RELEO - REpresentation Learning for Earth Observation

J. Inglada, N. Dobigeon, M. Fauvel, S. Valero, T. Oberlin, J. Michel, S. Gürol

CESBIO (CNES/CNRS/INRAe/IRD/UPS), INPT - IRIT, ISAE, CERFACS, Toulouse, FRANCE

Land surface monitoring with satellites

Land surface monitoring - What we need

Essential Climate and Biodiversity variables

- Above-ground biomass, albedo, evaporation from land, fire, land cover, land surface temperature, leaf area index, soil carbon, soil moisture,
- River discharge, terrestrial water storage, glaciers, permafrost, snow
- Species distributions and abundances, physiology, phenology, primary productivity, ecosystem distribution, ecosystem vertical profile

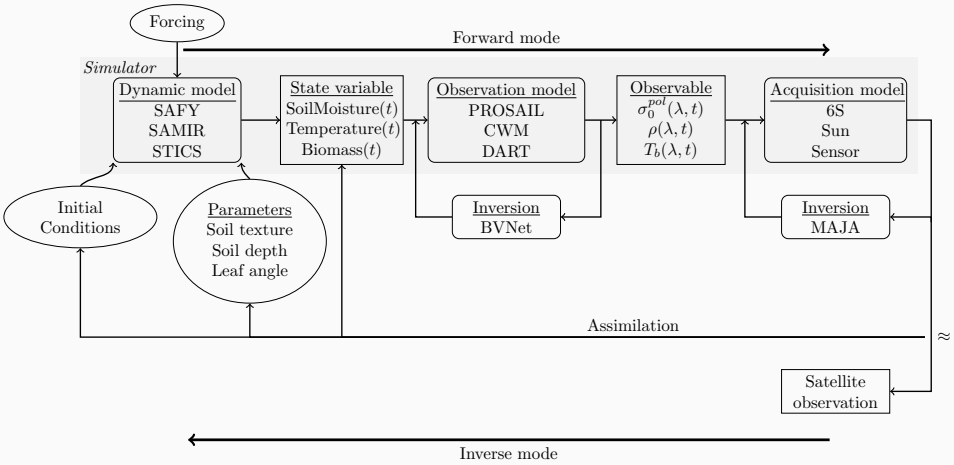
Monitoring

- Where: every point on land up to 10 m
- When: up to once a week
- How: on demand resolution (temporal, spatial)

Forecasting

- Realistic evolutions
- Scenario generation

Land surface monitoring with satellites - Data assimilation



A foundation model for land surface monitoring

Large Representation Models for EO - RELEO

Industrial Chair within ANITI IA Cluster

- 4 year project

Academic partners

- CNES (4 ½ grants)
- CESBIO (INRAE, Université de Toulouse)
- INPT - IRIT
- ISAE

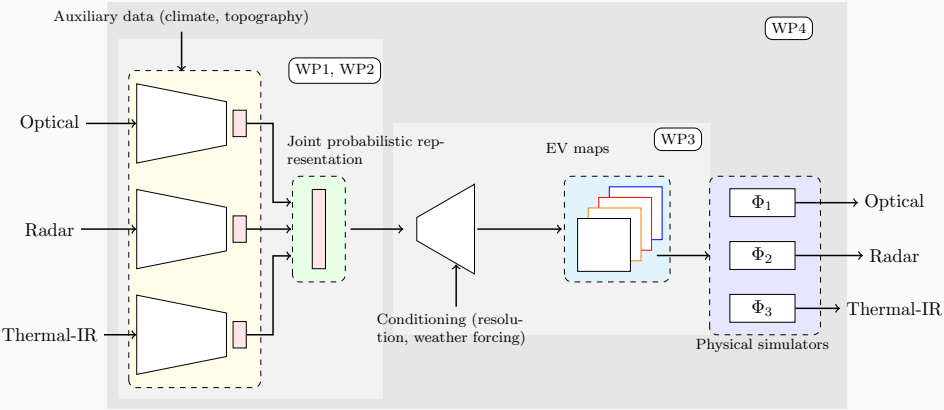
ANR - French National Research Agency

- 572 k€
- 1 + ½ PhD + 2 post-doc

Industrial partners

- CERFACS (½ grant + Eng.)
- CS Sopra Steria (½ grant + Eng.)
- Magellium (½ grant)
- Thales Alenia Space (½ grant)
- Thales Services Numériques (½ grant)

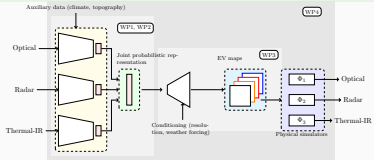
Large Representation Models for EO - RELEO



Existing works

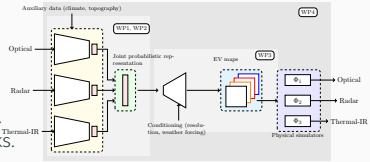
- Dumeur, I., Valero, S., & Inglada, J. (2024). Paving the way toward foundation models for irregular and unaligned satellite image time series. CoRR, http://arxiv.org/abs/2407.08448v1b.
- Dumeur, I., Valero, S., & Inglada, J. (2024). Self-supervised spatio-temporal representation learning of satellite image time series. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, (), 1–18. http://dx.doi.org/10.1109/jstars.2024.3358066
- Yoël Zérah, Valero, S., & Inglada, J. (2024). Physics-constrained deep learning for biophysical parameter retrieval from sentinel-2 images: inversion of the prosail model. Remote Sensing of Environment, 312(), 114309. http://dx.doi.org/10.1016/j.rse.2024.114309
- Zérah, Yoël, Valero, S., & Inglada, J. (2023). Physics-driven probabilistic deep learning for the inversion of physical models with application to phenological parameter retrieval from satellite times series. IEEE Transactions on Geoscience and Remote Sensing, 61(), 1–23. http://dx.doi.org/10.1109/tgrs.2023.3284992
- Bellet, V., Fauvel, M., Inglada, J., & Michel, J. (2023). End-to-end learning for land cover classification using irregular and unaligned sits by combining attention-based interpolation with sparse variational gaussian processes. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, (), 1–16. http://dx.doi.org/10.1109/jstars.2023.3343921
- Baudoux, L., Inglada, J., & Mallet, C. (2022). Multi-nomenclature, multi-resolution joint translation: an application to land-cover mapping. International Journal of Geographical Information Science, 37(2), 403–437. http://dx.doi.org/10.1080/13658816.2022.2120996
- Michel, J., Vinasco-Salinas, J., Inglada, J., & Hagolle, O. (2022). Sen2venμs, a dataset for the training of Sentinel-2 super-resolution algorithms. Data, 7(7), 96. http://dx.doi.org/10.3390/data7070096
- Inglada, J., Michel, J., & Hagolle, O. (2022). Assessment of the usefulness of spectral bands for the next generation of Sentinel-2 satellites by reconstruction of missing bands. Remote Sensing, 14(10), 2503. http://dx.doi.org/10.3390/rs14102503

How RELEO will work



- Self-Supervised Learning: RELEO will primarily use self-supervised learning, where the model learns from the data itself without extensive manual labeling.
- Multi-Modal Data: The system will process data from multiple satellites (different sensors) to capture complementary information across space and time.
- Sensor genericity: The system will be able to perform inference on data from sensors not seen during training.

Key Steps



1. Pre-training: Blend physical models and DL to pre-train models that:
   - Learn representations related to EVs (e.g., carbon content, vegetation indices).
   - Create “AI-ready data” (generic embeddings) that can be used for various tasks.
2. Decoding & Forecasting: Use physics-guided DL to condition generic embeddings for:
   - Retrieving (estimating current values) and forecasting EVs.
   - Estimating the uncertainties in those forecasts.
3. Data Assimilation: Combine satellite observations with pre-trained model forecasts to ensure continuous monitoring.
4. Continual Learning: Continuously update models with new EO data and data from new satellites.
5. Account for Long-Term Trends: Handle changes over time that go beyond the initial training data.

Expected Outcomes & Scientific Questions

- Deliverables:
  - Improved ability to monitor EVs with greater accuracy and confidence.
  - A system that can handle diverse, irregular EO data.
  - A flexible AI-ready data system that can be adapted to new tasks.
  - Contribute to improved understanding of long-term environmental changes.
- Scientific Questions:
  - How to best combine probabilistic representation learning and physical knowledge.
  - How to efficiently use pre-trained models.
  - How to continually update models with new data and sensors.

What is missing in current works

What we have learned so far

When we say multi-modal...

- OK, we don't have video, audio, tweets, kitties, ...

What we have learned so far

When we say multi-modal...

- OK, we don't have video, audio, tweets, kitties, ... but we have: visible, infrared, thermal, microwave, lidar ... in 4D (x, y, z, t, λ)

What we have learned so far

When we say multi-modal...

- OK, we don't have video, audio, tweets, kitties, ... but we have: visible, infrared, thermal, microwave, lidar ... in 4D (x, y, z, t, λ)

We have plenty of open data for pre-training, but building good pre-training datasets is not easy

- Use all available spectral bands, with appropriate levels of corrections, look for temporal co-occurrences
- Add auxiliary data: weather, climate, topography, etc.

What we have learned so far

When we say multi-modal...

- OK, we don't have video, audio, tweets, kitties, ... but we have: visible, infrared, thermal, microwave, lidar ... in 4D (x, y, z, t, λ)

We have plenty of open data for pre-training, but building good pre-training datasets is not easy

- Use all available spectral bands, with appropriate levels of corrections, look for temporal co-occurrences
- Add auxiliary data: weather, climate, topography, etc.

Latent spaces are not $\mathcal{N}(0, I)$

- Physical magnitudes may be bounded, have assymetric distributions, be correlated, ...

What we have learned so far

When we say multi-modal...

- OK, we don't have video, audio, tweets, kitties, ... but we have: visible, infrared, thermal, microwave, lidar ... in 4D (x, y, z, t, λ)

We have plenty of open data for pre-training, but building good pre-training datasets is not easy

- Use all available spectral bands, with appropriate levels of corrections, look for temporal co-occurrences
- Add auxiliary data: weather, climate, topography, etc.

Latent spaces are not $\mathcal{N}(0, I)$

- Physical magnitudes may be bounded, have assymetric distributions, be correlated, ...

We need fully differentiable physical simulators...

What we have learned so far

When we say multi-modal...

- OK, we don't have video, audio, tweets, kitties, ... but we have: visible, infrared, thermal, microwave, lidar ... in 4D (x, y, z, t, λ)

We have plenty of open data for pre-training, but building good pre-training datasets is not easy

- Use all available spectral bands, with appropriate levels of corrections, look for temporal co-occurrences
- Add auxiliary data: weather, climate, topography, etc.

Latent spaces are not $\mathcal{N}(0, I)$

- Physical magnitudes may be bounded, have assymetric distributions, be correlated, ...

We need fully differentiable physical simulators... and all models are wrong!

What we have learned so far

When we say multi-modal...

- OK, we don't have video, audio, tweets, kitties, ... but we have: visible, infrared, thermal, microwave, lidar ... in 4D (x, y, z, t, λ)

We have plenty of open data for pre-training, but building good pre-training datasets is not easy

- Use all available spectral bands, with appropriate levels of corrections, look for temporal co-occurrences
- Add auxiliary data: weather, climate, topography, etc.

Latent spaces are not $\mathcal{N}(0, I)$

- Physical magnitudes may be bounded, have assymetric distributions, be correlated, ...

We need fully differentiable physical simulators... and all models are wrong!

Pre-training strategies from NLP or CV may not work...

- Pretext tasks need the right amount of difficulty to be pertinent.

What we have learned so far

When we say multi-modal...

- OK, we don't have video, audio, tweets, kitties, ... but we have: visible, infrared, thermal, microwave, lidar ... in 4D (x, y, z, t, λ)

We have plenty of open data for pre-training, but building good pre-training datasets is not easy

- Use all available spectral bands, with appropriate levels of corrections, look for temporal co-occurrences
- Add auxiliary data: weather, climate, topography, etc.

Latent spaces are not $\mathcal{N}(0, I)$

- Physical magnitudes may be bounded, have assymetric distributions, be correlated, ...

We need fully differentiable physical simulators... and all models are wrong!

Pre-training strategies from NLP or CV may not work...

- Pretext tasks need the right amount of difficulty to be pertinent.

Most published models in EO are not validated on meaningful downstream tasks

- We want to accurately map the continental biosphere at high resolution and with uncertainty estimation.
- Hallucination is not an option 🙂

Datasets

Table 1: Description of Sentinel-2 SITS data-sets used to pretrain large SSL models. DEM stands for Digital elevation model, LOC for longitude latitude coordinates, DW for Dynamic World Land Cover classes, LS for Landsat and WV for Worldview very high resolution images.

Data-Set Name	Data	Temporal Extent	Temporal length	Acquisitions	Geographical extent	ROI size	Available	Cloud filter
SSL4EO-S12	S2, S1	2020	1 year	4, (1/season)	Worldwide	264×264	✓	yes, ≤ 10 %
Presto Data-Set	S2, S1, DEM, ERA5, DW, LOC	2020-2021	2 years	24 (1/month)	Worldwide	1×1		yes
Prithvi data-set	NASA HLS[1] V2 L30	?	?	?	USA	64×64		yes
SkySense data-set	S2, S1, WV	?	?	~ 65	Wordlwide	$64 \times \times 64$		yes, ≤ 1 %
Clay data-set[2]	LS(8,9), S2, S1, NAIP[3], LINZ[4]	2018-2023	2 years	8, (1/quarter)	Worldwide	224×224	✓	yes

[1] Harmonized Landsat Sentinel-2 data-set

[2] https://clay-foundation.github.io/model/release-notes/data_sampling.html

[3] https://eos.com/find-satellite/naip/

[4] https://basemaps.linz.govt.nz/

Datasets

RGB is not realistic data for us

L1C + L2A + NDVI is not multi-modal

Sentinel-2 NIR and Landsat 9 NIR are not the same band

Sensor and solar angles are more useful meta-data than a text caption describing the image content

Realistic downstream tasks for evaluation

Table 2: Description of the downstream tasks with high resolution EO data (i.e. Sentinel or Landsat) employed to assess the different so-called Foundation Models. Tasks relevant to land monitoring are sorted based on the number of inputs SITS (1 or 2) and on whether the temporal information is exploited (TS yes, No T no). Tasks with SITS with less than 6 acquisitions are not-considered as tasks exploiting the temporal information. "?" corresponds to unknown information. Total tasks refers to the total number of downstream tasks presented in the original study.

	1, TS	1, No T	2, TS	2, No T	Total tasks
Skysense	3			2	16
Presto	3				5
Prithvi		4		1	5
Clay	?	?	?	?	?
DOFA		2			12
Spectral GPT		1		1	4

Summary

Summary

- Land surface monitoring: EVs at 10m, once a week + on-demand resolution + forecasting.
  - We don't do CV on EO data

Summary

- Land surface monitoring: EVs at 10m, once a week + on-demand resolution + forecasting.
  - We don't do CV on EO data
- Self-supervised learning by inclusion of physical models into a VAE-like architecture.
- Multi-modal data: multiple satellites (different sensors) even those not seen during training.
- Probabilistic representations for uncertainty evaluation in retrieval and forecast.

Summary

- Land surface monitoring: EVs at 10m, once a week + on-demand resolution + forecasting.
  - We don't do CV on EO data
- Self-supervised learning by inclusion of physical models into a VAE-like architecture.
- Multi-modal data: multiple satellites (different sensors) even those not seen during training.
- Probabilistic representations for uncertainty evaluation in retrieval and forecast.
- Most published FM for EO are not validated on meaningful tasks *for our goal*.
- Most available training datasets are simplified, partial or not realistic *for our goal*.

Acknowledgements

- Yoël Zerah
- Iris Dumeur
- Ekaterina Kalinicheva
- Kevin De Sousa
- Julien Prissimitzis
- Richard Faucheron
- Sasha Troncy-Portier