Progress in hybrid models for applications in remote sensing of vegetation

Jochem Verrelst, Juan Pablo Rivera, Miguel Morata, Jose Estévez, Matias Salinero, Pablo Reyes, Enrique Portales, Ana Belen Pasqual, Jorge Vicent, Santiago Belda, Bastian Siegmann, Katja Berger, and colleagues

07 June 2021, Warsaw, Poland
Any difference? Which model would you choose?

SCOPE (RTM)

Emulator (emulated SCOPE)

@760nm
Multiple models exist with diverse complexity.
Advanced RTMs: generation of a large LUT (>1000#)

SCOPE

DART

MODTRAN

Advanced RTMs: more realistic but slow
Emulators are statistical models that approximate the processing (input-output) of a physical model (e.g. RTM) - at a fraction of the computational cost:

**creating a statistical model from a physical model**
Hybrid models: Regression vs. Emulation:

Common use of hybrid models in optical RS:

Statistical regression method:
• Variable/data-driven, 1 output, portability is questionable

Emulation in optical RS:

Replace RTM:
• Multiple applications, e.g. inversion
  ✓ Radiometric method: Spectral fitting
  ✓ Portable: generally applicable

“spectral redundancy” is a blessing
Processing steps emulation

Input (variables + spectra)
Splitting into training/validation
PCA on spectra
MLRA training looping over components
Prediction of components
Reconstruction of spectra
Validation
Emulator

PCA on spectra

MLRA training looping over components

Prediction of components

Reconstruction of spectra

$\mathbf{S}_c = \mathbf{U} \cdot \mathbf{X}$

$\mathbf{W} = (\mathbf{Y} + \lambda \mathbf{I})^{-1} \cdot \mathbf{S}_c$

$\mathbf{S}_p = \mathbf{S}_c \cdot \mathbf{W}$

$\mathbf{X}_r = \mathbf{U}^\top \cdot \mathbf{S}_p$
With ARTMO’s emulation processing chain any RTM can be converted into an emulator.

\[ X_r = U^\top \cdot S_p \]


http://artmotoolbox.com
Errors can go down < 3\% (NRMSE) by optimizing ML, sample size, DR,...
Emulators great idea... what about accuracy?

1) Role of machine learning regression algorithm?

2) Role of dimensionality reduction (DR) method?

3) Role of LUT size training?

4) Role of data type?

If OK with losing some accuracy, various applications are possible:

- Fast RTM output generation:
  1. Fast spectral generation
  2. Fast scene generation
  3. Fast global sensitivity analysis
  4. Fast approximation of retrieval strategies

Applications (1/4)

Field data
Example of #500 emulated SPARC campaign spectra based on varying 6 field variables.
Scene generation
Emulation for hyperspectral scene generation

- GPR emulator applied for scene generation
- Compared against RTM scenes: PROSAIL & SCOPE

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Symbol</th>
<th>Units</th>
<th>Class 1</th>
<th>Class 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leaf area index</td>
<td>LAI</td>
<td>m²/m²</td>
<td>Uniform: 3 - 6</td>
<td>Uniform: 0 - 2</td>
</tr>
<tr>
<td>Chlorophyll a+b content</td>
<td>C_{ab}</td>
<td>μg/cm²</td>
<td>Uniform: 20 - 60</td>
<td>Uniform: 10 - 35</td>
</tr>
</tbody>
</table>

Parameter values compared against RTM scenes:
- PROSAIL: 45s
- SCOPE: 65min
- GPR emulator: 8s
Emulation for realistic synthetic scene generation.

A S2 image is used as input to emulate an hyperspectral image with a S2 texture.

**Image with high quality spatial information**
- Large image (e.g. S2)
- High spatial but low spectral resolution (e.g. Worldview, UAV)

**Image with high spectral information**
- Hyperspectral image

---

**Emulation for image fusion**

**Emulator**
- Training
- Testing

**Emulated image with high spectral and spatial information**
S2 texture, APEX-like hyperspectral info

Original APEX image
Global Sensitivity Analysis (GSA)

GSA techniques quantify the relative importance of each input variable to model outputs.

Applications (3/4)

PROSAIL

Emulator KRR (28s)
Emulator GPR (32s)
Emulator VHGP (37s)

1000#/variable
Applying emulation to $L_{\text{TOA}}^{(1/3)}$

PROSAIL spectra

$L_{\text{TOA}} = L_0 + \frac{T_{\text{gas}} \cdot E_{\text{tot}} \cdot T_{\text{tot}} \cdot \rho}{\pi(1 - S\rho)}$

6S or MODTRAN spectra

2500 TOA radiance spectra

Emulator

- 30 PCA
- 70/30% train/test
Some regions are perfectly fit for variable retrieval from TOA radiance data.

Emulation for GSA of atmospheric RTMs (3/3)

\[ L_{\text{toe}} = L_0 + \frac{E_{\text{dir} \mu_1} + E_{\text{dir} \mu_2} (T_{\text{dir}} + T_{\text{dio}}) \rho}{\pi (1 - S \rho)} \]

<table>
<thead>
<tr>
<th>Variable name</th>
<th>Min–max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elevation (h)</td>
<td>0–3 km</td>
</tr>
<tr>
<td>Aerosol optical thickness (AOT)</td>
<td>0.05–1</td>
</tr>
<tr>
<td>Ångström exponent (α)</td>
<td>0.1–1.5</td>
</tr>
<tr>
<td>Asymmetry parameter (G)</td>
<td>0.6–1</td>
</tr>
<tr>
<td>Single scattering albedo (SSA)</td>
<td>0.75–1</td>
</tr>
<tr>
<td>Water vapor (CWV)</td>
<td>1–4 g cm²</td>
</tr>
<tr>
<td>Ozone (O₃)</td>
<td>0.25–0.45 atm·cm</td>
</tr>
</tbody>
</table>

Applications (4/4)

(SIF) Retrieval
Emulation of sun-induced fluorescence (SIF) data

Emulator workflow: radiance to SIF with experimental data

INPUT → PCA Components → MLRA Training → Predicted Components → Reconstructed OUTPUT

SFM @760nm → Emulator

Absolute error

1-2min @760nm
Conclusions emulation

Emulation approximates physical models with sufficient accuracy and tremendous gain in speed thanks to ML.

- Emulation permits fast rendering of optical data
- Emulation permits fast calculation of global sensitivity analysis
- Emulation can provide a fast alternative of tedious retrieval routines
Hybrid retrieval methods

- BOA retrieval
- TOA retrieval

Variables (e.g. LAI, chlorophyll,...)

Regression

Emulation

Spectra

Variable (e.g. LAI)
BOA & TOA retrieval from S2 L2A and L1C data

Biochemical leaf traits and FVC

Canopy traits

Consistent BOA & TOA retrievals at the canopy scale: **no need for atmospheric correction** *(given a clear sky)*
TOA retrieval from S2 L2A and L1C data in GEE

Crucial for implementing GPR into GEE is reducing its size.

**Active learning**

For GPR a few hundred samples enough for optimal retrieval performances: **quality more important than quantity**. Light models can then be implemented into GEE.
TOA-based retrieval from S2 L1C data: Germany
zoom-ins @20m
S2 time series processing: gap filling and phenology indicators mapping

Hybrid retrieval models

EBD-GPR retrieval models
C_ab, C_aw, C_rm, LAI, LaiC_ab, LaiC_w, LaiC_m, FVC

GPR Training and validation
Active learning (AL)
Dataset optimization using EBD

Spatio-temporal Gapfilling
- GPR-based gapfilling strategy based on pre-optimized kernel hyperparameters.

Length-scale \( l \)
Signal variance \( \sigma^2_s \)
Noise variance \( \sigma^2_n \)

GEE GPR implementation
L2A product

Predicted mean
Gapfilling

Pixel-wise time series analysis

LAI/Gr (nm/ha)

Date
Gap-filled TS products: cloud-free vegetation products on a regular basis.
GEE calculation of S2 phenology indicators (SOS & EOS) based on vegetation products

In GEE, SOS and EOS can be determined anywhere in the world
TOA retrieval from S3 OLCI L1C data: Europe

FAPAR

LAI

FVC

LCC
The Copernicus Hyperspectral Imaging Mission, CHIME, will carry a visible to shortwave infrared spectrometer to provide routine hyperspectral observations to support new and enhanced services for sustainable agricultural and biodiversity management, as well as soil property characterisation.

**Technical concept**

Routine spectroscopic observation in contiguous spectral bands:
- Instrument: Pushbroom Imaging Spectrometer 400 – 2500 nm, \( \Delta \lambda \leq 10 \text{nm} \)
- Revisit 10 – 15 days
- GSD (spatial resolution) 20 – 30 m
- Sun synchronous orbit (LTDN 10:30 – 11:30)
- Nadir view covering land and coastal areas
- High radiometric accuracy, low spectral/spatial misregistration

**CORE Data Products:**

The mission shall provide access to Level-1B, Level-1C and Level-2A products accessible via DIAS and with API support:
- Bottom-of-Atmosphere (BOA) reflectance (atmospheric corrected)
- Ortho rectified geometry
- Basic pixel classification (opaque clouds, thin clouds, cloud shadows, vegetation, water, snow etc.)
- **Additionally -> Vegetation products (Level-2B)**
Hybrid workflow CHIME vegetation models

Use RTMs (e.g. SCOPE, PROSPECT DyN – SAIL) to generate a LUT composed by pairs (e.g. 1000) of vegetation parameters and spectra. 

Select the most representative samples from the LUT via a diversity or entropy criteria. Later, add non-vegetated spectra.

Dimensionality reduction with PCA (20 components) or band selection based on GSA and kernel sensitivity ranking.

With the LUT optimized for vegetation and non-vegetated surfaces, train probabilistic ML algorithms.

Apply to new observations

Final outputs of the workflow.

Maps + uncertainties

Apply to new observations

PRISMA images resampled to CHIME band settings

Validate the models

Assess models’ performance against field data and vegetation reference scenes.
Maps: results with 20 PCA

- **PRISMA** image resampled to CHIME band settings, heterogeneous spatial subset to test the performance in vegetation, buildings and water
- Canopy variables 😊
- Some leaf variables 😞

GPR hybrid models powerful for vegetation trait mapping (with inclusion of uncertainty estimates)
Conclusions hybrid models for retrieval

✓ Hybrid models powerful for vegetation properties mapping: generic, adaptive, competitive, fast and provision of uncertainty estimates (GPR)

✓ Hybrid models at both BOA and TOA scale (S2, S3)

✓ Framework developed for running GPR models into GEE. Any GPR model can process anywhere and anytime within the GEE catalogue.

✓ Also gap-filling processing in GEE & calculation of phenology indicators

✓ Hybrid GPR models under development for vegetation traits retrieval from the future CHIME mission
Thanks!

Questions?

http://artmotoolbox.com
https://ipl.uv.es/sentiflex/