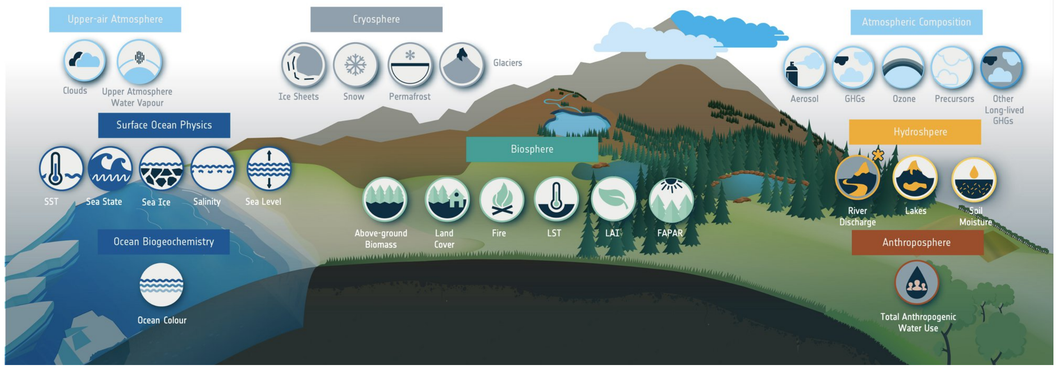
Learning Generic Probabilistic Latent Representations for Multi-Modal Earth Observation Forecasting and Reconstruction

Richard Faucheron[1,2,4,*] Silvia Valero[1] Selime Gürol[3,4] Vincent Poulain[5]

[1] *CESBIO, Univ Toulouse/CNES/CNRS/INRAE/IRD, Toulouse, France,* [2] *CNES,* [3] *CERFACS,* [4] *CECI, Univ Toulouse/CERFACS/CNRS/IRD,*
[5] *Thales Services Numériques,* * *Correspondence: richard.faucheron@utoulouse.fr*

Context and Motivation

Recent Earth Observation (EO) systems provide critical data for monitoring Essential Climate and Biodiversity Variables (EVs), supporting climate change mitigation and adaptation. However, retrieving EVs from multi-source EO data remains challenging due to indirect observations of interest variables, varying spatial resolutions, and irregularly sampled time series.

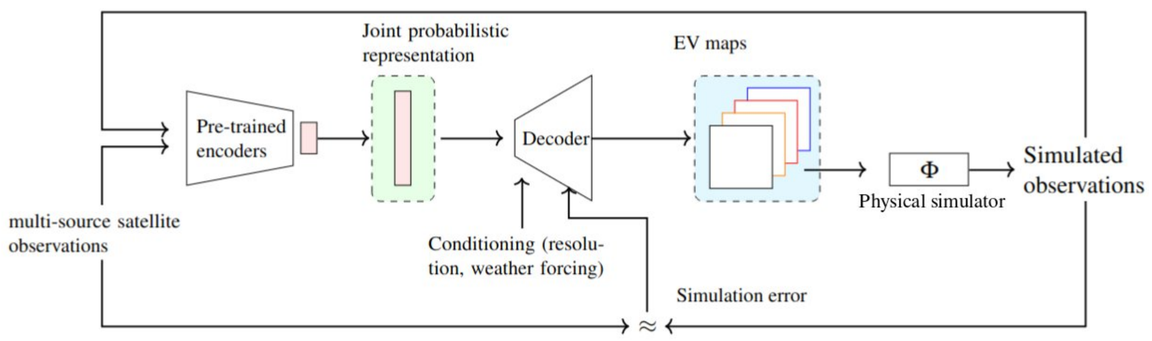


While deep learning (DL) models show promise in extracting complex patterns, they often lack physical consistency and interpretability, struggle with irregular and unaligned sampling, and require large labeled datasets, resources that are scarce and noisy in EO. Additionally, most methods underutilize the complementarity of diverse sensors, limiting their ability to capture the complexity of the Earth system. Finally, the reliability of EO-derived information depends on both precise representations and uncertainty quantification. Yet, existing foundation models typically provide deterministic embeddings without addressing representation uncertainty.

RELEO Framework

The REpresentation Learning for Earth Observation (RELEO) project aims at developing new self-supervised representation learning methods that :

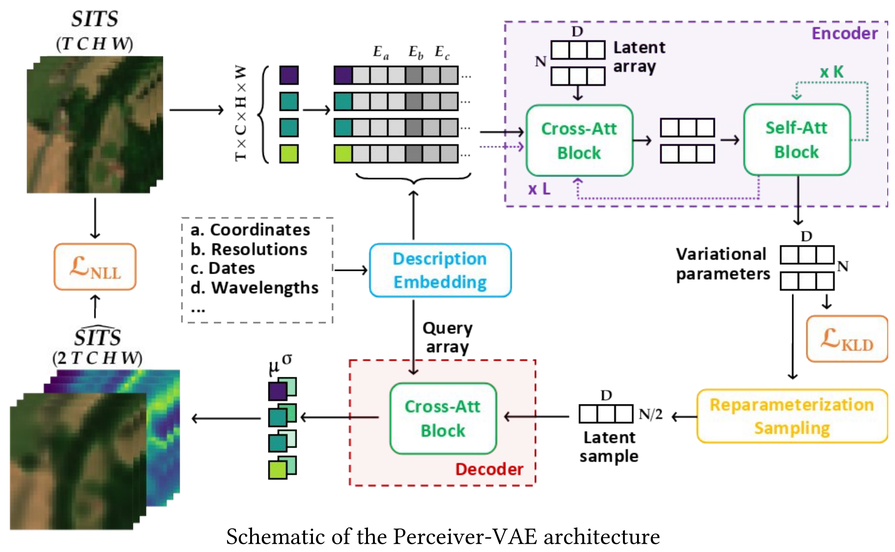
- Combine **physical priors and DL**
- Process **multi-modal** EO data to capture complementary spatio-temporal patterns
- Generate **task-agnostic, probabilistic, semantically meaningful embeddings**
- Use physics-guided decoding to retrieve and forecast EVs with **quantified uncertainties and improved interpretability**
- Account for long-term trends and ensure continuous land monitoring using **data assimilation** strategies and **continual learning**



Probabilistic Autoencoder

This work introduces **Perceiver-VAE**, a self-supervised framework that maps satellite image time series (SITS) into a probabilistic latent space, enabling uncertainty quantification and multi-modal data fusion.

- Input SITS are flattened into a pixel sequence, **augmenting each pixel with contextual metadata** embedded via Fourier positional encoding.
- The attention-based encoder maps the tokenized representation into a **fixed-dimensional probabilistic latent space**.
- The decoder **models explicitly the conditional distribution** of the input data given the latent sample, providing reconstructed observation and uncertainty estimates.



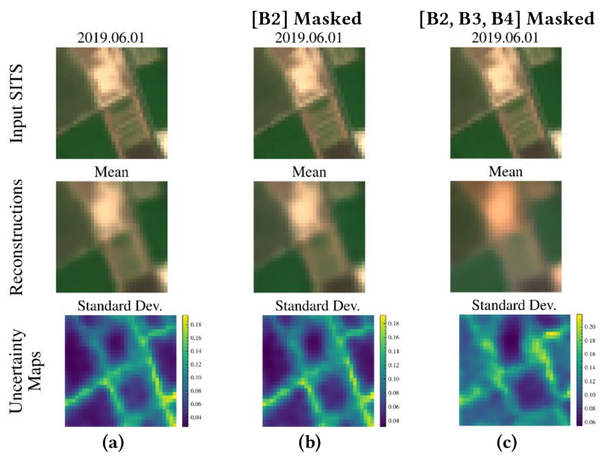
Schematic of the Perceiver-VAE architecture

The model processes batches of Sentinel-2 data, where each batch consists of 32x32 spatial patches spanning 3 dates and including all 10 spectral bands.

Meaningful Representations

The focus of this work is not on retrieving the sharpest reconstruction, but rather on **ensuring that the proposed model provides meaningful latent embeddings**, following some key criteria :

- Capture **essential features** and **dependencies** of the input data
- Reliable **uncertainty** estimates
- Latent space that **maximizes information** content with **compact dimensionality and low redundancy**

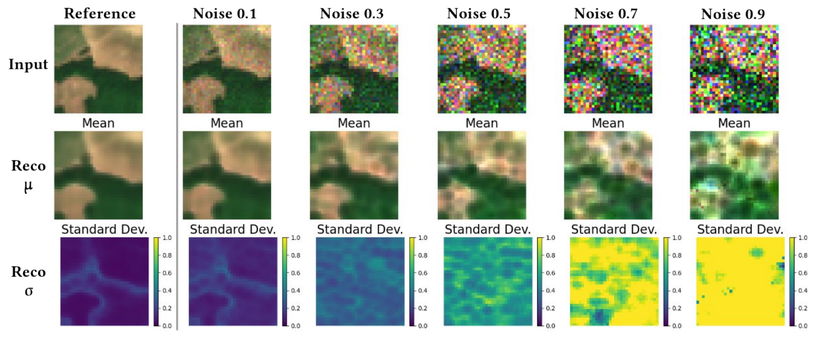


A model that captures essential features and their dependencies should produce **faithful reconstructions** and handle **missing data recovery**

→ Produces **coherent reconstructions** and **structured uncertainty estimates**

→ Despite degraded reconstruction quality, the model effectively **leverages remaining Sentinel-2 bands** to recover missing spectral information

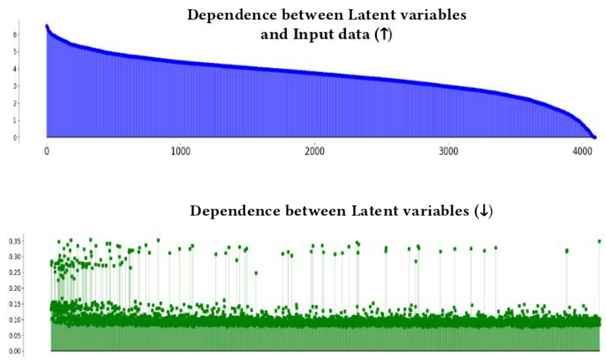
Uncertainty should **increase as input data quality degrades**, reflecting reduced confidence in reconstructions



→ High uncertainty in **high-constrast** areas

→ Uncertainty **increases gradually** as noise is progressively added to the input data during inference

An efficient latent space should be **compact and meaningful**, with **low redundancy** to ensure optimal input representation



→ Relatively **balanced distribution of information** encoded by latent variables, indicating effective latent space usage and limited posterior collapse

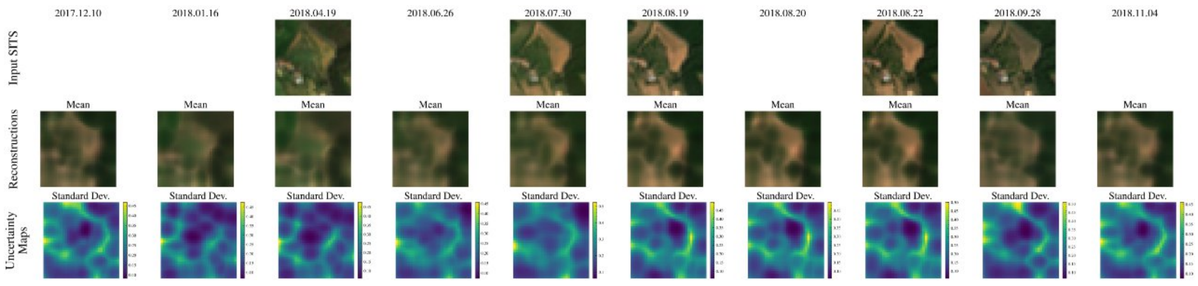
→ Latent variables exhibit **low inter-dependence**, capturing distinct, non-redundant features

Future direction : *Towards Multi-Modal Forecasting*

→ We assessed that Perceiver-VAE produces coherent reconstructions and uncertainty estimates, capturing essential features and dependencies of input SITS into a compact, informative latent space.

Future work will extend this framework into a probabilistic forecasting system for multi-modal EO data.

1. Input data will be organized into **spatio-temporal observation chunks (STOCs)**—contiguous batches of acquisitions within fixed-length spatial and temporal windows—initially integrating Sentinel-2, Sentinel-1, MODIS, and AgERA5
2. By encoding consecutive STOCs of the same area of interest using the Perceiver-VAE encoder, we will generate a **time-series of probabilistic embeddings**.
3. A **latent attention-based model** will then iteratively cross-attend to past STOCs, refining a **unified probabilistic state** that summarizes historical context. The most recent STOC may also incorporate weather forecasts to provide additional predictive context.
4. Finally, the Perceiver-VAE decoder will use this latent state to **reconstruct future satellite images with quantified uncertainty**



Reconstruction of a SITS. Top: Input SITS. Blank images indicate dates excluded from the input but targeted for reconstruction. Middle: Model's estimated reconstructed values. Bottom: Estimated reconstruction uncertainty. The model reconstructs missing dates and quantifies uncertainty coherently with input structures

References

- [1] I. Higgins et al., "beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework," presented at the International Conference on Learning Representations, Feb. 2017.
- [2] A. Jaegle et al., "Perceiver IO: A General Architecture for Structured Inputs & Outputs," Mar. 15, 2022, doi: 10.48550/arXiv.2107.14795.
- [3] S. Liu et al., "Discovering influential factors in variational autoencoders," Pattern Recognition, vol. 100, p. 107166, Apr. 2020, doi: 10.1016/j.patcog.2019.107166.

Acknowledgement

- Our work has benefitted from the AI Interdisciplinary Institute ANITI. ANITI is funded by the France 2030 program under the Grant agreement n°ANR-23-IACL-0002.
- This work was performed using HPC resources from CNES Computing Center.