Simple Models or Deep Learning Architectures? A Study on Intergranular Temperature Prediction in Vertical Grain Silos

Gustavo Queiroz Silveira[1], **Danilo Sipoli Sanches**[1],
Rafael Staiger Bressan[2], **André Robero Ortoncelli**[1]

[1]Universidade Tecnológica Federal do Paraná (UTFPR)
[2]Universidade Federal do Paraná (UFPR)
gustavoqueiroz@alunos.utfpr.edu.br, danilosanches@utfpr.edu.br
profbressan@gmail.com, ortoncelli@utfpr.edu.br

Abstract. *This study presents a comparative analysis of accurate intergranular temperature prediction using real data from three Brazilian vertical silos, comparing statistical methods, tabular machine learning, temporal deep learning, and spatiotemporal deep learning models under a unified evaluation protocol. To avoid overestimating performance, results are analyzed by aeration condition and compared with a persistence baseline. The results show that predictive performance saturates at the tabular level: Random Forest and XGBoost with aeration-related features consistently outperform persistence during aeration. At the same time, temporal and spatiotemporal deep learning models incur higher computational costs without improving accuracy.*

1. Introduction

Grain storage is critical in the agribusiness. In Brazil, approximately 20% of annual production is lost between harvesting and storage, largely due to the development of microorganisms, insects, and deteriorative processes favored by inadequate thermal conditions within the grain mass [Reginato et al. 2014]. Intergranular temperature is, in this context, one of the main indicators of quality, directly related to metabolic activity and the risks of deterioration of the stored product [Paraginski et al. 2015].

Aeration, the forced passage of ambient air through the grain mass by means of fans installed at the base of the silo, is the main mechanism for thermal control, reducing and standardizing the temperature and dissipating hot spots. Continuous temperature monitoring through thermometry and proper aeration activation enable preventive action before problems reach critical values [Reginato et al. 2014]. In this scenario, predicting the evolution of intergranular temperature is a valuable resource for supporting management decisions, and recent studies highlight the potential of machine learning and deep learning techniques for optimizing these processes [Junior et al. 2024].

Recent studies have proposed deep learning models for the thermal dynamics of silos; however, most focus on horizontal silos, while vertical silos remain underexplored, despite their widespread use in Brazil. These approaches generally entail high computational costs and, in the context of grain storage, are rarely systematically compared with simpler alternatives within the same experimental protocol, such as classical time-series methods, tabular regressors, and simpler temporal neural networks. Furthermore, the limited availability of data to replicate experiments, coupled with the predominance of studies conducted on horizontal grain silo structures abroad, leaves significant gaps.

This gap is not unique to the domain. The time-series forecasting literature has repeatedly questioned whether increased complexity actually translates into predictive gain: competitions and comparative studies show that simple statistical methods and models often constitute strong baselines, sometimes difficult to surpass with deep architectures [Makridakis et al. 2018, Elsayed et al. 2021]. Without a fair and standardized comparison, it is impossible to determine whether the gain obtained from using more robust deep-learning architectures is justified for predicting the temperature of a grain silo, whether these models are based on recurrent neural networks that exploit spatio-temporal convolutions in the data, or on transformers [Li et al. 2025].

In the specific case of intergranular temperature, there is an additional, little-discussed methodological complication. Because it acts as a thermal insulator, the grain mass maintains a practically static internal temperature for most of the time, varying significantly only during aeration periods, when the forced airflow couples the mass to the external environment. As a consequence, any model — even a trivial predictor that repeats the last value or the average — obtains a low absolute error when evaluated over the entire series, since the static regime dominates the global statistics. This effect masks the differences between families of models and overestimates the performance of complex architectures: the real challenge lies not in the trivially predictable average series, but in the dynamics induced by aeration. A benchmark that does not separate these two regimes or report the relative gain over naive baselines tends to produce misleading conclusions.

Problem. This work starts from the following question: *in the prediction of intergranular temperature in Brazilian vertical silos, how much predictive gain does each level of architectural complexity add over simple baselines — particularly in the dynamic aeration regime — and does that gain stem more from the complexity of the model or from making it aware of aeration onset?*

Objective. Establish, on real thermometry data from vertical silos (provided by a partner company), a benchmark of model families for temperature prediction, evaluated to isolate the aeration regime and to measure the gain relative to naive baselines. The investigation is organized around two crossed factors. The first is a "ladder" of increasing complexity: (i) naive and statistical methods (persistence, ARIMA/SARIMA); (ii) tabular machine learning (Random Forest, gradient boosting); (iii) temporal deep learning (LSTM, 1D-CNN); and (iv) 2D spatio-temporal models (ConvLSTM). The second factor contrasts aeration-aware models — which incorporate the motor state and the external temperature lagged by the physical thermal lag (the delay between external-air variation and the response of the sensors in the mass) — with aeration-blind, purely autoregressive models. All are evaluated under a leak-free temporal split, with metrics stratified by regime (during and outside aeration) and relative to the naive predictor (MASE and skill score), together with computational-cost metrics.

Contributions. The main contributions of this work are: (1) a comparative, reproducible benchmark of model families for temperature prediction in Brazilian vertical silos under a unified protocol; (2) an evaluation protocol that isolates the aeration regime and reports the gain relative to naive baselines, avoiding the accuracy overestimation caused by the near-static nature of the series; and (3) empirical evidence on the extent to which the predictive gain stems from architectural complexity *versus* from making the model aware of aeration onset, through the engineering of physical thermal-lag features.

Our results show that predictive accuracy saturates at the tabular level: Random Forest and XGBoost with physical thermal-lag features are the only models that consistently beat persistence in the aeration regime (MASE down to ≈ 0.52 at the 12 h horizon), at roughly one to two seconds per fit, whereas the temporal (1D-CNN/LSTM) and 2D spatio-temporal (ConvLSTM) deep models cost up to three orders of magnitude more and are no better — often worse — in the regime that matters. Aeration awareness, in turn, converts into predictive gain *only* at the tabular level, where the physical lag is supplied as an explicit feature.

2. Theoretical Aspects

Thermometry is the science of temperature measurement, fundamental for monitoring stored agricultural products. In metal silos, grain silos, and other storage structures, temperature is one of the main indicators of quality, as it is directly related to microorganism metabolic activity, grain respiration, and the risk of deterioration [Paraginski et al. 2015].

2.1. Importance of Thermal Monitoring

During storage, the grain mass is constantly subjected to external factors that can compromise its quality: physical factors (temperature and humidity), chemical factors (oxygen supply), and biological factors (bacteria, fungi, insects, and rodents). Safe storage maintains the grains' qualitative and quantitative properties, preventing the development of these deteriorating agents. In Brazil, 20% of annual grain production is lost between harvest and storage.

Thermal monitoring is essential for the early detection of deterioration processes in stored grain masses. Elevated temperatures favor the proliferation of microorganisms (fungi and bacteria), increase grain respiration, cause weight and dry-matter losses, reduce commercial and nutritional quality, and may lead to changes in color, odor, and taste. In severe cases, self-heating can increase the risk of fire and stimulate the production of mycotoxins, which are harmful to human and animal health [Reginato et al. 2014].

Monitoring temperature increases is achieved through thermometry, enabling the identification and assessment of problem severity before they reach critical levels. Daily monitoring of these temperatures enables the preventive activation of the ventilation (aeration) system, preventing significant losses in the stored product [Reginato et al. 2014].

2.2. Aeration System

The aeration system consists of fans installed at the base of the silo that force ambient air through the grain mass by forced convection. Its main purpose is to even out and reduce the temperature of the grains, creating conditions unfavorable to the development of fungi, insects, and deterioration [Navarro and Noyes 2002, Montross et al. 2002].

Climatic criteria condition the system's activation: aeration is performed when the outside-air temperature is lower than the temperature of the grain mass, typically with a minimum difference of 4 to 6 °C, and when the relative humidity of the air does not pose a risk of wetting the grains. Automated systems use external temperature and humidity sensors to make this decision, activating the motor only during favorable windows [Reed and Harner 1998, Plumier et al. 2021].

The injected air percolates through the spaces between the grains, absorbing the accumulated heat and dissipating it through the upper surface of the mass. The cooling front advances from the base to the top of the silo [Navarro and Noyes 2002], which directly explains the temporal lag phenomenon observed in the sensors: sensors near the base respond more quickly to variations in the external air than sensors at the top [Plumier et al. 2021, Maier and Plumier 2022].

2.3. Temporal Lag in Temperature Measurements

Time lag, or time delay, represents the time interval required for changes in external temperature to be reflected in the reading of a sensor immersed in the grain mass during aeration periods. This physical phenomenon is a direct consequence of forced-convection heat transfer in the storage system.

During aeration, outside air is forced through the grain mass by the ventilation system. However, the sensors' thermal response is not instantaneous. The response time depends on multiple physical factors, including the distance of the sensor from the air inlet, the thermal resistance of the grain mass, the airflow rate of the system, and the thermal properties of the stored material.

Formally, the temporal lag τ of a sensor can be defined as the time shift that maximizes the cross-correlation between the external temperature series $T_{ext}(t)$ and the sensor temperature series $T_{sensor}(t)$ during periods of active aeration. In practice, lag calculation is performed only during periods of continuous aeration, since thermal coupling between the external environment and the grain mass occurs only when forced airflow is present. For periods without aeration, the correlation between T_{ext} and T_{sensor} tends to be low or nonexistent, as the grain mass acts as a thermal insulator.

Several physical and operational factors of the storage system determine the sensor's thermal response time. Sensors located closer to the air inlet and in central regions of the silo generally respond faster due to greater airflow exposure. Aeration-system flow rate and power also influence the propagation speed of the thermal front. In addition, grain properties such as porosity, density, moisture content, and thermal conductivity affect heat transfer within the grain mass. The duration of aeration is another important factor, since short periods may not allow the thermal signal to reach distant sensors. At the same time, larger temperature differences between the grain mass and incoming air tend to accelerate thermal equalization.

2.4. Related Work and the Complexity Debate

Recent work on grain-silo thermal modeling has shifted toward spatio-temporal deep learning, treating the sensor grid over time. In particular, represents the silo's thermal state as three-dimensional volumes evolving in time and combines 3D DenseNet and 3DCNN-LSTM networks for horizontal granaries; related studies extend this line with recurrent spatio-temporal convolutions and transformers [Li et al. 2025]. Such architectures are promising but computationally heavy and, in this domain, are rarely compared against simpler alternatives under a common protocol; code and data are seldom released, and the literature concentrates on foreign horizontal granaries, leaving the case of Brazilian vertical silos largely unestablished.

This issue is part of a broader forecasting debate. Large-scale comparisons and competitions repeatedly show that simple statistical methods are strong baselines, often hard to beat with deep architectures [Makridakis et al. 2018, Elsayed et al. 2021]; and, for tabular problems, tree-based ensembles still tend to outperform deep networks [Grinsztajn et al. 2022]. The model families on our ladder are well established: ARIMA/SARIMA [Box et al. 2015, Hyndman and Athanasopoulos 2018] for classical time series; Random Forests and gradient boosting [Géron 2022] for tabular learning; LSTM and one-dimensional convolutions [Ishida et al. 2024] for temporal modeling; and ConvLSTM [Shi et al. 2015] for spatio-temporal grids — the direct 2D predecessor of the 3D models above. This work addresses that gap through a regime-stratified comparison of these families on real Brazilian vertical-silo data.

3. Methodology

3.1. Experimental Database

The data used in this study comes from a partnership with a specialized company, that provides a monitoring platform for clients in the grain-storage sector. To strengthen external validity and ensure that aeration episodes are represented in evaluation, the benchmark pools data from three real storage units (vertical silos) (here denoted *C1*, *C2*, and *C3*), treated as independent series, each with its own observation window.

The reference unit (*C1*) contains two silos with 6 internal thermometric cables arranged vertically, with 6 sensors each (arranged in a 6×6 spatial grid – cable $\times$ height), in which the vertical position is oriented from 0 (base, near the air inlet) to 5 (top). The other two units contribute vertical silos with their own (taller) grid geometries: *C2* contributes two silos and *C3* three silos, both with up to 7 cables and 11 vertical positions per cable; for this reason, sensor and silo identifiers are namespaced per unit to prevent silent collisions when integrating series from different clients. Sensor readings are collected approximately every hour. In addition to the internal temperature measurements, the dataset also includes external environmental variables (air temperature, relative humidity, and rainfall) and aeration system status (motor operation and mode).

The database integrates five source tables per unit: (i) metadata about the cable and sensor registries;(ii) the main series of intergranular temperature per sensor; (iii) the external-climate series; and (iv) the event log of the aeration motor. As there is a single weather station per site, the external climate is applied to all silos of that unit. By temporally intersecting the available series, a synchronized database is obtained for each unit, considering the period between 07/05/2025 to 12/05/2026).

Data preparation follows integration, cleaning, and transformation steps. The main cleaning decisions, were: (i) mapping the sentinel values of the DS18B20 sensor (255, indicating no reading) and spurious readings of $0\,^{\circ}\mathrm{C}$ to `NaN`, avoiding artificial inflation of statistics; (ii) applying physical domain limits (temperature in $[-10, 70]\,^{\circ}\mathrm{C}$, humidity in $[0, 100]\%$) and handling a one-off glitch from the climate sensor; (iii) discarding the remaining historical identifiers from sensor changes (about 4.4% of the records), keeping only the current sensors; and (iv) reindexing each sensor on a regular hourly grid, imputing only short gaps (up to 3 h) by temporal interpolation and keeping larger gaps as absent, so as not to train through them.

The motor state is decoded from the event log and converted into a contin-

uous hourly state by forward-fill, generating the covariates `motor_on` (binary) and `aeration_mode`. The external climate and the aeration state are then aligned to the grid of each sensor by asof-join (tolerance of 1 h), materializing an integrated, tidy analytical base, with one row per (sensor, instant) pair containing the target temperature, the climate and aeration covariates, and the lag derivatives.

Because it acts as a thermal insulator, the grain mass only couples to the external environment during aeration; outside aeration, and when the silo is emptied, the sensors stop measuring the mass. To distinguish these regimes, a label `in_grain` is assigned based on the 24-hour moving standard deviation of each sensor's temperature: sensors with persistently very low variability are considered exposed (outside the mass) and are discarded from the modeling. This labeling delimits the effective storage episode.

3.2. Problem Formulation and Feature Engineering

The problem is formulated as a multi-horizon prediction of intergranular temperature. Given a window of past hourly readings from a sensor (and, depending on the model, of the exogenous covariates), the temperature is predicted at horizons $H \in \{1, 6, 12, 24\}$ h. Classical/tabular models operate **per sensor**; temporal models operate on the multivariate series; and the 2D spatio-temporal model operates on the frames of the sensor grid of each silo, projected onto a common 11×7 frame (Section 3.3), and is therefore the only level whose target — and whose error — is the whole grid rather than an individual sensor.

The cross-cutting experimental factor is aeration awareness. Because the coupling between external climate and the grain mass occurs only under forced airflow, the thermal lag between external-air variation and the sensor response is estimated by aeration-conditioned cross-correlation — that is, computed only during the hours when the motor is on. The estimated lag defines the lagged-climate covariates. We then contrast two configurations: aeration-blind models, which receive only seasonality terms (Fourier series with a 24 h period); and aeration-aware models, which additionally receive the motor state and the interaction between the lagged climate and the motor state (climate only "enters" as an exogenous driver when there is airflow). This contrast isolates the gain attributable to making the model aware of aeration onset, separating it from the gain attributable to architectural complexity.

3.3. Model Ladder (Experimental Setup)

The comparison is organized as a "ladder" of increasing architectural complexity:

- **Level 0 — naive and statistical**: uses Persistence and Seasonal Persistence as baseline benchmarks, alongside ARIMA/SARIMA. Daily seasonality is handled via Fourier terms to save computational costs, and external climate is included as an exogenous variable (SARIMAX) during aeration.
- **Level 1 — Tabular machine learning**: *Random Forest* and *gradient boosting* (XGBoost) [Géron 2022], with no explicit memory: temporal dependence enters only through features. Models are trained pooled over all sensors, with one model per horizon.
- **Level 2 — Temporal deep learning (1D).** *1D-CNN* and *LSTM* [Ishida et al. 2024], which model temporal dependence explicitly over an input window, treating the sensors as channels.

- **Level 3 — 2D spatio-temporal.** *ConvLSTM2D* [Shi et al. 2015], which models space and time in a single operator (convolution inside the LSTM gates) over the grid tensor of each silo. Since the three units have different geometries, all silos are projected onto a common 11×7 (position $\times$ cable) grid with a validity-mask channel, and the loss is masked to the immersed cells; the model predicts the *whole* grid at the four horizons at once, which is what makes its errors scored over cells rather than over selected sensors (Section 4).

Evaluation protocol. We use a leak-free temporal split. We adopt rolling-origin validation (an expanding window with a minimum training fraction of 50% and four consecutive test folds per series, with a 24 h embargo), which preserves temporal order (no leakage; the origin only advances) and distributes the aeration episodes across the test folds. Any normalization is fit on the training portion only. Metrics are stratified by regime — during aeration (motor on at some point within the prediction window) and at rest (motor off throughout) — recording, for each prediction, the aeration path of the target window. This stratification is essential: because the series is near-static at rest, global metrics are dominated by the trivial regime and mask the differences between models.

Evaluation Metrics. Predictive performance is assessed using absolute metrics (MAE and RMSE) and metrics relative to the naive persistence baseline, including MASE and a skill score based on MSE. Since the series are predominantly near-static, relative metrics are essential to quantify the effective gain over trivial predictions. All results are reported separately for aeration and resting periods. Statistical significance of differences between models is evaluated using the paired Diebold–Mariano [Diebold and Mariano 1995]test, while the Friedman test with Nemenyi post-hoc can be used for comparisons across multiple series [Demšar 2006].

4. Results and Analysis

Evaluation was conducted on the three real storage units, treated as independent series (pooling), under the rolling-origin protocol of Section 3.3 and stratified by regime. Aeration accounts for only 4.0% of the immersed test hours (Section 3.3), so the global metric is dominated by the near-static rest regime; we therefore always report stratified errors. For cross-level comparison, the persistence-relative **MASE** is the honest comparator: absolute MAE is not directly comparable across levels, since Levels 0–2 score per sensor while Level 3 scores over the whole grid (mixing easy and hard cells). The consequences of this scope difference, and the reason the ranking survives it, are stated in the note to Table 2; the tuning budget given to each level, which is what licenses reading the ranking as evidence about complexity rather than about effort, is described in Section 3.3.

4.1. Level 0 — naive and classical baselines

Table 1 reports the MAE (in °C) of each Level 0 model, by horizon and regime. All models, including naive persistence, achieved low absolute errors, confirming the near-static nature of the intergranular temperature series. This result highlights that absolute metrics alone are insufficient to distinguish predictive performance, particularly because the resting regime dominates the data and is easily predicted by trivial baselines.

In the aeration regime, where temperature variations are significant, classical models do not consistently outperform the persistence baseline. SARIMA shows only modest

Table 1. MAE (in °C) of each Level 0 model, by horizon and regime. Errors are scored per sensor, over the 6 series fitted individually at this level.

	Aeration (motor on)				**Rest** (motor off)			
Model	h=1	h=6	h=12	h=24	h=1	h=6	h=12	h=24
Persistence	**0.488**	2.042	2.603	**1.336**	0.530	2.411	3.305	1.314
Seasonal persistence	1.374	**1.345**	**1.351**	**1.336**	—	—	—	—
SARIMA (blind)	0.557	**1.644**	**2.163**	1.840	**0.479**	**1.394**	**1.749**	1.502
SARIMAX+aer (aware)	0.596	1.654	2.195	1.860	0.480	1.440	1.863	**1.496**

gains at intermediate horizons and is matched or surpassed by persistence at short and long horizons. Notably, SARIMAX+aer achieves a MASE of 1.303 at the 24-hour horizon, indicating worse performance than persistence and an inability to anticipate temperature changes caused by aeration onset.

The most relevant result, however, is the contrast between the blind and aware models. Contrary to expectation, making the model aware of the motor state and the lagged external climate **brought no gain**: SARIMAX+aer is statistically equivalent to — and at several horizons marginally worse than — the purely autoregressive SARIMA, and the Diebold–Mariano test indicated no significant, systematic advantage for the aware configuration. This suggests a limitation of Level 0 models, which are unable to effectively capture the abrupt and nonlinear temperature changes associated with aeration events. The aeration information is available to the model, but the Level 0 family lacks a mechanism to convert it into an accurate prediction of the shift.

4.2. Level 1 — Tabular machine learning is the accuracy plateau

The tabular models are the first to break the persistence floor in the regime that matters. Under aeration, both Random Forest and XGBoost beat persistence at horizons $h \geq 6$, reaching a MASE as low as 0.52 at h=12 (Table 2), with skill scores above 0.80 — more than double the corresponding Level 0 figures. At h=1 the naive baseline remains essentially unbeatable (MASE ≈ 0.97 for the best model): at one hour the temperature barely moves and the last value is already near-optimal. The two families tie within the margin on accuracy; XGBoost, running on GPU, is the cheaper option ($\approx$1.8$\times$ faster to fit) and barely grows in cost when the aeration features are added.

Table 2. Consolidated ladder: MASE under aeration (best variant of each level. Rows N0–N2 are scored per sensor; N3 row is scored over the whole grid.

Model (level)	h=1	h=6	h=12	h=24
Persistence (baseline, N0)	1.000	1.000	1.000	1.000
SARIMA / SARIMAX+aer (N0)	1.30	1.32	1.55	1.25
RF, aware (N1)	1.094	**0.517**	**0.568**	**0.937**
XGBoost, aware (N1)	1.105	**0.587**	**0.527**	**0.929**
1D-CNN, blind (N2)	1.285	**0.706**	**0.629**	**0.992**
LSTM, blind (N2)	1.279	**0.813**	**0.680**	**0.999**
ConvLSTM2D, blind-weighted (N3)	2.538	1.217	1.066	1.446

This is the only level where aeration awareness pays off. The Diebold–Mariano test on the blind $\times$ aware contrast favors the aware configuration in 6 of 8 comparisons, all significant at $p < 0.05$ for $h \geq 6$, with the gain growing with the horizon (at h=24, the

aware XGBoost reaches MASE 0.929 versus 1.017 for the blind, and a skill of 0.406 versus 0.191). The feature-importance analysis confirms that the aeration path, the planned motor state, and the lagged-climate$\times$motor interactions carry substantial weight — so the gain comes from modeling aeration onset and the lagged climate, not from adding trees. This confirms the sub-hypothesis that, given an explicit physical-lag representation, a memoryless tabular model converts aeration information into predictive gain.

4.3. Levels 2 and 3 — More complexity does not pay off

The temporal deep models (Level 2) do beat persistence in the mid horizons under aeration (best MASE 0.629 for the blind 1D-CNN at h=12), but they are **worse than the tabular plateau** and roughly an order of magnitude more expensive to train. Two findings stand out. First, with the input window and the capacity of each configuration selected on the internal validation split — separately for the blind and the aware variants, and with early stopping deciding the number of epochs (Section 3.3) — **aeration awareness no longer helps**: the blind configuration wins the Diebold–Mariano contrast in 16 of 16 cases (14 significant), and the validation search consistently selects *smaller* networks for the aware configuration (hidden size 64 against 128 for the 1D-CNN, 32 against 128 for the LSTM) — the extra aeration channels add more variance than signal. This is direct evidence that the negative result is not an artifact of under-tuning the aware variants: each of them was given its own search over the same grid, and the validation itself asked for less capacity. Second, the gain over persistence in the mid horizons comes from *temporal memory*, not from aeration awareness, mirroring the Level 0 finding.

The 2D spatio-temporal model (Level 3, ConvLSTM2D) is the most complex and, paradoxically, the **worst of the modern levels** in the regime of interest: it beats persistence only at rest (MASE 0.54–0.71 at h=6/12) and *never* under aeration or post-aeration, where it is catastrophically worse than persistence at h=1 (MASE 2.5–5.1). Training the full Level 3 grid took about an hour on GPU ($\approx$1000$\times$ the cost of a tabular fit). At this data scale — few immersed cells per silo and rare aeration — the spatial convolution has insufficient signal and regresses toward the mean; here, too, the blind configuration is at least as good as the aware one.

4.4. The accuracy–cost trade-off across the ladder

Table 3 pairs each level's best aeration-regime accuracy with its training cost. The picture is clear and non-monotonic in complexity: accuracy saturates at Level 1. The classical models (Level 0) never beat persistence under aeration; the tabular models (Level 1) are the only ones that beat it at h=6/12/24 and come closest at h=1, at $\approx$1–2 s per fit; the temporal models (Level 2) beat persistence at $h \geq 6$ but are worse than Level 1 and $\approx$10$\times$ costlier; and the 2D spatio-temporal model (Level 3) is the most expensive and the least accurate in the aeration/post-aeration regimes.

Figure 1 shows this trade-off graphically, plotting the persistence-relative MASE under aeration against training cost: the lower-left region (low cost, low error) is Pareto-optimal, and the aware XGBoost of Level 1 dominates it, sitting well below the persistence reference line while the spatio-temporal Level 3 models remain above it. The answer to the title question is therefore little: the useful accuracy plateau is reached already with tabular trees; climbing to temporal deep learning costs about an order of magnitude more and slightly loses accuracy, and climbing to 2D spatio-temporal deep learning

Table 3. Accuracy versus training cost for each model analyzed. The MASE for rows N0–N2 is scored per sensor, and for N3 over the whole grid.

Level	Representative model	Params	Train time/fit	Best MASE (aer.)
N0	Persistence	—	≈ 0	1.000 (ref)
N0	SARIMAX(2,1,1)+Fourier	—	seconds	1.25
N1	XGBoost (GPU)	300 trees	≈ 1.2 s	**0.527**
N1	Random Forest	100 trees	≈ 2.1 s	**0.517**
N2	1D-CNN (blind)	≈43 k	≈ 12 s	**0.629**
N2	LSTM (blind)	≈87 k	≈ 29 s	**0.680**
N3	ConvLSTM2D	≈14 k	$\approx$ 206–303 s	1.066

costs about three orders of magnitude more and loses further in the physically interesting regime.

A consistent pattern emerged across model levels: aeration-aware versions outperformed blind models only at Level 1. This suggests that the physical lag is converted into predictive gain only when it is handed to the model as an explicit feature. Sequence and grid models that must *re-learn* the thermal coupling from raw windows or frames fail to do so at this data scale, and the extra aeration channels add variance rather than signal.

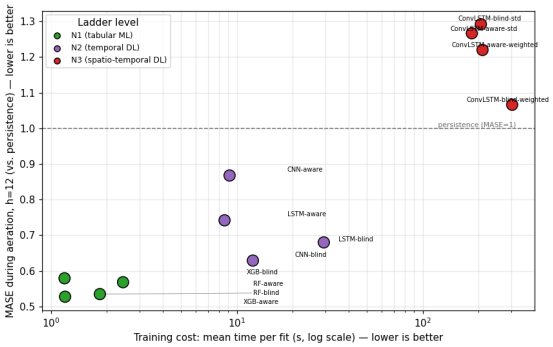


Figure 1. Accuracy–cost trade-off across the evaluated models.

5. Conclusion

This work established, based on real thermometric data, a set of reference models for predicting intergranular temperature, under a unified protocol, with metrics related to persistence and stratified by regime. The protocol itself is a contribution: since grain mass is a thermal insulator and remains practically static outside of aeration, the global metrics are dominated by the trivial regime and overestimate the model's competence; isolating the aeration regime and reporting the MASE/skill as a function of persistence is what reveals the real differences between the model families.

The central question *how much complexity is needed to predict silo temperature?* has a clear answer for this domain and data scale: **little**. Predictive accuracy saturates at Level 1. Random Forest and XGBoost are the only models that consistently beat persistence in the aeration regime (MASE down to ≈0.52 at h=12), at ≈1–2 s per fit. Temporal deep learning (1D-CNN/LSTM) costs about an order of magnitude more and is slightly less accurate; 2D spatio-temporal deep learning (ConvLSTM2D) costs about three orders of magnitude more and is the least accurate in the aeration and post-aeration regimes, beating persistence only at rest. The accuracy–cost trade-off is thus non-monotonic, and the aware XGBoost of Level 1 is Pareto-optimal.

The benchmark also answers the question on the source of the gain. Aeration awareness modeling the motor state and the lagged external climates converts into predictive gain *only* at the tabular level, where the physical lag is supplied as an explicit feature. At Levels 0, 2, and 3 the blind, purely autoregressive configuration is at least as good: models that must re-learn the thermal coupling from raw sequences or grids do not, at this data scale, turn it into a better prediction of the aeration-induced shift. Where deep models do help (mid horizons), the gain comes from temporal memory, not from awareness.

For the architectural decision that motivated this study, the implication is direct: at the current data scale, heavier spatio-temporal architectures are **not yet justified** for predicting temperature in Brazilian vertical silos; a tabular model with physical-lag features is the recommended operational choice, and the simple persistence baseline remains hard to beat at the one-hour horizon.

Limitations. Aeration is rare in absolute terms (4.0% of the immersed test hours, unevenly spread over the folds: 9.4%, 3.2%, 1.1% and 2.2%), This same scarcity most penalizes the data-hungry spatio-temporal model; the immersion detector uses a fixed threshold on the 24 h moving standard deviation; the hyperparameter search was not nested per fold (Section 3.3): Levels 2–3 were tuned once on the internal validation of the last fold and Levels 0–1 use fixed configurations, so the absolute figures could move under a per-fold nested search and the pooling treats the units as independent series, which is adequate for tabular models but is a simplification for the spatial grid, whose errors are also scored over cells rather than per sensor (note to Table 2).

The saturation at the tabular level is a statement about the current data scale, not about the intrinsic ceiling of spatio-temporal models: the aeration regime the very signal these architectures need is scarce, so the data-hungry models are precisely the most penalized. The hybrid architectures of Li et al. 2025 reach high accuracy supported by a much larger and richer thermal record, which suggests that the gap is one of data rather than of model family. Rather than discarding complex approaches, our result therefore motivates a deeper study of the database itself.

Future work. The main direction is to address the data scarcity problem. This involves collecting more aeration data, as well as applying data augmentation strategies tailored to thermal dynamics to better represent the aeration regime for spatio-temporal models. With a richer representation of this regime, it will be possible to fairly revisit more advanced spatio-temporal architectures, including Level-3 models such as 3D CNN-based approaches [Li et al. 2025]. Complementary directions include using the operational reference parameters as supervision for labeling aeration-episode quality, and extending the comparison with temporal convolutional networks and transformer-based forecasters under the same regime-stratified protocol. Additionally, due to page constraints, the data collection environment and time-series behavior are only briefly described; future work will provide a more detailed analysis.

References

Box, G. E. P. et al. (2015). *Time Series Analysis: Forecasting and Control*. John Wiley & Sons, 5th edition.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30.

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

Elsayed, S. et al. (2021). Do we really need deep learning models for time series forecasting? *arXiv preprint arXiv:2101.02118*.

Géron, A. (2022). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. " O'Reilly Media, Inc.".

Grinsztajn, L., Oyallon, E., and Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 507–520.

Hyndman, R. J. and Athanasopoulos, G. (2018). *Forecasting: Principles and Practice*. OTexts, Melbourne, Australia, 2nd edition.

Ishida, K. et al. (2024). Use of one-dimensional cnn for input data size reduction in lstm for improved computational efficiency and accuracy in hourly rainfall-runoff modeling. *Journal of Environmental Management*, 359:120931.

Junior, F. et al. (2024). Estimating energy efficiency of the aeration process of stored grains through machine learning. *Revista Brasileira de Engenharia Agrícola e Ambiental*, 28(11):e281001.

Li, X. et al. (2025). Smart grain storage solution: Integrated deep learning framework for grain storage monitoring and risk alert. *Foods*, 14(6):1024.

Maier, D. E. and Plumier, B. M. (2022). Grain operations: Temperature cable placement. *World Grain*. Acesso em: abr. 2026.

Makridakis, S., Spiliotis, E., and Assimakopoulos, V. (2018). Statistical and machine learning forecasting methods: Concerns and ways forward. *PLoS ONE*, 13(3):e0194889.

Montross, M. D., Maier, D. E., and Haghighi, K. (2002). Validation of a finite-element stored grain ecosystem model. *Transactions of the ASAE*, 45(5):1465–1474.

Navarro, S. and Noyes, R. T., editors (2002). *The Mechanics and Physics of Modern Grain Aeration Management*. CRC Press, Boca Raton, FL.

Paraginski, R. T. et al. (2015). Qualidade de grãos de milho armazenados em diferentes temperaturas. *Revista Brasileira de Engenharia Agrícola e Ambiental*, 19(4):358–363.

Plumier, B. M. et al. (2021). Effect of temperature sensor numbers and placement on aeration cooling of a stored grain mass using a 3D finite element model. *Agriculture*, 11(3):231.

Reed, C. and Harner, J. (1998). Cooling of stored wheat in multiple or single cycles using automatic aeration controllers. *Applied Engineering in Agriculture*, 14(5):495–500.

Reginato, M. P. et al. (2014). Boas práticas de armazenagem de grãos. *Anais do ENIC*, (6).

Shi, X. et al. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 28, pages 802–810.