Supplement of Biogeosciences, 22, 7687–7708, 2025
https://doi.org/10.5194/bg-22-7687-2025-supplement

Biogeosciences Open Access EGU

Supplement of

Utilizing probability estimates from machine learning and pollen to understand the depositional influences on branched GDGT in wetlands, peatlands, and lakes

Amy Cromartie et al.

Correspondence to: Amy Cromartie (aec277@cornell.edu)

Supplement

Supplementary Figures:

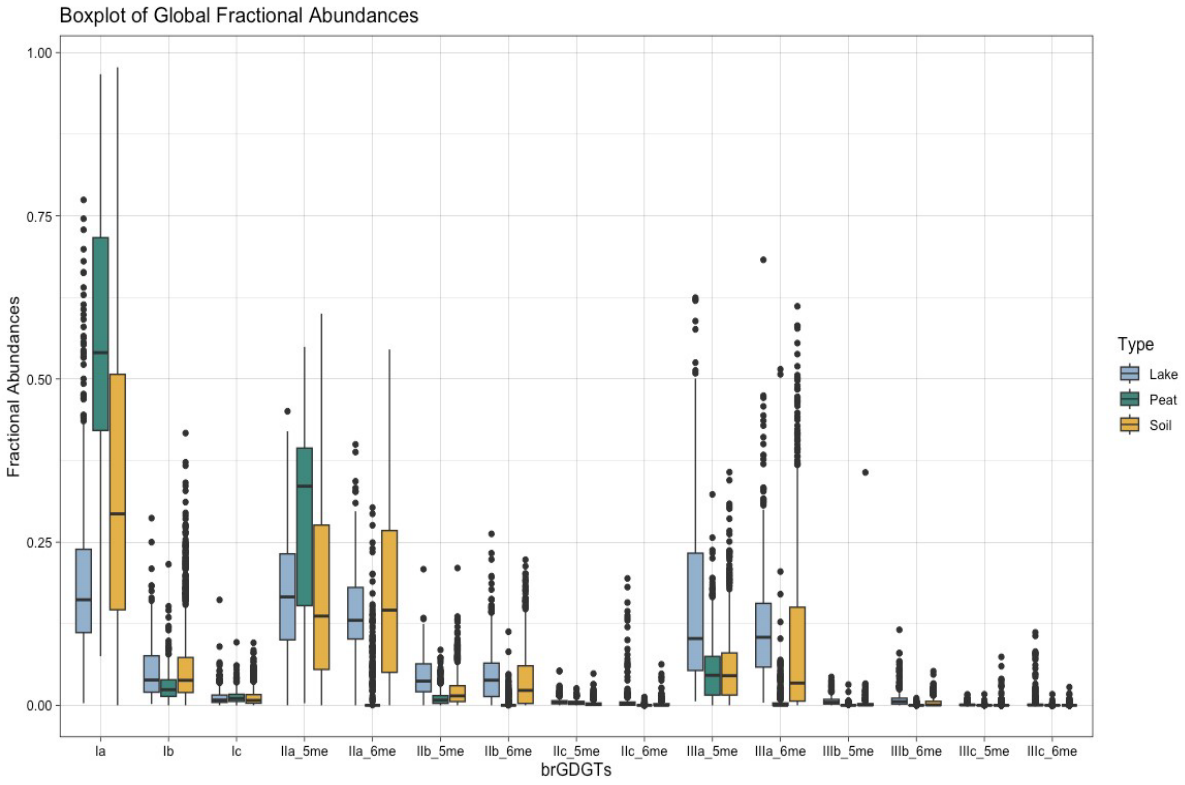
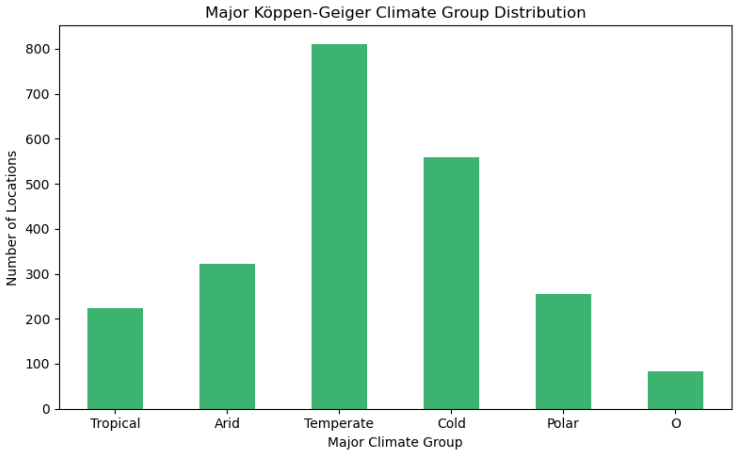
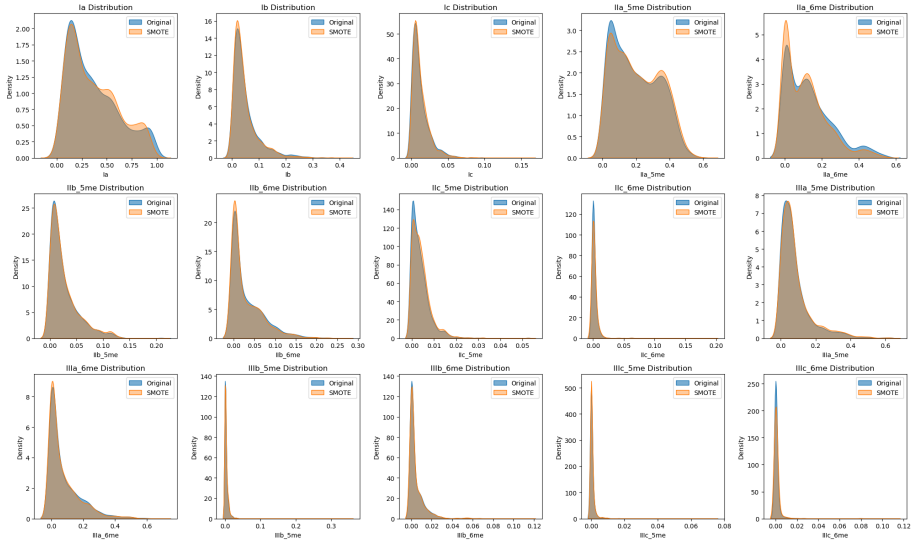


Figure S1. Box plot of the distribution of fractional abundances on the global brGDGT database classified by type (lake, peat, soil)



15

Figure S2. Köppen-Geiger climate distribution of samples in the brGDGT database created with the kgcpy (Yu et al. 2024) library in Python. O are samples without a label or coordinates assigned.



20

Figure S3. Distribution of brGDGTs between the original and smote datasets. Overlapping distributions suggests that bias was not introduced with the SMOTE samples.

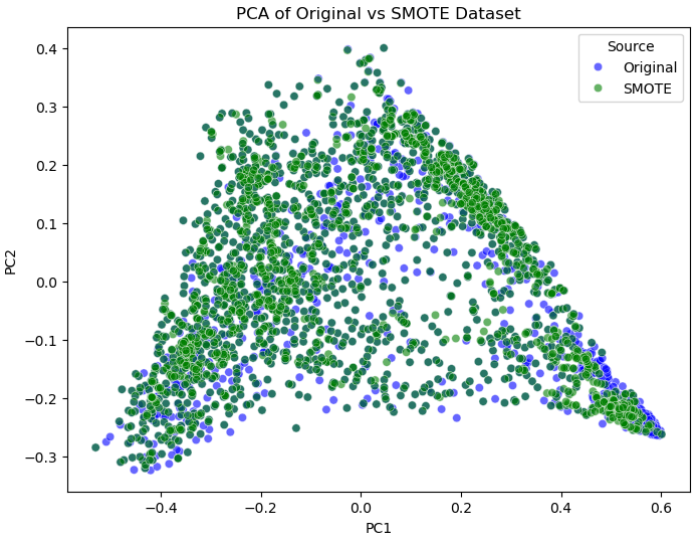


Figure S4. PCA distribution of the original and SMOTE datasets

A.	Original class distribution:	
	SampleTypeNum	
	1 0.507672	
	0 0.259097	
	2 0.233231	
	Name: proportion	dtype: float64
	SMOTE class distribution:	
	SampleTypeNum	
	0 0.333333	
	1 0.333333	
	2 0.333333	
	Name: proportion	dtype: float64

B.	Kolmogorov-Smirnov Test between original and SMOTE datasets (for each feature):	
	Ia: KS statistic = 0.0323	p = 1.8211e-01
	Ib: KS statistic = 0.0277	p = 3.4250e-01
	Ic: KS statistic = 0.0401	p = 5.0102e-02
	IIa_5me: KS statistic = 0.0300	p = 2.5261e-01
	IIa_6me: KS statistic = 0.0742	p = 6.9874e-06
	IIb_5me: KS statistic = 0.0161	p = 9.2414e-01
	IIb_6me: KS statistic = 0.0525	p = 3.6958e-03
	IIc_5me: KS statistic = 0.0658	p = 1.0150e-04
	IIc_6me: KS statistic = 0.0094	p = 9.9994e-01
	IIIa_5me: KS statistic = 0.0297	p = 2.6297e-01
	IIIa_6me: KS statistic = 0.0555	p = 1.7547e-03
	IIIb_5me: KS statistic = 0.0315	p = 2.0594e-01
	IIIb_6me: KS statistic = 0.0129	p = 9.9015e-01
	IIIc_5me: KS statistic = 0.0616	p = 3.4764e-04
	IIIc_6me: KS statistic = 0.0190	p = 7.9730e-01

Table S1. A counts of each class type in the dataset B. Kolmogorov-Smirnov Test between original and SMOTE datasets (for each brGDGT feature in the dataset)

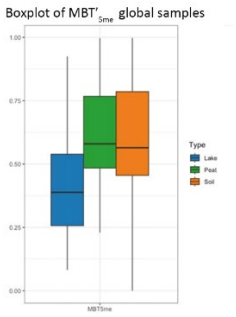


Figure S5. Boxplot of indexes utilized in paper on the global modern database

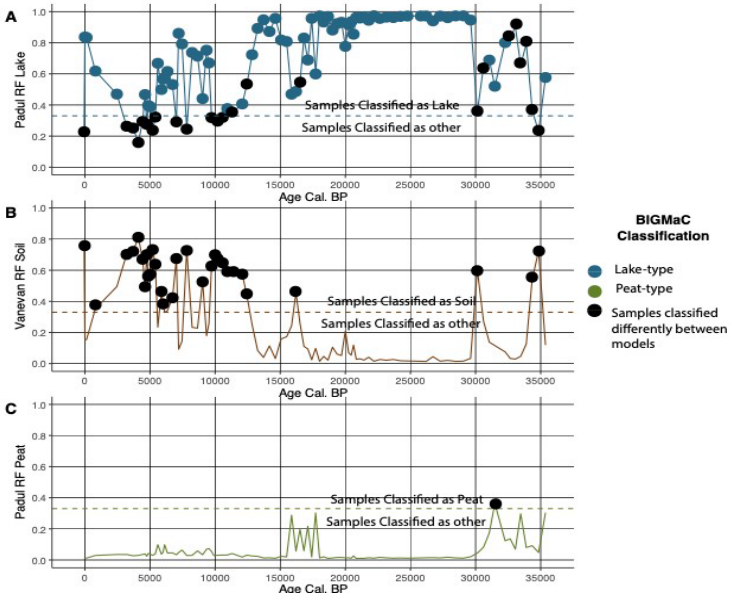


Figure S6. Probability estimates of brGDGTs from the Padul record from this article compared to the BIGMaC classification model of Martínez-Sosa et al. (2023). Lines are the probability estimates and the

dots are the classification from the BIGMaC. Colors are blue = lake, brown = soil, and green = peat. Black dots are samples that are not in agreement between models.

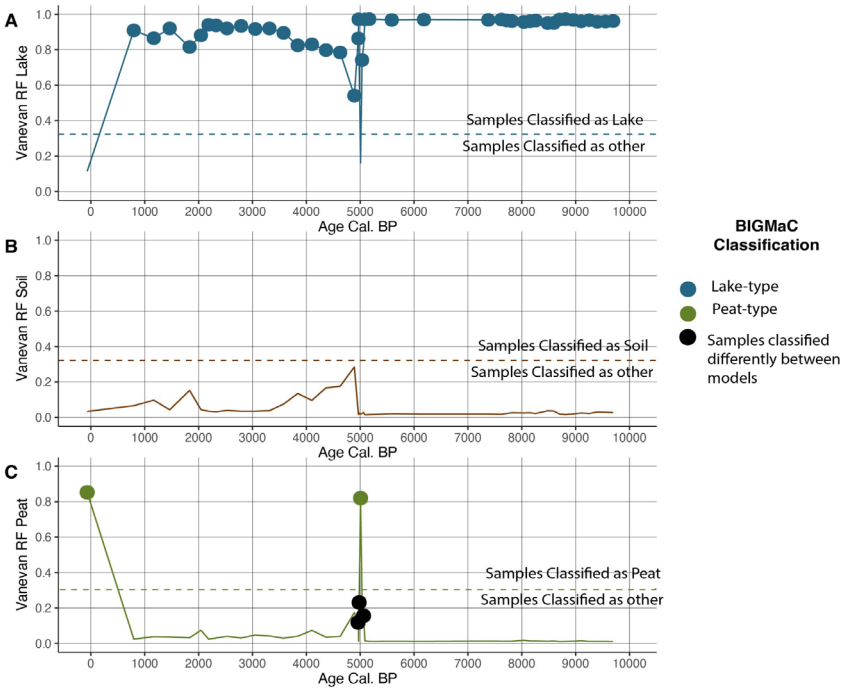


Figure S7. Probability estimates of brGDGTs from the Vanevan record from this article compared to the BIGMaC classification model of Martínez-Sosa et al. (2023). Lines are the probability estimates and the dots are the classification from the BIGMaC. Colors are blue = lake, brown = soil, and green = peat. Black dots are samples that are not in agreement between models.

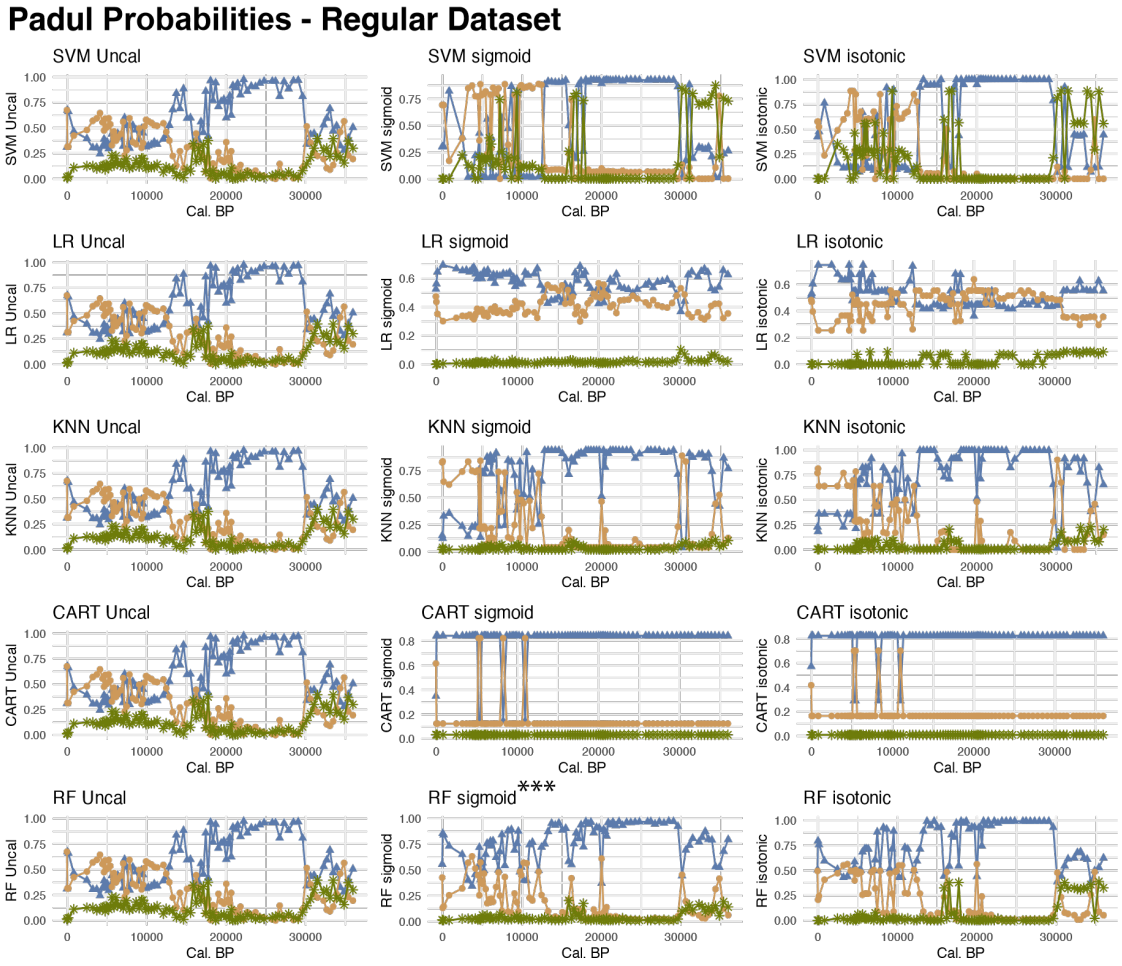
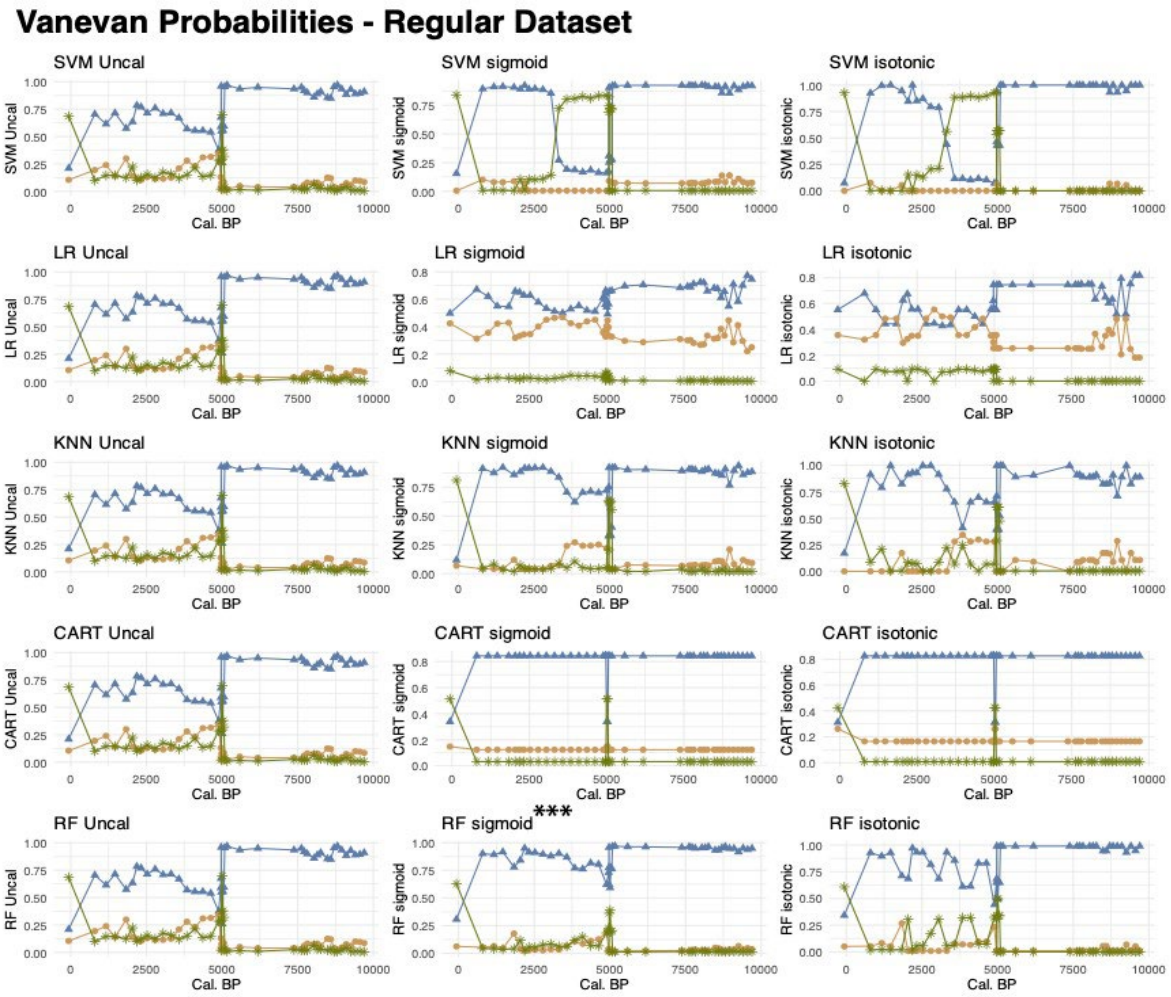
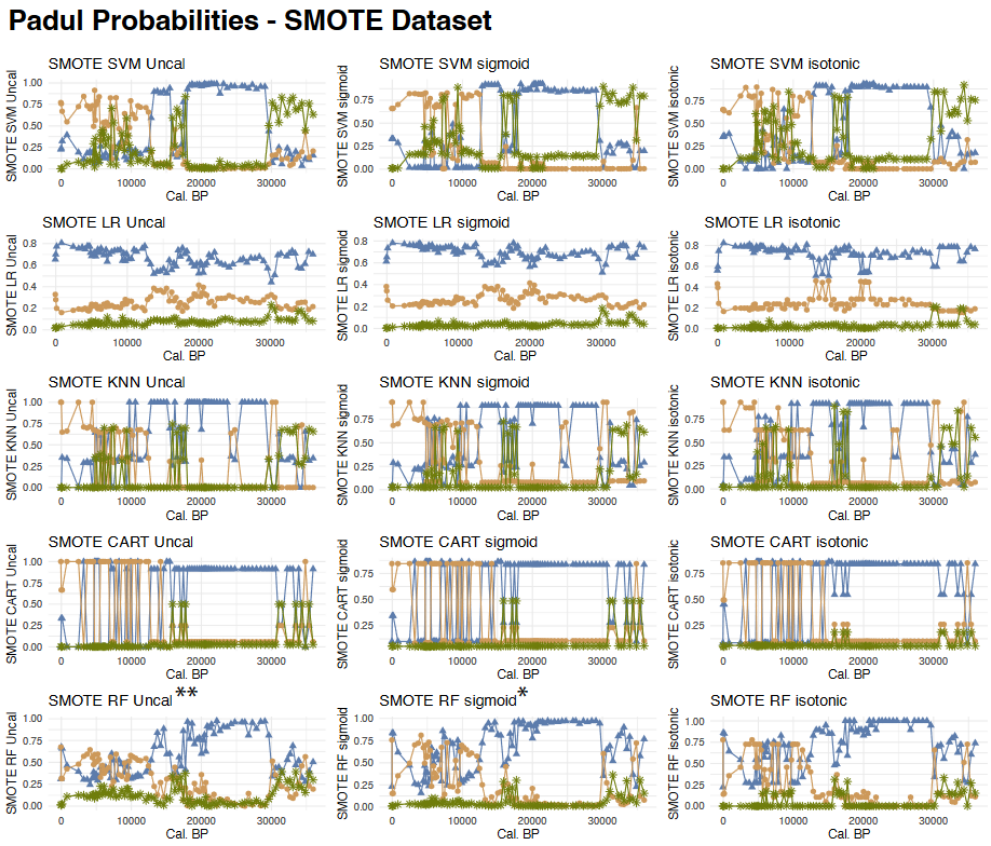


Figure S8. Probability estimates for the Padul record on models tested in the paper on the regular unsampled dataset



50

Figure S9. Probability estimates for the Vanevan record on models tested in the paper on the regular unsampled datasets.



55

Figure S10. Probability estimates for the Padul record on models tested in the paper on the resampled datasets.

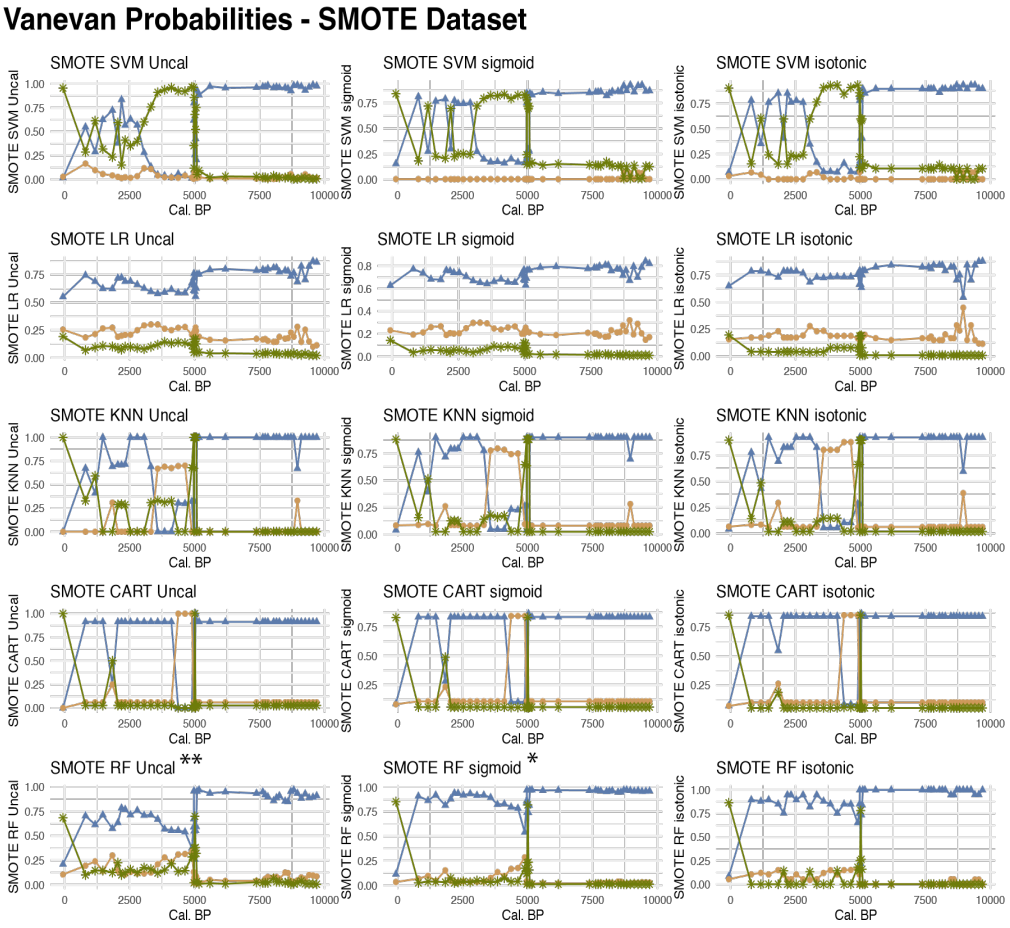


Figure S11. Probability estimates for the Vanevan record on models tested in the paper on the resampled datasets.

References:

Martínez-Sosa, P., Tierney, J. E., Pérez-Angel, L. C., Stefanescu, I. C., Guo, J., Kirkels, F., Sepúlveda, J., Peterse, F., Shuman, B. N., and Reyes, A. V.: Development and Application of the Branched and Isoprenoid GDGT Machine Learning Classification Algorithm (BIGMaC) for Paleoenvironmental Reconstruction, Paleoceanography and Paleoclimatology, 38, https://doi.org/10.1029/2023PA004611, 2023.

Yu, X., Ascencio, J., and French, R.: Open-Source Climate Classification Package: kgcPy, in: 2024 IEEE 52nd Photovoltaic Specialist Conference (PVSC), 1085–1085, 2024.