Predicting grape variety phenology with machine-learning: capturing latent eco-physiological traits with embeddings

1st Joel Z. Harms
Grape Climate
Kelowna, Canada
Vintality
Kaleden, Canada
grape-climate@protonmail.ch

2nd Nathaniel K. Newlands
Summerland R. & D. Centre
Agriculture and Agri-Food Canada
Summerland, Canada

3rd Gregory V. Jones
Abacela Vineyards and Winery
Roseburg, USA

4th Christopher Mark
Vintality
Kaleden, Canada

***Abstract*—Agricultural machine learning models typically require large, variety-specific datasets and extensive calibration, severely limiting their transferability beyond a small group of well-documented cultivars. In this work, we demonstrate that a three-dimensional grapevine variety embedding, extracted from a general recommendation system, encodes biologically meaningful signals for phenology prediction. By integrating these embeddings into ensemble machine learning models, we effectively address the "variety start-up" problem—the inability to predict for new varieties without years of historical data. Our embedding-enhanced models reduce maturity prediction error by 53% relative to a linear classification baseline and show a 32% average improvement over non-embedding ensemble models. We further demonstrate strong zero-shot performance: models trained without variety-specific calibration outperform linear baselines by 34% and non-embedding ensembles by 19% on average. These results provide a pathway toward general-purpose grapevine phenology models that generalize across varieties. There is a growing need for phenology models that can generalize across varieties without extensive calibration, as cultivated variety-diversity is shifting due to climate change. Finally, we show that framing phenology as a probabilistic classification problem provides more actionable decision support for vineyard management than traditional binary thresholds. Our results highlight the potential of embedding-based representation learning to modernize decision support in viticulture.**

***Index Terms*—Precision Agriculture, Viticulture, Representation Learning, General Purpose Models**

Introduction

Applying state of the art machine learning (ML) methods to help guide agricultural production is challenging. The challenge stems mainly from data availability. Generally speaking, for many management variables such as yield, phenology, disease, soil carbon, or nutrient status no more than a few data points per year are available for each location. This means many years of dedicated data collection are needed to assemble representative datasets. Further, applying general purpose models to specialized management problems is not feasible since they do not possess the specialized domain knowledge necessary [1]. This is true particularly when considering specialized agriculture, like viticulture (growing grapes for wine). Where data is available it is focused on a few common varieties or cultivars, but the current challenge in agriculture, due to climate change, is exactly in leveraging the diversity of crops [2], [3], which is not well represented in the data. Hence currently, critical processes in agriculture generally need to rely on physical models, which need to be re-calibrated for every new variety, which requires datasets covering multiple years at multiple locations. Datasets of sufficient size are only available for a few grape varieties [4].

An important step to overcome current challenges would be to develop transfer learning or domain adaption strategies to deal with agricultural complexity and data scarcity [5], [6]. Models need to be flexible, regionally adaptable and inclusive of present crop variety diversity [5], [7], which may require the development of agricultural foundation models [8]. To aid in this aim we propose a method to extract grape variety embeddings from a general climate-based grape variety recommendation model [7]. We hypothesize that these embeddings capture latent eco-physiological traits, effectively solving the 'variety start-up' problem by mapping 1,300+ varieties onto a shared biological embedding. We examine their potential for enhancing random forest (RF) models' phenology predictions for varieties that it was not trained on (zero-shot).

Predicting grape phenology is crucial for many management operations in viticulture (wine-grape growing) such as disease management, frost protection, pruning, irrigation and harvest decisions. Additionally, phenology models can be used to estimate the future suitability of a grape variety in a region affected by climate [2], [9]. It can therefore allow farmers to immediately adapt to future climate conditions. Physical phenology models have been developed for up to 65 grape varieties, depending on the specific stage predicted [2], [4], [10]. However, current grapevine diversity is estimated somewhere between 1300 - 1700 varieties [7]. Instead we propose to extract the variety embeddings from a general grape recommendation system. To ensure robustness in low-data regimes, we reduce these embeddings to a 3-dimensional latent space. Here we present results of probing these embeddings

2835-3420/26/$31.00 ©2026 IEEE
DOI 10.1109/SDS70563.2026.00030

— specifically their linear relationship to physical forcing thresholds (F*) — and their effect on predicting phenology. Through this we are able to predict phenology competitively even for varieties that the RF models have never seen. Thereby we essentially create a pathway towards a general phenology model for every possible grape variety, at least in theory, as the limits of the current system will yet need to be tested.

METHODS

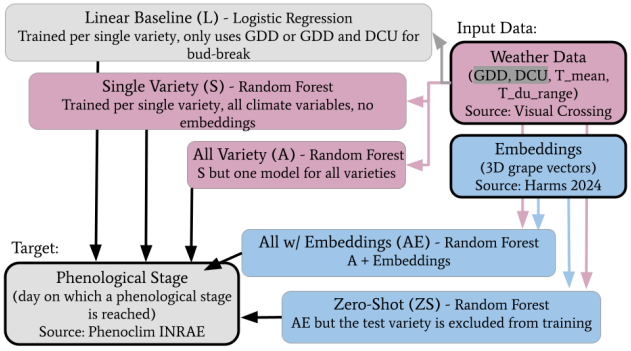

Fig. 1. The main training setup and data sources summarized.

The experimental framework utilizes five distinct modeling configurations to isolate the performance gains attributed to ensemble architectures versus latent variety representations. As illustrated in Figure 1, these include:

- Linear Baseline (L): A logistic regression model utilizing only traditional thermal-sum inputs (GDD or GDD and DCU for bud-break) to serve as a controlled, linear benchmark.
- Single-Variety (S): Random Forest (RF) models trained and tested on individual varieties using all climate features but no embeddings.
- All-Variety (A): A pooled RF model trained on the entire multi-variety dataset without variety-specific identifiers.
- All-Variety with Embeddings (AE): A pooled RF model that incorporates 3D variety embeddings as input features.
- Zero-Shot (ZS): A "leave-one-variety-out" configuration where the model is tested on a variety entirely excluded from the training set, relying solely on embeddings to generalize latent traits.

All models utilize a binary classification setup designed for operational online forecasting. For a window of ± 30 days around a recorded phenological event, the models predict whether the stage has been reached ($y = 1$) or not ($y = 0$). This produces a daily probability score, which we convert into a discrete date prediction by identifying the first calendar day where $P \geq 0.5$. Error is then calculated as the Root Mean Square Error ($RMSE$) between this predicted date and the observed ground truth. The embeddings are extracted from the grape variety recommender systems developed by [7], and were supplied for this purpose by the author. To ensure a balanced feature-to-sample ratio, the original 256-dimensional variety embeddings were compressed into a 3D representation. This was achieved via a symmetric bottleneck autoencoder (3-layer MLP) trained to reconstruct the original variety vectors provided by the latent space of Harms' models [7]. This ensures the 3D manifold retains maximum eco-physiological signal while preventing the high-dimensional latent vectors from drowning out the physical climate features during Random Forest training, since we only have 4 climate features. Aside from the ZS regime, all models utilize a consistent 80-20 train-test split.

The phenological data for the main stages of grape phenology were collected from the Phenoclim Agroclim INRAE database accessible through the TEMPO Portal [9]. The database contains data for 11 grape varieties across the 4 stages with most data: Bud-break, 50% flowering, 50% maturity and Maturity. We filter out only obvious outliers and erroneous phenology data manually and collect weather data for years 1970-2010 from Visual Crossing [11]. The prediction parameters calculated are the accumulated growing degree days (GDD), the accumulated daily Chilling Units (DCU), the mean temperature of the last 7 days (T_mean) and the average diurnal temperature range over the last 7 days (T_du_range). Most physical models use some combination of Chilling units and GDD but few consider the day and night-time temperatures which have important effects on phenology [12], [13]. We use DCU as an abstraction for the usual Utah chilling units used [13]. We apply the same addition and subtraction operation but utilize the mean temperature and only do one calculation per day. Multiple methods to calculate Chilling Units exist in literature, in this case we choose a method that allows us to use daily data for all metrics. Missing temperature values are linearly interpolated but years where frequent data was missing (most years 1972) the years were omitted. This led to a total dataset size of 1511 unique situations (stage x variety x year x location). We select a window of ± 30 days around when the stage occurred, this becomes our training data with 61 days per situation (30 negative (49%) and 31 positive (51%) samples). This leads to a total dataset size of 92171 samples. Data splits are done by situation, not sample, so that test data is not accidentally leaked.

DISCUSSION AND RESULTS

Prediction Improvement

The performance of our phenological models across five regimes is summarized in Table 1. To provide context, we include Physical Baselines (PB) representing averaged $RMSE$ across all varieties reported for a selection of papers on calibrated mechanistic models (one $RMSE$ value per paper) [2], [4], [9]. These PB values represent theoretical performance targets; however, they are not directly comparable to the current study as they typically rely on significantly larger datasets per variety. Consequently, we implemented a local linear classification baseline (L) to establish a direct internal benchmark. Utilizing only traditional thermal-sum inputs on our data, the L baseline represents the linear performance floor

(6.76 to 12.05 days $RMSE$) expected from standard viticultural threshold models on our dataset. Generally, acceptable model performance for bud-break and flowering is considered < 7 days, with 7–10 days for maturity predictions [2], [9], [14]. As shown in Table 1, the L benchmark, but even single-variety (S) and pooled (A) Random Forest models, struggle to consistently meet these benchmarks. Standard Random Forest configurations (S and A) demonstrate that moving to an ensemble architecture with additional climate variables, provides minor benefits. Including variety embeddings (AE); however, results in a significant accuracy gain. The AE model improves performance by 32% on average compared to the A setup and dramatically outperforms the L baseline, reducing error by up to 53%, for maturity prediction. This suggests that the non-linear integration of climate features and latent variety traits captures critical biological signals that simple linear thresholding cannot resolve. We further evaluated the model's ability to generalize to unseen cultivars via the Zero-shot (ZS) experiment. If the variety embeddings—originally extracted from a recommendation system—successfully encoded eco-physiological traits, the ZS model should maintain predictive power without variety-specific training data. Our results confirm this: the ZS model outperforms the linear baseline by an average of about 34%, additionally it also outperforms both the S and A regimes by 19% on average. Both AE and ZS models rival some of the best known PB's in most stages by taking advantage of the unified dataset. While ZS performance for late-season maturity ($RMSE = 10.2$ days) indicates that further architectural refinement is needed to rival highly calibrated physical models in that stage, without variety specific data, its superior performance in early stages like Bud-break and Flowering (3.6–5.9 days) highlights the potential for a general-purpose phenology framework. This demonstrates that variety embeddings effectively solve the "variety start-up" problem by mapping unknown cultivars onto a shared, biologically meaningful embedding. Further architectural optimizations are still needed before this is applied in practice.

TABLE I
COMPARISON OF PHYSICAL MODELS (PB) TO SHOW THEORETICAL MAXIMUM PERFORMANCE [2], [4], [9], WITH A LINEAR PHYSICAL BASELINE (L) ON OUR DATASET AND RF: S: SINGLE VARIETAL, A: ALL VARIETALS TREATED EQUALLY, AE: ALL VARIETALS WITH ENCODINGS, ZS: ZERO-SHOT. THE BEST PERFORMANCE ON MULTI-GRAPE DATASETS FOUND IS UNDERLINED AND BOLD.

Stage	PB	L	S(A)	AE	ZS
Bud-Break	7.1	9.3	9.2 (6.3)	**5.8**	5.9
Flowering	3.8-6.1	6.8	4.6 (4.7)	3.9	**3.6**
Ripening	**4.9**-6.0	10.8	7.3 (7.1)	5.9	7.0
Maturity	7.8-9.8	12.1	12.0 (13.0)	**5.6**	10.2

A. Latent space probing

Since it seems that the encodings do hold important characteristics that can be used by the RF models we further explore how much information about specific variety phenology is contained in the embeddings. We test whether the embeddings

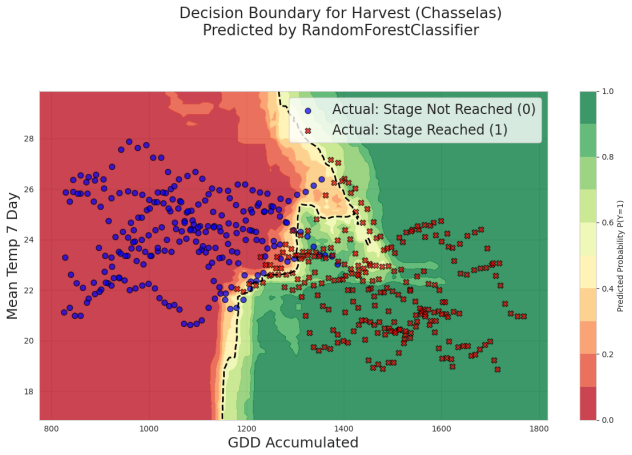

Fig. 2. Displaying the ripeness probability of Chasselas (also known as Fendant or Gutedel) as calculated by the RF model with encodings trained on all varieties. Only Chasselas data-points from the test-set are shown.

are able to be linearly fit to the physically derived forcing unit thresholds F* in general. We create a linear model that maps from the embeddings to the F* values for Bud-break, Flowering and Veraison according to the published data in [10]. We achieve a R^2 score of 0.69 showing that the encodings hold valuable information about the physical characteristics of the varieties. To see how far this can be pushed we also apply the encodings to predicting the F* of a larger dataset of detailed sugar loading requirements for each grape variety [4]. We compare the prediction to a prediction without encodings ($R^2 = 0.22$), as well as random encodings ($R^2 = 0.25$), to ours ($R^2 = 0.29$). The linear models did not perform well on these tasks but the improvement that was possible by using our encodings and their significance to the predicted variable ($p = 1.4 * 10^{-5}$) shows that some aspects of the complex sugar loading dynamics of grapes, which are inherently harder to predict [4], are captured in the encoding. It also, however, is shown that especially the reduced 3-dimensional embeddings may still require further improvement to more clearly capture precise climate-variety characteristics. We argue that such an effort is valuable as it potentially provides great advantages for vineyard managers as will be discussed in the next section.

B. ML models improve Vineyard Management

The development of agricultural models should not be the ultimate objective; rather, these models must serve a practical purpose for growers by enabling informed and timely interventions. Compared to traditional physical models, Machine Learning (ML) approaches offer significant advantages, provided that data scarcity challenges can be addressed. By utilizing a Random Forest (RF) classification approach, the model avoids returning a simple binary threshold. Instead, it provides a class score that serves as a proxy for prediction uncertainty and phenological progression. Because vineyard development is inherently continuous rather than binary, these

scores provide a more realistic representation of field conditions.

Furthermore, these scores capture non-linear relationships across all key features. Unlike many physical models, our approach incorporates diurnal dynamics [12] and accounts for short-term temperature fluctuations (e.g., the last 7 days) that may accelerate or delay development, rather than relying solely on seasonal accumulated variables. The models achieved area-under-the-precision-recall-curve (AUC-PR) values between 0.98 and 0.99. This indicates that the models are well-calibrated—neither over- nor under-confident—maximizing their utility despite inherent data uncertainty.

Figure 1 illustrates the probability of ripeness for the Chasselas variety relative to two RF input parameters, with test-set data points overlaid. The visualization demonstrates that the Growing Degree Day (GDD) threshold shifts based on recent temperature trends. Notably, the model reports non-zero scores when uncertain, signaling to growers when field scouting is most valuable. Given that scouting is labor-intensive and often inexact, maximizing the efficiency of each effort is critical. Incorporating current scouting data into ensemble modeling scores is technically straightforward, making ensemble models an ideal choice for modernizing vineyard management. By using variety encodings, we overcome the primary bottleneck of ML in agriculture: the ability to train on diverse datasets. This allows for accurate predictions even for varieties not present in the training set, enabling growers to evaluate the suitability of new varieties [9] even when local data is sparse.

Conclusion

Agricultural models must serve as tools for action. By integrating 3D variety embeddings into an ML framework, we get excellent phenology prediction performance that rivals comparable physical baselines. This approach overcomes the primary bottleneck of agricultural machine learning: the requirement for variety-specific calibration. Unlike traditional models that only provide a binary prediction for a specific date, our machine learning approach calculates a probability score that mirrors how grapes actually grow—as a gradual transition rather than a sudden switch. Because these scores update automatically based on recent weather and not just the thermal-sums, they are better able to incorporate current model certainty about the plants' progress. For a grower, this means they no longer have to guess when to send workers into the vineyard for manual inspections. Instead, they can prioritize their labor and only perform field scouting when the model indicates a high likelihood that a critical development stage is being reached. The primary contribution of this work; however, is the demonstration of zero-shot performance. The ability to make accurate predictions for varieties not present in the training set enables growers to evaluate new varieties without years of local historical data. This establishes a potential path toward general-purpose grape models that leverage a shared base of data to eliminate the need for individual variety calibration.

Acknowledgment

JZH would like to thank Vintality for the ability to publish this research that was developed in collaboration with Vintality, and financially supported by Vintality. JZH and CM were supported by the National Research Council of Canada Industrial Research Assistance Program (NRC IRAP). NKN was supported by funding from the Canadian Sustainable Agriculture Partnership (CSAP) Program (Agriculture and Agri-Food Canada) (Viticulture Cluster).

References

[1] B. Yang et al., "AgriGPT: a Large Language Model Ecosystem for Agriculture," Aug. 12, 2025, arXiv: arXiv:2508.08632. doi: 10.48550/arXiv.2508.08632.

[2] I. Morales-Castilla et al., "Diversity buffers winegrowing regions from climate change losses," Proc. Natl. Acad. Sci., vol. 117, no. 6, pp. 2864–2869, Feb. 2020, doi: 10.1073/pnas.1906731117.

[3] E. M. Wolkovich, I. García de Cortázar-Atauri, I. Morales-Castilla, K. A. Nicholas, and T. Lacombe, "From Pinot to Xinomavro in the world's future wine-growing regions," Nat. Clim. Change, vol. 8, no. 1, pp. 29–37, Jan. 2018, doi: 10.1038/s41558-017-0016-6.

[4] A. K. Parker et al., "Temperature-based grapevine sugar ripeness modelling for a wide range of Vitis vinifera L. cultivars," Agric. For. Meteorol., vol. 285–286, p. 107902, May 2020, doi: 10.1016/j.agrformet.2020.107902.

[5] T. Chen et al., "Empowering Agrifood System with Artificial Intelligence: A Survey of the Progress, Challenges and Opportunities," ACM Comput Surv, vol. 57, no. 2, p. 49:1-49:37, Nov. 2024, doi: 10.1145/3698589.

[6] P. Aniruddha, A. S. Valiyaparambil, and K. Sornalakshmi, "Evaluating Multimodal Fusion Strategies for Resilient Agricultural Sensing Systems," in Data Science and Applications, S. J. Nanda, R. P. Yadav, M. Prasad, and M. Saraswat, Eds., Cham: Springer Nature Switzerland, 2026, pp. 249–259. doi: 10.1007/978-3-032-10783-1_19.

[7] J. Harms, "Development of a generative AI-based model for guiding grape variety selection under contemporary climate dynamics," 2024, Accessed: Feb. 20, 2025. [Online]. Available: https://escholarship.mcgill.ca/concern/theses/z316q7030

[8] S. Yin, Y. Xi, X. Zhang, C. Sun, and Q. Mao, "Foundation Models in Agriculture: A Comprehensive Review," Agriculture, vol. 15, no. 8, p. 847, Jan. 2025, doi: 10.3390/agriculture15080847.

[9] A. Parker et al., "Adaptation to climate change by determining grapevine cultivar differences using temperature-based phenology models," Oct. 2020, Accessed: Dec. 15, 2025. [Online]. Available: https://hdl.handle.net/10182/13559

[10] A. k. Parker, I. g. De Cortázar-Atauri, C. Van Leeuwen, and I. Chuine, "General phenological model to characterise the timing of flowering and veraison of Vitis vinifera L.," Aust. J. Grape Wine Res., vol. 17, no. 2, pp. 206–216, 2011, doi: 10.1111/j.1755-0238.2011.00140.x.

[11] Visual Crossing Corporation. Visual Crossing Weather. 1971-2008, https://www.visualcrossing.com/. Accessed November 27, 2025.

[12] L. Meng et al., "Divergent responses of spring phenology to daytime and nighttime warming," Agric. For. Meteorol., vol. 281, p. 107832, 2020.

[13] Y. Mo, J. Zhang, H. Jiang, and Y. H. Fu, "A comparative study of 17 phenological models to predict the start of the growing season," Front. For. Glob. Change, vol. 5, Jan. 2023, doi: 10.3389/ffgc.2022.1032066.

[14] S. Reis et al., "Grapevine Phenology in Four Portuguese Wine Regions: Modeling and Predictions," Appl. Sci., vol. 10, no. 11, p. 3708, Jan. 2020, doi: 10.3390/app10113708.