction and its effect reaches the minimum relevant effect declared for its effect-size family; statistical significance alone is not sufficient, and that requirement is part of the rule rather than an exception introduced for any particular block. Demonstrated equivalence is a separate state and can arise only from an equivalence test against a declared region, never from non-significance.

Order	Condition	State	Reading
1	equivalence test performed and significant	EQ	equivalence formally demonstrated
2	significant after Holm and $\lvert\text{effect}\rvert \geq \text{MRE}$, supportive direction	D+	conclusive difference in the supportive direction
3	significant after Holm and $\lvert\text{effect}\rvert \geq \text{MRE}$, opposite direction	D−	conclusive difference in the opposite direction
4	otherwise, effect $\geq \tau$	IND+	inconclusive, comparatively large effect estimate
5	otherwise, effect $< \tau$	IND−	inconclusive, comparatively small effect estimate

Table 2. Stage A: the classification rule for an elementary contrast. It is evaluated in the order shown and is the same for the five evidence blocks. No hypothesis enters at this stage.

The minimum relevant effect is a declared constant of each effect-size family, set at the conventional medium benchmark: 0.30 for the paired rank-biserial correlation of E1 and E5, 0.06 for the η^2 of E2, 0.30 for the absolute difference in correlation of E4, and, for E3, the half-width $\lvert log_2 \text{ ratio}\rvert = 1$ of the declared region of practical equivalence, that is, a factor of two in dispersion. Each block also declares which sign is supportive. In E1 the supportive direction is a difference favouring the VAE-LSTM, and in E2 a significant separation of the scenarios. In E3 and E4 both signs are failures — over-dispersion and collapse, positive and negative distortion of a correlation — so both signs are sent to D−; that is a declared property of those two blocks, stated here and not a departure from the rule. The threshold τ only distinguishes the two indeterminate labels and is descriptive: Section 3.7 shows that it cannot change any degree, index or rank.

Stage B: hypothesis-specific interpretation.

Only now do the hypotheses enter. The three readings in competition are H1, that the VAE-LSTM is distributionally superior to the reference models; H2, that no distributional difference between the VAE-LSTM and the reference models is detectable at the resolution of this experiment; and H3, that the potential utility of the VAE-LSTM is capability-dependent, its distinctive value, if any, lying in latent-space scenario generation rather than in marginal distributional fidelity. The mapping $m(e, h, z) \rightarrow \{T, I, F\}$ is given in full in Table 3, with no entry left to be inferred.

Three features of Table 3 are modelling decisions and not consequences of the data, and are stated as such. First, E1 is the only block that compares the models with one another, so it is the only block that can adjudicate H1 or H2: a scenario result carries no information about whether the VAE-LSTM beats a bootstrap on marginal distance. Second, E4 measures how faithfully the VAE-LSTM reproduces the real dependence structure in absolute terms; the same dependence-preservation quantity was not computed for the LSTM autoencoder, the conditioned bootstrap and TimeVAE, so this block cannot support a claim of relative superiority and is mapped to indeterminacy under all three hypotheses. It is retained in the audit because making it comparative is the natural extension, and because concealing a block that was measured would defeat the purpose of the procedure. Third, E5 is an internal ablation of the model under study and informs robustness rather than the ranking, so it too maps to indeterminacy throughout. The consequence is an asymmetry that Table 3 makes visible and Section 3.7 quantifies: H1 and H2 are adjudicated by the twelve contrasts of E1, H3 by the twenty-eight contrasts of E1, E2 and E3.

Block	State	H1	H2	H3
E1 (marginal distance)	D+	T	F	F
	D−	F	F	T
	EQ	F	T	T
	IND+	I	T	I
	IND−	I	T	I
E2 (scenario separation)	D+	I	I	T
	D−	I	I	F
	EQ	I	I	F
	IND+	I	I	I
	IND−	I	I	I
E3 (within-scenario realism)	D+	I	I	T
	D−	I	I	F
	EQ	I	I	T
	IND+	I	I	I
	IND−	I	I	I
E4 (correlation preservation)	D+	I	I	I
	D−	I	I	I
	EQ	I	I	I
	IND+	I	I	I
	IND−	I	I	I
E5 (latent-dimension ablation)	D+	I	I	I
	D−	I	I	I
	EQ	I	I	I
	IND+	I	I	I
	IND−	I	I	I

Table 3. Stage B: the complete mapping from statistical state to degree of truth (T), indeterminacy (I) or falsity (F), for every evidence block and every hypothesis. This table is the whole interpretive content of the synthesis**.**

Degrees, aggregation and profile distance.

For a block of n contrasts, the degree of truth of a hypothesis is the proportion of contrasts whose Stage A state maps to T under that hypothesis, as in Eq. (4):

$$T(e,h) = \frac{1}{n}\sum_{c=1}^{n} 1[m(e,h,z_c) = T] \tag{4}$$

where $1[m(e,h,z_c) = T]$ equals 1 when the Stage A state of contrast c maps to T and 0 otherwise; $I(e,h)$ and $F(e,h)$ are defined in the same way with I and F in place of T. The three degrees sum to one by construction, so no normalization step is required and none is applied. Blocks are then aggregated with weights w_i, uniform in the base case ($w_i = 1$ for all five), as in Eq. (5):

$$\overline{T} = \frac{\sum_{i=1}^{k} w_i T_i}{\sum_{i=1}^{k} w_i} \tag{5}$$

Uniform weighting is a declared convention, not an objective weighting, and the sensitivity analysis is what tests whether it matters. Four aggregation rules are computed and reported separately, because their ranges and meanings differ and their raw values are not comparable across rules. The first two are given by Eq. (6):

$$R = \overline{T}(1-\overline{I})(1-\overline{F}), \quad S = \frac{2+\overline{T}-\overline{I}-\overline{F}}{3} \tag{6}$$

The remaining two are the net truth $T - F$ and the indeterminacy-discounted truth $T(1 - I)$, computed on the same weighted means. The separation between two hypothesis profiles is reported as the normalized profile distance of Eq. (7), which is a distance between triples and carries no probabilistic or confidence interpretation:

$$d(h,h') = \frac{|\Delta T| + |\Delta I| + |\Delta F|}{3} \tag{7}$$

Worked example.

Block E1 contains 12 contrasts. Stage A classifies 2 as D+, 5 as D−, 0 as EQ and 5 as indeterminate, which is the E1 row of Table 12. Row E1 of Table 3 maps, for H1, D+ to T, D− and EQ to F, and both indeterminate states to I. Equation (4) therefore gives $T = \frac{2}{12} = 0.167$, $F = \frac{(5+0)}{12} = 0.417$ and $I = \frac{5}{12} = 0.417$, so the E1 entry of the H1 column of Table 13 reads $\langle 0.167, 0.417, 0.417 \rangle$. Every other entry of that matrix is obtained the same way, from the counts of Table 12 and the mapping of Table 3.

The choices that remain with the analyst are listed here rather than denied: the statement of the three hypotheses; the partition of the evidence into five blocks; the mapping of Table 3; the minimum relevant effects; the equivalence regions; the block weights; and the set of aggregation rules. This procedure does not eliminate subjectivity. What it does is make each such choice explicit and separable, and submit the mathematically active ones to sensitivity analysis over the space set out in Table 4.

Parameter	Base case	Alternatives sampled	Range	Role
Block weights w_i	1 for the five blocks	flat Dirichlet, rescaled to sum 5	simplex	relative evidential importance of the blocks
Aggregation rule	all four reported	R, S, net truth, discounted truth	own scale each	how the three degrees are condensed into one number
Minimum relevant effect	medium benchmark per family	not varied	fixed	part of the Stage A rule; declared, not fitted
τ (IND+ / IND− label)	declared per family	not varied	fixed	descriptive label only; shown inactive in Section 3.7
Incompleteness penalty	removed	not applicable	—	no longer needed once E3 measures within-scenario realism directly

Table 4. The sensitivity space. Only the mathematically active parameters are sampled; the remaining rows record what was held fixed and why.

The Monte Carlo experiment draws 20,000 weight vectors from the flat Dirichlet distribution with seed 20260904, evaluates the four rules on each draw and credits the leading hypothesis under each; where two or more hypotheses lead within 10^{-12}, the credit is split equally and the proportion of such ties is reported. No penalty transferring mass from truth and falsity to indeterminacy is applied when a block's own validation is incomplete: within-scenario realism is measured directly and block E3 carries that evidence, so such an axis would have no justification in a sensitivity analysis.

3. RESULTS

3.1 Latent dimension

Validation loss over fifteen seeds was 0.02541 for $d = 4$, 0.02527 for $d = 8$ and 0.02681 for $d = 16$; $d = 8$ was selected on that basis and used throughout. Figure 1 shows the distribution of epochs to early stopping and of validation loss by latent dimension.

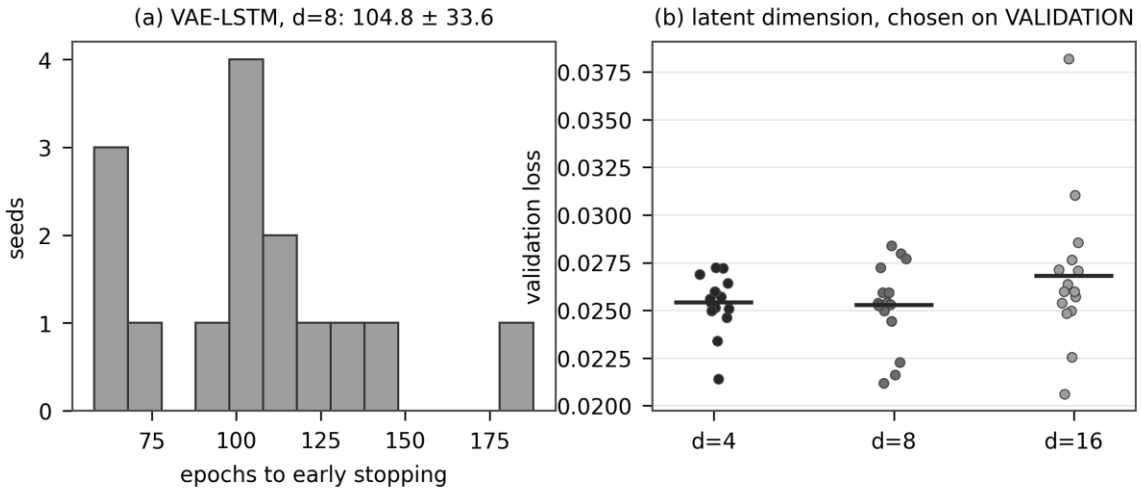


Figure 1. (a) Epochs to early stopping for the VAE-LSTM at the selected latent dimension, over fifteen independent seeds. (b) Validation loss by latent dimension; horizontal bars are means. Both panels use the validation year only.

No statistically detectable differences were found among the tested latent dimensions at the current experimental resolution: none of the three Holm-corrected paired contrasts of Table 5 reaches significance. That is not equivalence, and the two-one-sided-test column separates the two readings. Against a region of practical equivalence of ± 10 % of the mean validation loss, equivalence is established for $d = 4$ vs. $d = 8$ alone; the two comparisons involving $d = 16$ are neither significantly different nor demonstrably equivalent, and remain indeterminate.

Comparison	p (Holm)	rank-biserial r	mean difference	ROPE δ	TOST p (Holm)	Equivalence
d=4 vs d=8	0.804	+0.083	+0.00014	±0.00258	0.005	yes
d=4 vs d=16	0.779	−0.267	−0.00140	±0.00258	0.290	no
d=8 vs d=16	0.688	−0.367	−0.00154	±0.00258	0.290	no

Table 5. Latent dimension: paired contrasts on validation loss and equivalence tests. Non-significance and equivalence are reported separately**.**

3.2 Marginal distributional comparison

Table 6 reports all four metrics for every model and variable as mean ± standard deviation over the fifteen seeds. Figure 2 displays the primary metric.

Model	Variable	Wasserstein	MMD	CRPS	KS
VAE-LSTM	T2M	1.2225 ± 0.0469	0.6625 ± 0.0441	1.0143 ± 0.0518	0.6000 ± 0.0268
VAE-LSTM	QV2M	0.5979 ± 0.0621	0.2665 ± 0.0307	1.1392 ± 0.0323	0.2862 ± 0.0325
VAE-LSTM	PRECTOTCORR	1.7810 ± 0.3543	0.3349 ± 0.0872	2.2575 ± 0.2213	0.4113 ± 0.0732
VAE-LSTM	WS2M	0.1363 ± 0.0232	0.3423 ± 0.0668	0.2145 ± 0.0069	0.2930 ± 0.0416
LSTM-AE	T2M	1.2030 ± 0.0483	0.9166 ± 0.0199	1.1914 ± 0.0621	0.8952 ± 0.0127
LSTM-AE	QV2M	1.6478 ± 0.0046	0.8505 ± 0.0088	1.6335 ± 0.0336	0.5147 ± 0.0052
LSTM-AE	PRECTOTCORR	2.7898 ± 2.8163	0.6667 ± 0.0514	2.7022 ± 1.1006	0.6026 ± 0.0824
LSTM-AE	WS2M	0.2676 ± 0.0038	0.7707 ± 0.0095	0.2728 ± 0.0079	0.5081 ± 0.0191
Bootstrap	T2M	1.2043 ± 0.0469	0.5619 ± 0.0343	0.9270 ± 0.0530	0.5298 ± 0.0174
Bootstrap	QV2M	0.4246 ± 0.0302	0.2010 ± 0.0230	1.1016 ± 0.0228	0.1972 ± 0.0088
Bootstrap	PRECTOTCORR	1.1688 ± 0.2906	0.0581 ± 0.0212	1.9432 ± 0.1506	0.1070 ± 0.0104
Bootstrap	WS2M	0.0448 ± 0.0120	0.0596 ± 0.0172	0.1925 ± 0.0047	0.0586 ± 0.0127
TimeVAE	T2M	1.2097 ± 0.0481	0.6283 ± 0.0406	0.9751 ± 0.0532	0.5914 ± 0.0222
TimeVAE	QV2M	0.6107 ± 0.0324	0.3018 ± 0.0257	1.1443 ± 0.0268	0.2565 ± 0.0133
TimeVAE	PRECTOTCORR	1.4442 ± 0.1927	0.2750 ± 0.0428	2.1274 ± 0.1520	0.2832 ± 0.0232
TimeVAE	WS2M	0.0919 ± 0.0084	0.2126 ± 0.0291	0.2007 ± 0.0060	0.1665 ± 0.0139

Table 6. Distributional metrics on the test year (mean ± SD over fifteen seeds). Wasserstein is the primary metric; the other three are reported in full.

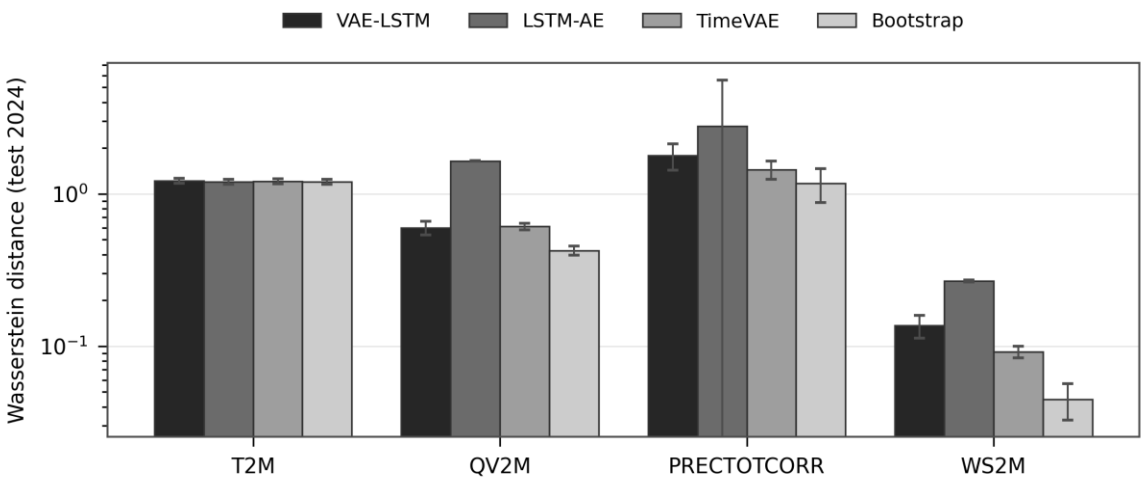


Figure 2. Wasserstein distance to the real test distribution by model and variable, mean ± SD over fifteen seeds. Logarithmic vertical axis.

Evidence block E1, on which the comparison between models rests, is built on the primary metric alone: the twelve paired contrasts below are computed on the 1-Wasserstein distance and on nothing else. The three secondary metrics are reported in full in Table 6 and are read as corroborating context; they do not enter the neutrosophic synthesis, and the count of conclusive contrasts is therefore not inflated by treating four correlated metrics as four independent pieces of evidence.

The paired comparison of Table 7 does not favour the model under study. Of the twelve contrasts, two are conclusive in favour of the VAE-LSTM, both of them against the deterministic autoencoder (QV2M, WS2M), which confirms that the variational component contributes; five are conclusive against it, three against the conditioned bootstrap and two against TimeVAE; and five do not reach significance after Holm correction. The direction of the significant results is the finding that matters: on marginal distributional distance a resampling procedure with no learned parameters outperforms the generative model on three of the four variables, and a convolutional variational baseline outperforms it on two.

Reference	Variable	p (raw)	p (Holm)	rank-biserial r	Mean difference	Seeds won by VAE	Favours
LSTM-AE	T2M	0.4543	1.0000	+0.233	+0.0194	7/15	base
LSTM-AE	QV2M	0.0001	0.0007	−1.000	−1.0500	15/15	VAE
LSTM-AE	PRECTOTCORR	0.1354	0.6769	−0.450	−1.0088	9/15	VAE
LSTM-AE	WS2M	0.0001	0.0007	−1.000	−0.1312	15/15	VAE
Bootstrap	T2M	0.4212	1.0000	+0.250	+0.0182	6/15	base
Bootstrap	QV2M	0.0001	0.0007	+1.000	+0.1733	0/15	base
Bootstrap	PRECTOTCORR	0.0012	0.0081	+0.883	+0.6123	1/15	base
Bootstrap	WS2M	0.0001	0.0007	+1.000	+0.0915	0/15	base
TimeVAE	T2M	0.4887	1.0000	+0.217	+0.0128	6/15	base
TimeVAE	QV2M	0.2524	1.0000	−0.350	−0.0128	11/15	VAE
TimeVAE	PRECTOTCORR	0.0026	0.0157	+0.833	+0.3368	2/15	base
TimeVAE	WS2M	0.0001	0.0007	+1.000	+0.0445	0/15	base

Table 7. Paired Wilcoxon signed-rank contrasts on the Wasserstein distance, VAE-LSTM against each reference, over fifteen matched seeds. Holm correction across the twelve contrasts.

3.3 Correlation structure

A synthetic series may match every marginal distribution and still misrepresent how the variables move together. The comparison in Table 8 includes the uncertainty of the real coefficient rather than treating it as a fixed number: both the real and the synthetic coefficients are resampled, the real one by a moving-block bootstrap with blocks of thirty days appropriate for daily series [21], and the difference Δr is reported with its interval.

Pair	r real	95 % CI (real)	r synthetic	SD	Δr	95 % CI (Δr)	Excludes 0
T2M-QV2M	−0.172	[−0.437, 0.172]	−0.521	0.059	−0.344	[−0.700, −0.073]	yes
QV2M-PRECTOTCORR	+0.448	[0.336, 0.565]	+0.850	0.132	+0.394	[+0.043, +0.585]	yes
QV2M-WS2M	−0.048	[−0.308, 0.293]	−0.355	0.099	−0.317	[−0.713, +0.003]	no
PRECTOTCORR-WS2M	+0.057	[−0.102, 0.192]	−0.220	0.103	−0.256	[−0.477, +0.019]	no

Table 8. Difference between synthetic and real cross-correlation, with the uncertainty of both coefficients included. Real coefficients resampled by moving-block bootstrap (block = 30 days, 2000 replicates).

Two of the four pairs show a difference whose interval excludes zero: the temperature–humidity association is reproduced far more strongly negative than it is in the data ($\Delta r = -0.344$), and the humidity–precipitation association far more strongly positive ($\Delta r = +0.394$). This is consistent with a structural limitation of the present reconstruction objective, which rewards fidelity to each marginal and to the mean joint tendency rather than to the full dependency structure; the design of this study does not establish that mechanism, and no causal attribution is made. Figure 3 shows the two correlation matrices and their absolute difference, where the two affected cells stand out against a background of small discrepancies.

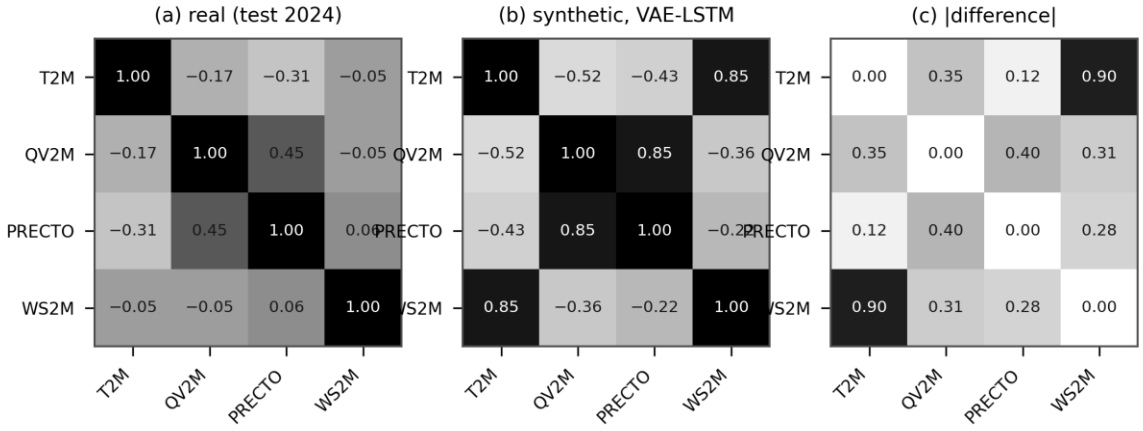


Figure 3. Cross-correlation matrices: real test year, synthetic mean over fifteen seeds, and the absolute difference between them.

3.4 Separation between typified scenarios

Three scenarios — drought, heat stress and heavy rain — are generated from the latent regions occupied by training windows meeting thresholds declared on the training period: the 20th percentile of the 14-day mean precipitation for drought, the 80th percentile of the 14-day mean temperature for heat stress, and the 80th percentile of precipitation for heavy rain. The same thresholds label the real test windows.

Variable	F (mean ± SD)	η^2 (mean ± SD)	ω^2 (mean)	Seeds with $p <$ .05	Levene violated	Welch-type $p <$ .05
PRECTOTCORR	932.6 ± 228.7	0.602 ± 0.052	0.601	100 %	100 %	100 %
QV2M	968.6 ± 206.5	0.612 ± 0.049	0.611	100 %	100 %	100 %
T2M	191.1 ± 53.0	0.239 ± 0.050	0.237	100 %	100 %	100 %
WS2M	99.3 ± 59.8	0.137 ± 0.067	0.135	100 %	93 %	100 %

Table 9. Separation between the three generated scenarios, by variable. Effect sizes accompany the F statistic, and the homogeneity assumption is checked rather than assumed.

The scenarios separate, as Table 9 reports: the one-way test is significant in every seed for all four variables, with η^2 of 0.60 for precipitation and 0.61 for humidity, both large by the conventional benchmark, 0.24 for temperature and 0.14 for wind speed, the last of which falls just below it. The homogeneity-of-variance assumption fails in essentially every seed, so the ordinary F statistic cannot be read at face value; a heteroscedasticity-robust Alexander–Govern test [4] remains significant in 100 % of seeds for all four variables, which is what supports the conclusion. Figure 4 shows

the per-seed scenario means, whose three clusters are visibly disjoint for precipitation and humidity and overlap for wind speed.

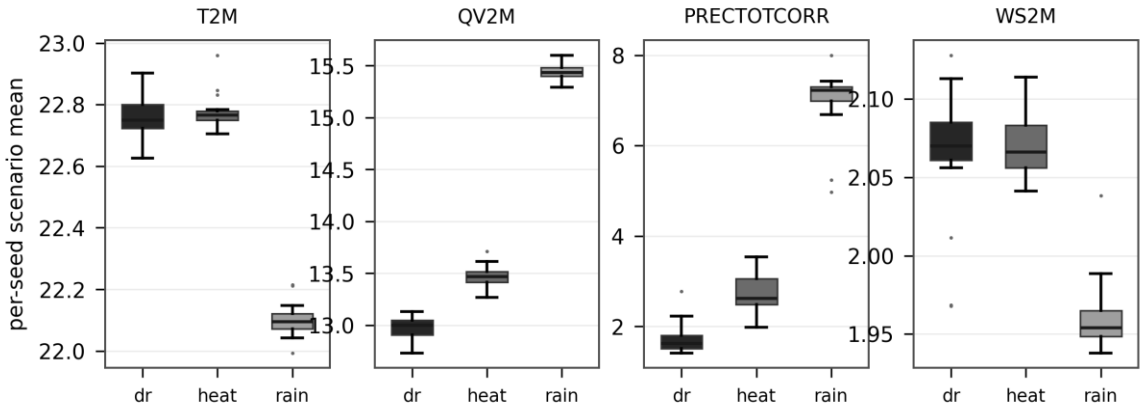


Figure 4. Typified scenarios: distribution across fifteen seeds of the per-seed scenario mean, by variable.

3.5 Realism within each scenario

A large separation statistic is compatible with mode collapse: three tight, well-separated clusters would produce it while representing none of the three regimes faithfully. The test reported in Table 10 closes that gap by comparing each generated scenario against the real test windows meeting the same threshold, variable by variable, on the same distributional metrics used in the model comparison plus a variance ratio, a coverage fraction and a relative range. Figure 5 displays the variance ratios and the coverage against the declared region of practical equivalence.

Scenario	Variable	Wasserstein	MMD	Var ratio	Coverage	Relative range	n real
drought	PRECTOTCORR	1.280	0.914	37.98	0.10	9.87	91
drought	QV2M	0.546	0.508	1.07	0.94	1.18	91
drought	T2M	1.202	0.660	0.85	0.55	0.87	91
drought	WS2M	0.152	0.480	0.46	1.00	0.59	91
heat stress	PRECTOTCORR	1.056	0.392	1.40	0.99	0.86	294
heat stress	QV2M	1.119	0.416	0.39	1.00	0.61	294
heat stress	T2M	0.954	0.553	1.98	0.45	0.98	294
heat stress	WS2M	0.086	0.305	0.49	1.00	0.57	294
heavy rain	PRECTOTCORR	1.364	0.425	4.62	0.53	3.62	10
heavy rain	QV2M	1.113	0.804	2.61	0.03	2.72	10
heavy rain	T2M	1.197	0.798	0.85	0.11	1.18	10
heavy rain	WS2M	0.097	0.437	5.30	0.51	2.91	10

Table 10. Within-scenario realism: generated scenario against the real threshold-conditioned distribution, averaged over fifteen seeds. Var ratio is Var(synthetic)/Var(real); coverage is the fraction of generated values inside the real range.

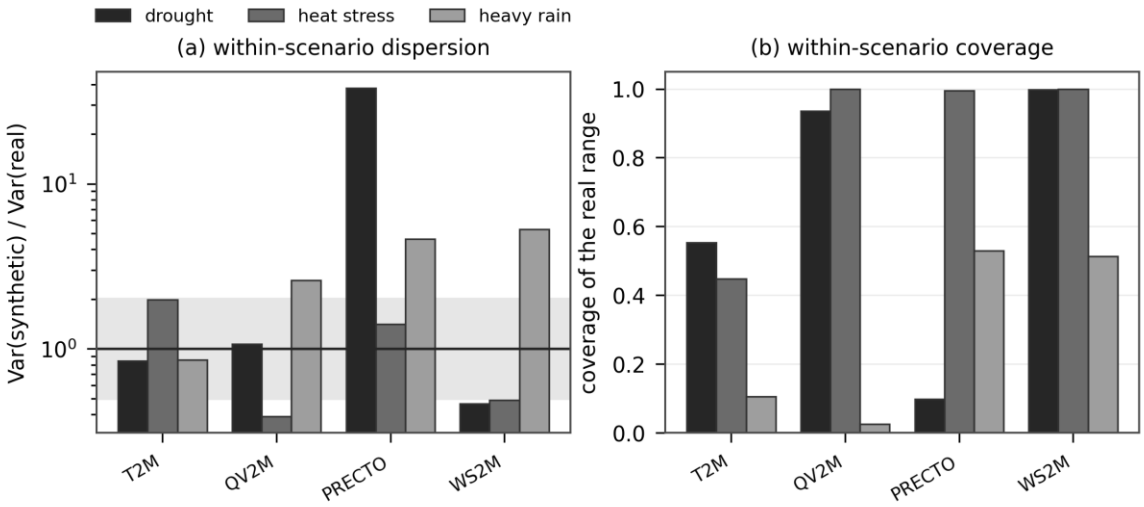


Figure 5. Within-scenario dispersion and coverage. The shaded band in (a) is the declared region of practical equivalence, a factor of two in either direction around a ratio of one.

The result does not support an unqualified scenario-generation advantage. Of the twelve scenario-by-variable cells, four establish equivalence of dispersion against the declared region, seven are conclusively outside it and one is inconclusive. The failures are of both kinds. Three cells show variance ratios below one half — wind speed under drought and under heat stress, and humidity under heat stress — the signature of collapse towards the scenario centre. At the other extreme, precipitation under drought is generated with 38 times the real variance while covering only 10 % of the real range, and the heavy-rain scenario, which rests on only 10 real windows, is over-dispersed on three variables and covers little of the observed range. Scenario separation and within-scenario fidelity are therefore not the same property, and only the first is achieved here.

3.6 Environmental indices

Reference evapotranspiration computed from the daily thermal amplitude averages 4.177 ± 0.667 mm d^{-1}. A fixed amplitude of 1.59 °C for every day would not be appropriate here: the observed daily amplitude at this site averages 9.43 °C; because ET_0 scales with the square root of the amplitude, such a fixed value would understate ET_0 by a factor of 2.41. All ET_0 and $P - ET_0$ values reported here use the daily maxima and minima.

Month	n	Zero months	Shape a	Scale	KS D	Asymptotic p	Bootstrap p	Mean (mm)
1	30	0	3.637	56.37	0.137	0.584	0.1672	205.0
2	30	0	3.156	84.41	0.246	0.044	0.0004	266.4
3	30	0	4.420	63.63	0.080	0.982	0.8988	281.2
4	30	0	2.546	92.21	0.157	0.411	0.0642	234.7
5	30	0	2.174	58.83	0.075	0.991	0.9452	127.9
6	30	0	2.650	19.34	0.195	0.181	0.0066	51.2
7	30	0	1.960	15.24	0.207	0.133	0.0042	29.9
8	30	0	2.402	8.06	0.171	0.311	0.0314	19.4
9	30	0	1.441	21.30	0.232	0.068	0.0002	30.7
10	30	0	3.167	12.44	0.122	0.716	0.3073	39.4
11	30	0	1.222	51.08	0.230	0.071	0.0008	62.4
12	30	0	2.556	38.58	0.116	0.776	0.4017	98.6

Table 11. Gamma fits for the standardized precipitation index, one per calendar month, over the separate thirty-year climatological reference dataset used exclusively for SPI calibration. The asymptotic column is shown only to document why it cannot be used.

The goodness-of-fit test had to be recomputed. The Kolmogorov–Smirnov statistic was being compared against its asymptotic null distribution while the gamma parameters were estimated from the same sample that was being tested. That is the Lilliefors problem: the null distribution no longer applies, the p values come out systematically too large, and the single rejection they produced was not interpretable. The column reported here is a parametric bootstrap with B = 5,000 resamples and seed 20260904, in which each simulated sample is refitted before its statistic is computed, so that the estimation of the parameters is inside the null distribution rather than outside it.

The corrected result is materially worse than the one it replaces. The asymptotic column rejects one month; the valid bootstrap rejects six — February, June, July, August, September and November — with a minimum p of 0.0002 in September. The rejected months are the dry half of the year, where monthly totals are small and the gamma fits the upper tail poorly. The consequence is stated plainly: the SPI values reported below for those six months rest on a distributional fit that the data reject, and should be read as indicative only. The same fitted parameters are applied to the real and to the synthetic series, so the comparison between them remains on a common scale, but the absolute calibration of the index in the dry season is not established by these data. A distribution with a heavier tail, or a non-parametric empirical transformation, is the natural remedy and is left to future work rather than fitted post hoc here.

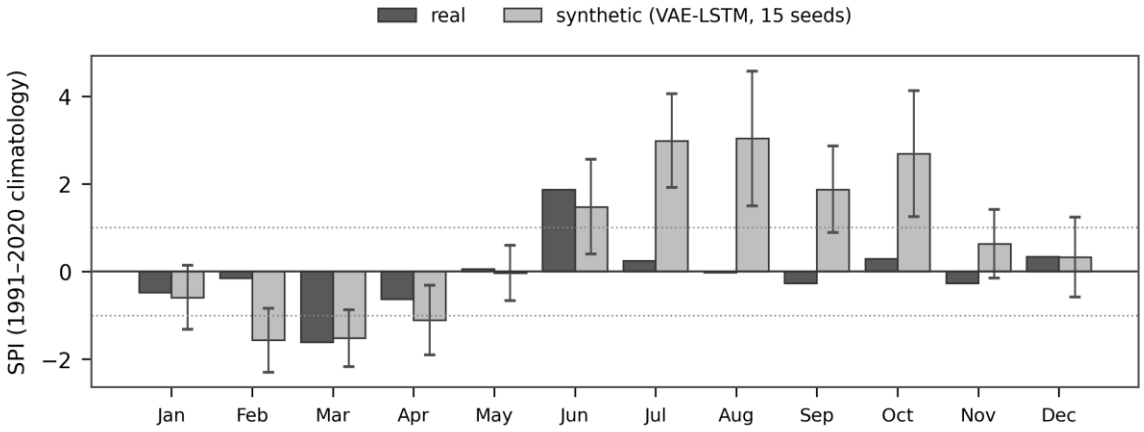


Figure 6. Monthly standardized precipitation index for the test year, real and synthetic (mean ± SD over fifteen seeds), on the 1991–2020 climatology.

Figure 6 places the two monthly series side by side. The synthetic index tracks the sign of the real one in the first half of the year and departs from it in the second, where it places the dry months of July to October in strongly wet territory. The current synthetic SPI should not be used as a standalone wet-extreme indicator until it has been evaluated against a longer climatological reference and a multi-year test period. Operational drought monitoring products built on reanalysis [19] and on deep learning [6, 39] provide the reference against which a future multi-year evaluation should be run.

3.7 Neutrosophic synthesis

Table 12 gives the Stage A classification of the thirty-five elementary contrasts by the rule of Table 2, and Table 13 the matrix that results from applying the Stage B mapping of Table 3 to those counts, with the base-case weights and the four aggregate indices.

Block	D+	D−	EQ	IND+	IND−	n	Content
E1	2	5	0	2	3	12	marginal distance, 3 references × 4 variables
E2	4	0	0	0	0	4	scenario separation, by variable
E3	0	7	4	1	0	12	within-scenario realism, 3 scenarios × 4 variables
E4	0	2	0	1	1	4	correlation preservation, key pairs
E5	0	0	1	1	1	3	latent-dimension ablation

Table 12. Stage A classification of the elementary contrasts by evidence block. States IND+ and IND− are reported separately but both denote inconclusive evidence.

Block	n	w_i	H1 $\langle T, I, F \rangle$	H2 $\langle T, I, F \rangle$	H3 $\langle T, I, F \rangle$
E1	12	1	$\langle 0.167, 0.417, 0.417 \rangle$	$\langle 0.417, 0.000, 0.583 \rangle$	$\langle 0.417, 0.417, 0.167 \rangle$
E2	4	1	$\langle 0.000, 1.000, 0.000 \rangle$	$\langle 0.000, 1.000, 0.000 \rangle$	$\langle 1.000, 0.000, 0.000 \rangle$
E3	12	1	$\langle 0.000, 1.000, 0.000 \rangle$	$\langle 0.000, 1.000, 0.000 \rangle$	$\langle 0.333, 0.083, 0.583 \rangle$
E4	4	1	$\langle 0.000, 1.000, 0.000 \rangle$	$\langle 0.000, 1.000, 0.000 \rangle$	$\langle 0.000, 1.000, 0.000 \rangle$
E5	3	1	$\langle 0.000, 1.000, 0.000 \rangle$	$\langle 0.000, 1.000, 0.000 \rangle$	$\langle 0.000, 1.000, 0.000 \rangle$
Weighted means $\langle \overline{T}, \overline{I}, \overline{F} \rangle$			$\langle 0.033, 0.883, 0.083 \rangle$	$\langle 0.083, 0.800, 0.117 \rangle$	$\langle 0.350, 0.500, 0.150 \rangle$
R, Eq. (6)			0.0036	0.0147	0.1487
S, Eq. (6)			0.3556	0.3889	0.5667
Net truth $\overline{T} - \overline{F}$			−0.0500	−0.0333	0.2000
Discounted truth $\overline{T}(1 - \overline{I})$			0.0039	0.0167	0.1750

Table 13. Complete NACH matrix. Every entry follows from Table 12 and Table 3 by the calibration of Section 2.7; the last four rows are the aggregate indices, each on its own scale and not comparable across rules.

In the base case the capability-dependent reading leads under all four rules. Its robust-plausibility index is 0.149, against 0.015 for the no-detectable-difference reading and 0.004 for distributional superiority; the ordering H3 > H2 > H1 is the same under the linear score, under net truth and under indeterminacy-discounted truth. The normalized profile distance between the leading and the second reading is 0.200. Two things about that lead should be read together with it. The profiles are dominated by indeterminacy — 0.500 for H3 and 0.800 for H2 — so the leading hypothesis leads within a body of evidence that is mostly uninformative about it. And the lead is not uniform across blocks: H3 is fully supported by the separation block, where all four contrasts are conclusive in its favour, and is opposed by the realism block, where seven of twelve contrasts fall against it.

The asymmetry of Table 3 has a quantitative consequence that must be stated before the robustness figures, because it conditions them. H1 and H2 are adjudicated by the twelve contrasts of E1 alone; H3 is additionally adjudicated by E2 and E3, twenty-eight contrasts in all. Blocks E4 and E5 map to indeterminacy for every hypothesis and therefore raise the indeterminacy of all three profiles without separating them. Restricting the comparison to E1, the one block that all three hypotheses share, changes the picture: the robust-plausibility index becomes 0.203 for H3, 0.174 for H2 and 0.057 for H1; the linear score is exactly equal for H2 and H3 at 0.611; net truth favours H3 (+0.250 against −0.167); and indeterminacy-discounted truth reverses the order, favouring H2 at 0.417 against 0.243. On the shared evidence the two readings are close to indistinguishable, and the margin H3 enjoys in the full matrix comes from the blocks that only H3 is exposed to. That is a property of the hypothesis set, not a stronger result, and it is reported here rather than left for the reader to reconstruct.

Aggregation rule	Range	H1 first (%)	H2 first (%)	H3 first (%)	Ties (%)
$R = \overline{T}(1 - \overline{I})(1 - \overline{F})$	[0,1]	0.0	0.9	99.1	0.0
$S = \frac{2 + \overline{T} - \overline{I} - \overline{F}}{3}$	[0,1]	0.0	0.0	100.0	0.0
$\overline{T} - \overline{F}$	[−1,1]	0.0	7.3	92.7	0.0
$\overline{T}(1 - \overline{I})$	[0,1]	0.0	2.7	97.3	0.0

Table 14. Rank robustness under resampled block weights, reported separately for each aggregation rule over 20 000 sampled weight vectors. Raw index values are not compared across rules, whose ranges differ.

Table 14 gives the rank robustness rule by rule. H3 ranks first in 99.1 % of the sampled weight vectors under the robust-plausibility index, 100.0 % under the linear score, 92.7 % under net truth and 97.3 % under indeterminacy-discounted truth, with no ties. Four qualifications belong with those figures and are not rhetorical. The sampling moves the block weights only; the mapping of Table 3, the minimum relevant effects and the equivalence regions are held fixed, and they are the choices that carry the interpretive weight. The sampling is Monte Carlo and not exhaustive. The robustness is conditional on the hypothesis set, and the preceding paragraph shows that on the evidence all three hypotheses share, the leading pair is nearly tied. And a high proportion here means

that the ordering is stable against how the blocks are weighted; it is not a posterior probability that H3 is true, and nothing in this procedure licenses reading it as one.

The threshold τ, which separates the IND+ from the IND− label, was varied from a fifth to three times its declared value in every block. No degree, no aggregate index and no ranking changed anywhere in that range, which is what the construction predicts: both labels map to the same degree under Stage B, so τ is descriptive and mathematically inactive. It is retained in Table 12 because the distinction between a large and a small inconclusive effect is informative to a reader, and it is excluded from Table 4 because varying an inactive parameter would overstate the extent of the sensitivity analysis.

4. DISCUSSION

The clearest empirical result is one the earlier design could not have produced. Under a chronological split in which the test year is genuinely held out, a conditioned bootstrap with no learned parameters attains a smaller Wasserstein distance than the VAE-LSTM on three of the four variables, and a convolutional variational baseline does so on two. The variational component is not idle — the VAE-LSTM beats its own deterministic ablation on two variables — but the comparison that matters for practice, against a resampling procedure that costs nothing to fit, does not favour the generative model on marginal fidelity. The caution generalizes beyond this case: under a genuinely prospective, chronological split, a parameter-free resampling baseline is exactly the comparison a generative model must beat to justify its added complexity, and here it does not, on the same marginal Wasserstein distance reported above for three of the four variables.

The second result concerns what scenario generation is worth. The three typified scenarios separate with large effect sizes and under a heteroscedasticity-robust test, and that is a capability the bootstrap does not have in the same form: resampling remains restricted to the empirical support of the training record, so it can recombine observed windows but cannot place mass outside the values that were observed, whereas a latent-space model can. But the realism test shows that the generated regimes do not reproduce the dispersion of the real conditioned data. Wind speed collapses towards the scenario centre under two of the three scenarios, precipitation under drought is generated with roughly thirty-eight times the real variance while covering a tenth of the real range, and the heavy-rain scenario rests on ten real windows and is over-dispersed on three variables. Scenario generation is therefore a capability that this model has and the baselines lack, and one that it does not yet exercise faithfully.

The methodological point is the one that generalizes. Each of these findings, taken alone, invites a different verdict, and a study free to choose its headline could report any of them. The synthesis of Section 3.7 exists to prevent that choice from being made after the fact: the statistical classification is fixed and hypothesis-free, the interpretive mapping is tabulated in full and separately, every matrix entry is derivable from the reported contrasts, and the ranking is examined under four aggregation rules rather than one. What that procedure returns here is not a decisive verdict but a well-characterized indecision. The capability-dependent reading leads under every rule and under almost every weighting, but it does so within profiles whose largest component is indeterminacy, with its own realism block counting against it, and with a lead that comes from the two evidence blocks to which only it is exposed: on the block that all three readings share, it and the no-detectable-difference reading are separated by nothing worth interpreting. A procedure that reported the ranking without that decomposition would be reporting the shape of the hypothesis set as though it were a finding about the models.

One architectural limitation should be read alongside the distributional results, because it bears on them directly. All four generators operate on variables scaled to the unit interval by a min–max transformation fitted on the training period, and the two variational models and the autoencoder emit their reconstruction through a sigmoid. That output is bounded by construction, so no generated value can fall outside the training minimum and maximum of its variable, and the sigmoid saturates near both ends, which compresses the density where the extremes of precipitation and wind speed live. Part of the marginal-fidelity gap reported above is attributable to that choice rather than to the latent model itself, and the effect is strongest for the two most skewed variables. A bounded output was retained because it

guarantees physically admissible values, but an unbounded head with an explicit positivity constraint, or a scaling that does not compress the tails, is the first thing to vary in a follow-up.

Four further limitations bound these results. The test period is a single year, so the multi-seed protocol resolves variability between training runs but not year-to-year representativeness; a rolling-origin design over several test years is the natural extension. All series come from a single site, La Maná, so the transferability of the ranking to other locations is not assessed here. The comparison covers one recurrent generator, one deterministic ablation, one resampling baseline and one convolutional variational baseline, and does not include adversarial or diffusion-based generators, which would strengthen the reference set. And the scenarios are mapped to proposed flight-decision protocols for the UAV platform; end-to-end validation against NDVI, NDRE and GNDVI computed from actual flights has not been performed, and no operational claim is made beyond that mapping. Deep learning on UAV imagery for crop condition [40] and vegetation-index estimation under fault-tolerant control [29] describe the pipeline into which such a mapping would feed, and hybrid forecasting-and-recommendation systems [27] the decision layer above it.

The monitoring framework has potential operational alignment with Sustainable Development Goals 2, 13 and 15. Environmental, agronomic and resource-use impacts were not directly measured in this study, and no reduction in irrigation or agrochemical application, no yield protection and no improvement in adaptive capacity is claimed as an observed outcome.

5. CONCLUSIONS

Four generators of daily agroclimatic series for La Maná were compared over fifteen matched seeds under a strictly chronological design in which the scaler, the early stopping, the latent dimension and the architecture were all fixed on a validation year preceding the test year, and the test year was read once. Four distributional metrics are reported in full for every model and variable.

On marginal distributional distance the VAE-LSTM is outperformed by a conditioned bootstrap on three of four variables and by a convolutional variational baseline on two, while outperforming its own deterministic ablation on two. It reproduces the cross-correlation structure imperfectly: two of four key pairs show a difference whose bootstrap interval excludes zero, a result consistent with a structural limitation of the present reconstruction objective. It separates the three typified scenarios with large effect sizes under a robust test, but does not reproduce the within-scenario dispersion of the real conditioned data, with collapse on some cells and gross over-dispersion on others.

The neutrosophic synthesis condenses that mixture into an auditable verdict whose every entry is derivable from the reported contrasts through a statistical stage and an interpretive stage that are kept apart. The reading that the model's potential utility is capability-dependent leads in the base case under all four aggregation rules (R = 0.149, against 0.015 and 0.004) and ranks first in between 93 % and 100 % of resampled block weightings. That lead is real and it is also bounded: indeterminacy is the largest component of all three profiles, and on the single evidence block that all three readings share, the leading pair is effectively tied. The evidence therefore points towards, without settling on, the reading that the model's value would lie in scenario generation rather than in marginal fidelity — and the realism test says that this value is potential rather than realized in the present model.

The procedure is the transferable part. When contrasts disagree and many are inconclusive, the choice among generative models can still be made explicit, reproducible and auditable, provided the classification rule is fixed in advance of the ranking, the analyst-dependent choices are listed rather than denied, and the robustness of the conclusion is reported separately for each aggregation rule instead of pooled into a single number. This comparative evaluation among competing statistical and machine learning architectures aligns with multicriteria frameworks for selecting models under operational constraints [23].

ACKNOWLEDGMENTS

Data availability: the NASA POWER records used here are public.

RECEIVED: APRIL, 2026.
REVISED: SEPTEMBER, 2026.

REFERENCES

[1] S. ACHARKI et al., "Comparative assessment of empirical and hybrid machine learning models for estimating daily reference evapotranspiration in sub-humid and semi-arid environments," **Model. Earth Syst. Environ.**, vol. 11, art. 21, 2025.

[2] M. AFONSO, V. GIUFRIDA, "Synthetic data for computer vision in agriculture," **Front. Plant Sci.**, vol. 14, art. 1277073, 2023. https://doi.org/10.3389/fpls.2023.1277073

[3] Y. AKKEM, S. K. BISWAS, A. VARANASI, "A comprehensive review of synthetic data generation in smart farming using variational autoencoder and generative adversarial network," **Eng. Appl. Artif. Intell.**, vol. 131, art. 107881, 2024.

[4] C. ALEXANDER, D. M. GOVERN, "A new and simpler approximation for ANOVA under variance heterogeneity," **J. Educ. Behav. Stat.**, vol. 19, no. 2, pp. 91–101, 1994. https://doi.org/10.3102/10769986019002091

[5] R. G. ALLEN, L. S. PEREIRA, D. RAES, M. SMITH, **Crop Evapotranspiration: Guidelines for Computing Crop Water Requirements**. FAO Irrigation and Drainage Paper 56. Rome, Italy: FAO, 1998.

[6] J. A. Anochi, M. H. Shimizu, "Precipitation forecasting and drought monitoring in South America using a machine learning approach," **Meteorology**, vol. 3, no. 1, pp. 82–96, 2024.

[7] A. DESAI, C. FREEMAN, Z. WANG, I. BEAVER, "TimeVAE: A variational auto-encoder for multivariate time series generation," **arXiv:2111.08095**, 2021. https://doi.org/10.48550/arXiv.2111.08095

[8] T. GNEITING, A. E. RAFTERY, "Strictly proper scoring rules, prediction, and estimation," **J. Am. Stat. Assoc.**, vol. 102, no. 477, pp. 359–378, 2007. https://doi.org/10.1198/016214506000001437

[9] A. GRETTON, K. M. BORGWARDT, M. J. RASCH, B. SCHÖLKOPF, A. SMOLA, "A kernel two-sample test," **J. Mach. Learn. Res.**, vol. 13, pp. 723–773, 2012.

[10] A. GROTH, E. CHAVEZ, "Efficient inference and learning of a generative model for ENSO predictions from large multi-model datasets," **Clim. Dyn.**, vol. 53, pp. 7359–7376, 2019.

[11] G. H. HARGREAVES, Z. A. SAMANI, "Reference crop evapotranspiration from temperature," **Appl. Eng. Agric.**, vol. 1, no. 2, pp. 96–99, 1985. https://doi.org/10.13031/2013.26773

[12] R. J. HEUER, **Psychology of Intelligence Analysis**. Washington, DC, USA: Center for the Study of Intelligence, Central Intelligence Agency, 1999.

[13] S. HOLM, "A simple sequentially rejective multiple test procedure," **Scand. J. Stat.**, vol. 6, no. 2, pp. 65–70, 1979.

[14] INAMHI, **Anuario Meteorológico del Ecuador 2023**. Quito, Ecuador: Instituto Nacional de Meteorología e Hidrología, 2023.

[15] S. JEON, J. T. SEO, "A synthetic time-series generation using a variational recurrent autoencoder with an attention mechanism in an industrial control system," **Sensors**, vol. 24, no. 1, art. 128, 2024.

[16] W. JIA, Y. ZHANG, Z. WEI, Z. ZHENG, P. XIE, "Daily reference evapotranspiration prediction for irrigation scheduling decisions based on the hybrid PSO-LSTM model," **PLoS ONE**, vol. 18, no. 4, art. e0281478, 2023.

[17] A. I. KAFI, A. Sanfilippo, and S. A. H. Shannak, "Evaluating variational autoencoders for synthetic time series data generation in agricultural applications," **Comput. Electron. Agric.**, vol. 226, art. 109412, 2024.

[18] C. KARAM, M. AWAD, Y. ABOU JAWDAH, N. EZZEDDINE, AND A. FARDOUN, "GAN-based semi-automated augmentation online tool for agricultural pest detection," **Front. Plant Sci.**, vol. 13, art. 813050, 2022.

[19] J. KEUNE, F. DI GIUSEPPE, C. BARNARD, E. DAMASIO DA COSTA, F. WETTERHALL, "ERA5-Drought: Global drought indices from reanalysis," **Earth Syst. Sci. Data**, vol. 17, pp. 1–24, 2025.

[20] D. P. KINGMA, M. WELLING, "Auto-encoding variational Bayes," **arXiv:1312.6114**, 2013. https://doi.org/10.48550/arXiv.1312.6114

[21] H. R. KÜNSCH, "The jackknife and the bootstrap for general stationary observations," **Ann. Stat.**, vol. 17, no. 3, pp. 1217–1241, 1989. https://doi.org/10.1214/aos/1176347265

[22] D. LAKENS, "Equivalence tests: A practical primer for t tests, correlations, and meta-analyses," **Soc. Psychol. Personal. Sci.**, vol. 8, no. 4, pp. 355–362, 2017. https://doi.org/10.1177/1948550617697177

[23] M. Y. LEYVA VÁZQUEZ, L. A. BRIONES PEÑAFIEL, S. X. SÁNCHEZ MUÑOZ, M. A. QUIROZ MARTÍNEZ, "A framework for selecting machine learning models using TOPSIS," **Advances in Intelligent Systems and Computing**, vol. 1213, pp. 119-126, 2021, doi: 10.1007/978-3-030-51328-3_18.

[24] M. Y. LEYVA VÁZQUEZ, F. SMARANDACHE, "Neutrosophic analysis of competing hypotheses (NACH): A novel framework for reasoning under contradictory and indeterminate evidence," **Neutrosophic Sets Syst.**, vol. 85, pp. 1–20, 2026.

[25] M. Y. LEYVA VÁZQUEZ, F. SMARANDACHE, "Neutrosophic paraconsistent logic: Evidence degrees, ontological indeterminacy, and scientific evidence synthesis," **Neutrosophic Comput. Mach. Learn.**, vol. 43, pp. 211–221, 2026.

[26] T. B. MCKEE, N. J. DOESKEN, J. KLEIST, "The relationship of drought frequency and duration to time scales," **in Proc. 8th Conf. Applied Climatology**, Anaheim, CA, USA, 1993, pp. 179–184.

[27] S. A. MOHAMED et al., "A hybrid deep learning and rule-based model for smart weather forecasting and crop recommendation," **Sci. Rep.**, vol. 15, art. 12045, 2025.

[28] A. V. NIKITIN, L. IANNUCCI, S. KASKI, "TSGM: A flexible framework for generative modeling of synthetic time series," **arXiv:2305.11567**, 2023.

[29] G. ORTIZ-TORRES et al., "Integrating actuator fault-tolerant control and deep-learning-based NDVI estimation for precision agriculture with unmanned aerial vehicles," **Agriculture**, vol. 15, no. 6, art. 620, 2025.

[30] L. PALMA et al., "Data-driven seasonal climate predictions via variational inference and transformers," **npj Clim. Atmos. Sci.**, vol. 8, art. 41, 2025.

[31] A. PASZKE et al., "PyTorch: An imperative style, high-performance deep learning library," **in Advances in Neural Information Processing Systems 32**, 2019, pp. 8024–8035.

[32] I. PRICE et al., "Probabilistic weather forecasting with machine learning," **Nature**, vol. 637, no. 8044, pp. 84–90, 2025.

[33] S. RANA, M. GATTI, "Comparative evaluation of modified Wasserstein GAN-GP and state-of-the-art GAN models for synthesizing agricultural weather data," **Comput. Electron. Agric.**, vol. 217, art. 108596, 2024.

[34] P. W. STACKHOUSE et al., "POWER release 8 methodology," **NASA Tech. Rep.**, Hampton, VA, USA, 2018. https://power.larc.nasa.gov

[35] M. TAILLARDAT, A.-L. FOUGÈRES, P. NAVEAU, AND R. DE FONDEVILLE, "Evaluating probabilistic forecasts of extremes using continuous ranked probability score distributions," **Int. J. Forecast.**, vol. 39, no. 3, pp. 1448–1459, 2023.

[36] M. L. TAN et al., "Evaluation of NASA POWER and ERA5-Land for estimating tropical precipitation and temperature extremes," **J. Hydrol.**, vol. 624, art. 129940, 2023.

[37] H. C. S. THOM, "A note on the gamma distribution," **Mon. Weather Rev.**, vol. 86, no. 4, pp. 117–122, 1958.

[38] C. VILLANI, **Topics in Optimal Transport**, 2nd ed. Providence, RI, USA: American Mathematical Society, 2021.

[39] Y. ZHANG et al., "Construction of an integrated drought monitoring model based on deep learning algorithms," **Remote Sens.**, vol. 15, no. 3, art. 667, 2023.

[40] H. ZHU et al., "Intelligent agriculture: Deep learning in UAV-based remote sensing imagery for crop diseases and pests detection," **Front. Plant Sci.**, vol. 15, art. 1435016, 2024.