







**BioSpace25 - Biodiversity insight from Space** 10 - 14 February 2025 | ESA-ESRIN | Frascati - Italy

### **How do Earth Observation Foundation Models Help to Predict Multi-Trophic Soil Biodiversity**

CERNA Selene, SI-MOUSSI Sara, MIELE Vincent, **THUILLER Wilfried** 

> Laboratoire d'Ecologie Alpine **Grenoble**, France









In Artificial Intelligence

→ THE EUROPEAN SPACE AGENCY

**CNrS** 

#### Why predicting soil biodiversity?

To address critical threats such as:

- Land use change and intensification. ٠
- Desertification ٠
- Increased levels of pollution.
- Climate change.

	LAND USE for			_			
REVIEW	нив	About the Hub 🗸	About Land Use & Net Zero 🗸	Resources			
Received 25 Apr 2015   Accepted 9 Oct 2015   Published 23 Nov 2015							
Extinction risk of soil biota	3. Soil biodiversity is likely in decline, but it is key to successful above ground biodiversity						
Stavros D. Veresoglou <sup>1,2</sup> , John M. Halley <sup>3</sup> & Matthias C. Rillig <sup>1,2</sup>		ICLE	OPEN	Check for			
Identifying potential threats to so biodiversity	il Cli	1 Climate change and cropland management compromise soil integrity and multifunctionality					
Mark Tibbett, Tandra D. Fraser and Sarah Duddigan	Marie Simon Thom Carlos	Sünnemann () <sup>1,2 M</sup> , Remy Beugnor te Cesarz () <sup>1,2</sup> , Arwyn Jones <sup>7</sup> , Anik as Reitz () <sup>1,5</sup> , Matthias C. Rillig () <sup>8</sup> 4. Guerra () <sup>1,2</sup> , <sup>1,1,2</sup> & Nico Eisenh	<sup>13,4</sup> , Claudia Breitkreuz <sup>5,6</sup> , François Buscoto <sup>5,1</sup> a Lehmann <sup>8,9</sup> , Alfred Lochner <sup>1,2</sup> , Alberto Orgiazzi <sup>9</sup> , Martin Schädler <sup>0</sup> <sup>1,10</sup> , Linnea C. Smith <sup>1,11</sup> , Anja auero <sup>1,2,12</sup>	l, ∋ <sup>7</sup> , Zeuner <sup>1,2</sup> ,			

Université Grenoble Alpes

**€€**3 B@N



CNIS

• e e sa

Main goal of our study: To predict the abundance of 51 soil trophic groups in the French Alps

Université Grenoble Alpes

European

**€€**) B€N

CNIS

· e esa

#### Main goal of our study: To predict the abundance of 51 soil trophic groups in the French Alps

Challenges when building predictive models:



he largely **unknown** diversity of **soil organisms.**  How to overcome them in this study?





UGA

Université Grenoble Alpes

aropean

**€€**) B€N

CNrs

· e esa

CECS

#### Main goal of our study: To predict the abundance of 51 soil trophic groups in the French Alps



How to overcome them in this study?



UGA

Jniversité Grenoble Alpes

**€€** D B ® N

CNrs

· e esa

CECS

#### Main goal of our study: To predict the abundance of 51 soil trophic groups in the French Alps

UGA

**€€**D B@N



CNrs

· e esa

#### **Main goal of our study:** To predict the abundance of 51 soil trophic groups in the French Alps



How to overcome them in this study?



UGA

**€€**) B€N

Leverage pretrained Earth **Observation Foundation models** to extract belowground features.

> Integrate **tabular** data with remote sensing features.

CN

· e esa

Main goal of our study: To predict the abundance of 51 soil trophic groups in the French Alps

UGA

**€€**) B€N



#### → THE EUROPEAN SPACE AGENCY

CN

eesa

### **Data Collection**



UGA

Université Grenoble Alpes

 $\bigcirc$ 

**€€D** B@N

Resulting dataset size: Aprox. 1000 samples

CN

· e esa

# Methodology: Levarage EOF Models



**Proposed solution**: Leverage pretrained EOF models to extract soil features (embeddings)



#### SatDINO Foundation Model by META AI Research:

Source: <u>https://github.com/facebookresearch/HighResCanopyHeight</u> Paper: <u>https://arxiv.org/abs/2304.07213</u> (Tolan et al.)

#### Why SatDINO?

- Model architecture: ViT.
- Training strategy: DINOv2 technique (self-supervised learning).

UGA

Université Grenoble Alpes

**€€**D B@N

 Dataset: forests, mountainous terrains, and high degree of tree biodiversity (high-resolution RGB images).



CESS

Overview of SatDINO approach for predicting canopy height. Author: Tolan et al.

CN

· e esa

#### Dynamic One-For-All (DOFA) Foundation Model:

Source: <u>https://huggingface.co/XShadow/DOFA</u> Paper: <u>https://arxiv.org/abs/2403.15356</u> (Xiong et al.)

#### Why DOFA?

- Model architecture: ViT.
- Training strategy: Masked modeling, wavelength-conditioned dynamic patch embedding, and multimodal distillation pretraining.
- Dataset: Setninel-1 (SAR), Sentinel-2 (multispectral), NAIP (RGB), EnMAP (hyperspectral).



DOFA's architecture emulating the neuroplasticity mechanism for processing multimodal EO data. Author: Xiong et al.

## **Methodology: Overview**





**€€**D B@N

Université Grenoble Alpes





CESS

CNrs

•eesa

Can embeddings match or **improve** the performance of models based on in-situ tabular data?

## Results: Performance of ML models

#### **ML techniques:**

- LGBM and Random Forest showed the best performances.
- LGBM is the fastest in all data configurations.



Université Grenoble Alpes

**€€**D B@N

LGBM	-> C1 and C2 = 0.71,	C3 and C4 = 6.37
RF	-> C1 and C2 = 2.43,	C3 and C4 = 35.61

12

CN

•eesa

# Results: Performance of ML models

#### **ML techniques:**

- LGBM and Random Forest showed the best performances.
- LGBM is the fastest in all data configurations.

Taxa:

- Bacteria, Fungi, and Protist show consistently higher Spearman's Rho correlation.
- Oligochaete and Insect perform poorly.



UGA

Université Grenoble Alpes

**€€**D B@N

Average execution time in sec: LGBM -> C1 and C2 = 0.71, C3 and C4 = 6.37 RF -> C1 and C2 = 2.43, C3 and C4 = 35.61 ATT-MLP -> C1 and C2 = 273.85, C3 and C4 = 354.83

CNrs

•eesa

CESS

## Results: Performance of ML models

#### **ML techniques:**

- LGBM and Random Forest showed the best performances.
- LGBM is the fastest in all data configurations.

Taxa:

- Bacteria, Fungi, and Protist show consistently higher Spearman's Rho correlation.
- Oligochaete and Insect perform poorly.

#### Data configurations:

- C2 generates the best performance.
- C3 (embeddings) captures relevant information but not enough to overcome C1 or C2 or C4... WHY?



UGA

Université Grenoble Alpes

**€€D** B@N

CN

·eesa

CECS

# Results: Evaluating Embeddings

#### PCA projection and visualization of land cover clustering based on the SatDINO embeddings.



 SatDINO captures some level of differentiation based on land cover types.

UGA

**€€**D B@N

 Grassland and Shrubs might have inherently similar visual features, making them harder to separate in the embedding space.

15

CNrs

· e esa

# **Results: Evaluating Embeddings**

#### Relationship between in-situ tabular data and PCA components of SatDINO embeddings.



UGA

**€€**) B€N

The color gradients in the embeddings align with patterns in the tabular data components (redundancy!). This explains why adding embeddings (configuration 4) doesn't improve performance.

16

CNrs

· e esa

### **Results: Environmental Drivers**

- Climate and soil are • main contributors in most cases.
- Soil variables • compensate the absence of embeddings in C2.
- Landscape ٠ contributions remain modest in most cases and phenology's relevance is highly group-specific.

C2 🔍		Climate 💻 S	Feature Group Soil 💼 Pheno	os logy 💼	Landscape		
B_Chemolithoautotroph	23%		41%		27%	8%	0.74
B_Copiotroph	23%		45%		19%	13%	0.63
B_Copiotropholigotroph	33%		37%		20%	10%	0.71
B_Oligotroph	- 4	10%	28%		20%	12%	0.65
B_Osmotroph	14%		60%		13%	13%	0.44
B_Photolithoautotroph	21%		45%		21%	12%	0.73
B_Phytoparasite	21%	1	39%	16	6	24%	0.64
B_Saprotroph	16%		53%		17%	14%	0.70
B_Zooparasite	14%		58%		13%	14%	0.81
F_Animalparasite	31%		45%		179	6 7%	0.49
F_Arbuscularmycorrhizal	24%		36%		22%	18%	0.67
F_Ectomycorrhizal	10%	20%	37%		33%		0.84
F_Lichenized	17%	23%		43%		16%	0.54
F_Lichenparasite	28%		35%		30%	8%	0.27
F_Littersaprotroph	22%		40%		24%	14%	0.41
F Mycoparasite	29%		38%		21%	12%	0.29
F Othersaprotroph	29%		37%		19%	15%	0.63
F Plantpathogen	18%	29%		28%		24%	0.60
F Rootendophyte	25%	29	196		33%	14%	0.64
F Soilsaprotroph	11%	29%	21%		39%		0.78
F Woodsaprotroph	26%		46%		15%	12%	0.40
P Animalparasite	26%		33%		28%	13%	0.53
P Bacterivor	20%		55%		17%	8%	0.60
P Fungivor	16%	32%			2	10%	0.55
P Mixotroph	21%		48%		22%	10%	0.53
P Omnivor	17%	1	62%			14% 6%	0.62
P Photoautotroph	26%		42%		21%	10%	0.51
P Plantparasite	23%		44%		22%	11%	0.52
P Protistivor	25%		42%		22%	11%	0.57
P Saprotroph	21%	35	8	22		21%	0.30
O Endoanecic	19%	1	47%		21%	14%	0.41
O Endogeic	25%		35%		25%	15%	0.23
O Epianecic	23%		50%		20%	8%	0.34
	20%		57%		153	6 8%	0.37
O Intermediate	24%		44%		25%	7%	0.25
L Detritivor	22%		41%		28%	9%	0.31
Europer	22%		40%		30%	8%	0.10
L Omnivor	20%		47%		26%	5%	0.22
L Dhutenhage	25%		40%		2010	1196	0.21
I_Prodotor	2.570		40%		26%	10%	0.25
I_Predator	2470		4079		2070	1070	0.25
M_Arachnidspredators	40/ 40/		04N		2070	070	0.37
M_NAnimalparasites	200		36%		220/	1.26	0.13
M_NBacterivores	2070		4600		22.00	10%	0.43
M_NFungivores	1694		4070		10%	10%	0.45
M_NOmeiveres	21%		40%		15%	12%	0.38
M_NOmnivores	21%		49%		15%	15%	0.28
M_NPredators	27%		38%		25%	10%	0.35
C_Epigeicanimal	21%		44%		24%	11%	0.44
C_Epigeicplant	24%		41%		26%	10%	0.37
C_Euedaphic	27%		35%		23%	17%	0.59
C_Hemiedaphic	16%	31%			46%	6%	0.11
		R	leiative importa	nce			spearman's khô



UGA

Université Grenoble Alpes

uropean

**€€**D B@N

17

CNrs

· e e sa

### **Results: Environmental Drivers**

٠

٠

٠

most cases.

Soil variables

absence of

Landscape

group-specific.

C2 🔍 C4 🕉 🗱 Feature Groups Feature Groups Phenology 📥 Landscape 📥 Orthophoto Climate Soil Soil Phenology Landscape **B\_Chemolithoautotroph** B Chemolithoautotroph 8% B Copiotroph B\_Copiotroph 1.3% B\_Copiotropholigotroph 0.63 B\_Copiotropholigotroph 10% 12% B Oligotroph 0.61 B\_Oligotroph 0.65 13% B Osmotroph B Osmotroph 0.44 Climate and soil are B\_Photolithoautotroph B Photolithoautotroph **B\_Phytoparasite** B Phytoparasite main contributors in B Saprotroph B\_Saprotroph 0.68 **B\_Zooparasite** B Zooparasite F\_Animalparasite F Animalparasite F\_Arbuscularmycorrhizal F\_Arbuscularmycorrhizal 0.62 F Ectomycorrhizal F\_Ectomycorrhizal F Lichenized 0.56 F\_Lichenized F Lichenparasite 0.31 F\_Lichenparasite 0.27 F\_Littersaprotroph 0.40 F\_Littersaprotroph F Mycoparasite 0.36 F\_Mycoparasite 0.29 F\_Othersaprotroph 0.62 F Othersaprotroph 0.63 F\_Plantpathogen 0.61 F\_Plantpathogen 0.60 compensate the 0.63 F Rootendophyte F\_Rootendophyte F\_Soilsaprotroph F\_Soilsaprotroph F Woodsaprotroph 0.40 F\_Woodsaprotroph 0.37 P\_Animalparasite 0.53 P\_Animalparasite 0.52 P Bacterivor 0.60 P\_Bacterivor embeddings in C2. P\_Fungivor P\_Mixotroph 0.52 P Omnivor P\_Omnivor P Photoautotroph 0.51 P\_Photoautotroph P Plantparasite 0.53 P Plantparasite 11% P\_Protistivor 0.53 P Protistivor P\_Saprotroph P Saprotroph 0.29 0.30 contributions remain O Endoanecic O\_Endoanecic 0.44 O\_Endogeic 0.21 O Endogeic 0.23 O Epianecic 0.34 O\_Epianecic 0.31 modest in most cases O\_Epigeic 0.36 O\_Epigeic 0.3 0.25 O\_Intermediate 0.25 O\_Intermediate 0.33 and phenology's I Detritivor I Detritivo 0.07 I\_Fungivor 0.10 I\_Fungivor 0.22 I\_Omnivor 0.15 I Omnivor relevance is highly I\_Phytophage 0.21 I\_Phytophage 0.26 0.23 I Predator 0.25 I Predator 0.37 M Arachnidspredators 0.36 M Arachnidspredators 0.13 M NAnimalparasites 0.10 M NAnimalparasites M\_NBacterivores M\_NBacterivores 0.42 M NFungivores 0.43 M NFungivores M\_NHerbivores M\_NHerbivores 0.54 M NOmnivores 0.28 M\_NOmnivores 0.23 15% M NPredators 10 0.35 M NPredators 0.39 C\_Epigeicanimal 11% 0.44 C\_Epigeicanima 0.40 C\_Epigeicplant 10% 0.37 C\_Epigeicplant 17% C\_Euedaphic C\_Euedaphic 0.08 C\_Hemiedaphic 0.11 C Hemiedaphic Spearman's Rho Relative Importance Spearman's Rho Relative Importance

UGA

Université Grenoble Alpes

**€€**D B@N

CNrs

· e e sa

CECS

## **Results: Environmental Drivers**

	C2 🖳	- Clin	Feature Groups mate 🚃 Soil 💼 Phenolog	y 💻 Landscape			C4 🕉 🧱	Feature Groups Climate Soil Phenology Landsca	e 🗕 Ortho	photo	
	B_Chemolithoautotroph	23%	41%	27%	8%	0.74	B_Chemolithoautotroph	22% 19% 35%	8%	16%	0.71
	B_Copiotroph	23%	45%	19%	13%	0.63	B_Copiotroph	17% 31% 24%	8%	20%	0.62
	B_Copiotropholigotroph	33%	37%	20%	10%	0.71	B_Copiotropholigotroph	39% 18%	31%	7% 4%	0.63
	B_Oligotroph	40%	28%	20%	12%	0.65	B_Oligotroph	18% 13% 21% 25%		4%	0.61
	B_Osmotroph	14%	60%	13%	13%	0.44	B_Osmotroph	19% 39% 15%	12%	15%	0.40
	B_Photolithoautotroph	21%	45%	21%	12%	0.73	B_Photolithoautotroph	25% 19% 22%	16%	17%	0.69
	B_Phytoparasite	21%	39%	16%	24%	0.64	B_Phytoparasite	14% 17% 20% 3		14%	0.62
1	B_Saprotroph	16%	53%	17%	14%	0.70	B_Saprotroph	24% 32% 17%	11%	16%	0.68
	B_Zooparasite	14%	58%	13%	14%	0.81	B_Zooparasite	31% 28% 15%	14%	12%	0.77
	F_Animalparasite	31%	45%		% 7%	0.49	F_Animalparasite	35% 34%	13% 3%	15%	0.47
_	F Arbuscularmycorrhizal	24%	36%	22%	18%	0.67	F_Arbuscularmycorrhizal	20% 17% 27%	24%	11%	0.62
	F_Ectomycorrhizal	10% 20%	37%	3	%	0.84	F_Ectomycorrhizal	3% 8% 42%	45%	2%	0.85
_ L	F Lichenized	17%	23%	43%	16%	0.54	F_Lichenized	12% 7% 54%	13%	14%	0.56
	F_Lichenparasite	28%	35%	30%	8%	0.27	F_Lichenparasite	24% 33%	7% 51	% 12%	0.31
	F_Littersaprotroph	22%	40%	24%	14%	0.41	F_Littersaprotroph	10% 20% 25% 25%		21%	0.40
	F_Mycoparasite	29%	38%	21%	12%	0.29	F_Mycoparasite	24% 27% 26%	4%	18%	0.36
	F_Othersaprotroph	29%	37%	19%	15%	0.63	F_Othersaprotroph	42% 22%	3% 7%	16%	0.62
	F_Plantpathogen	18%	29%	28%	24%	0.60	F_Plantpathogen	3/%	28%	13%	0.61
	F_Rootendophyte	25%	29%	33%	14%	0.64	F_Rootendophyte	26% 12% 35%	13%	13%	0.63
	F_Soilsaprotroph	11% 29	9% 21%	39%		0.78	F_Soilsaprotroph	4% 10% 29%	1%	3%	0.79
	F_Woodsaprotroph	26%	46%	15%	12%	0.40	F_Woodsaprotroph	14% 52%	11% 5%	19%	0.37
	P_Animalparasite	26%	33%	28%	13%	0.53	P_Animalparasite	32% 15% 34%	7%	12%	0.52
	- P Bacterivor	20%	55%	1	6 8%	0.60	P Bacterivor	31% 20% 19%	9%	(2%)	0.57
. L	P_Fungivor	16%	32%	42%	10%	0.55	P_Fungivor	28% 18% 35%	83/	9 11%	0.51
	GMixotroph	21%	48%	22%	10%	0.53	B Demokroph		070	20%	0.52
	P_Omnivor	17%	62%	210	14% 6%	0.62		2370 2270 2170	775	2175	0.57
	P_Photoautotroph	20%	42%	21%	10%	0.51	G P Plantnarasite		796	1.5%	0.51
	P_Plantparasite	2370	4470	2470	1170	0.52		15% 47%	7.4%	586 086	0.53
	P_Protistivor	2370	9470	2270	2144	0.30	P_Protistivor	164. 204. 204.	16%	15%	0.29
	P_Saprotroph	10%	47%	2270	14%	0.30		26% 15% 30%	7%	22%	0.44
	O_Endoanecic	25%	266	2170	15%	0.23	O Endogaic	19% 8% 34% 9%	309	*	0.21
	O Epianocic	23%	50%	200	8%	0.34	O Enjanecic	13% 19% 42%	5%	22%	0.31
	O_Epianecic	20%	576	20	0% M 0%	0.34	O Epigeic	22% 34% 11%	14%	19%	0.36
	0 Intermediate	24%	44%	25%	7%	0.25	0 Intermediate	20% 25% 28%	5%	22%	0.25
	L Detritivor	22%	41%	28%	9%	0.31	L Detritivor	20% 15% 30%	28	8	0.33
	L Eungivor	22%	40%	30%	8%	0.10	L Eungivor	13% 35% 25%	5%	22%	0.07
	L Omnivor	20%	47%	26%	6%	0.22	L Omnivor	52% 14%	23%	1% 10%	0.15
	L Phytophage	25%	40%	24%	11%	0.21	I Phytophage	37% 14% 16%	% 2	7 5	0.26
	I Predator	24%	40%	26%	10%	0.25	I Predator	31% 11% 19% 12%	2	7 5	0.23
	M Arachnidspredators	35%	31%	26%	8%	0.37	M Arachnidspredators	23% 30% 30	6	11%	0.36
<u>г</u>	M NAnimalparasites	4% 6%	84%		6%	0.13	M NAnimalparasites	3% 2% 83%		1% 11%	0.10
	M_NBacterivores	28%	26%	32%	13%	0.49	M_NBacterivores	16% 15% 41%	12%	16%	0.50
	 M_NFungivores	22%	46%	22%	10%	0.43	 M_NFungivores	28% 18% 19% 8	21	8'	0.42
	M_NHerbivores	16%	53%	19%	12%	0.56	M_NHerbivores	24% 21% 24%	10%	21%	0.54
	M_NOmnivores	21%	49%	15%	15%	0.28	M_NOmnivores	23% 17% 13% 23%		2 596	0.23
	M_NPredators	27%	38%	25%	10%	0.35	M_NPredators	16% 19% 32%	6 28	B	0.39
	C_Epigeicanimal	21%	44%	24%	11%	0.44	C_Epigeicanimal	18% 15% 30% 1	6	2 %	0.40
	C_Epigeicplant	24%	41%	26%	10%	0.37	C_Epigeicplant	25% 23% 24%	9%	19%	0.37
_	C Euedaphic	27%	33%	23%	17%	0.59	C Euedaphic	42% 21%	13%	10%	0.60
<u>г</u>	C_Hemiedaphic	16%	31%	46%	6%	0.11	C_Hemiedaphic	6% 16% 54%	3%	20%	0.08
_			Relative Importance	e		Spearman's Rho		Relative Importance			Spearman's Pho

MA

Université Grenoble Alpes

\*\*\*\* \*\*\*\* European **€€**) B€N

- Climate and soil are main contributors in most cases.
- Soil variables compensate the absence of embeddings in C2.
- Landscape contributions remain modest in most cases and phenology's relevance is highly group-specific.

\*

CNIS

· e esa

### Conclusions

### **Conclusions:**

- There are **group-specific** prediction **challenges** across trophic groups.
- Tabular data outperforms embeddings.
- Embeddings capture partial information.
- Embeddings offer an alternative where in-situ data is scarce.

#### **Future Work:**



- Model trophic networks to uncover species interdependencies.
- Investigate the feasibility of incorporating highresolution remote sensing data.

#### Three Recommendations for the conference organizers:

**€€**D B@N

Traditional tabular environmental data remains essential for robust biodiversity modeling.

UGA

- Integrating higher-resolution hyperspectral/multispectral remote sensing data could refine environmental characterizations for biodiversity assessment.
- A coordinated evaluation of different EOF models across multiple research teams would facilitate standardized comparisons.

CN

eesa







### Thank you for your attention!



selene.cerna@univ-grenoble-alpes.fr











22