Machine Learning-Based Observation Operators to Assimilate Microwave and SIF Satellite Observations into the ECMWF Integrated Forecast System (IFS)

Sébastien Garrigues (1), Patricia de Rosnay (1), Ewan Pinnington (1), Peter Weston (1), Anna Agusti-Panareda (1), Souhail Boussetta (1), *Jean-Christophe Calvet (2), David Fairbairn (1), Cédric Bacour (3), Richard Engelen (1), Stephen English (1).*

(1) *ECMWF, Reading UK*
(2) *CNRM, Université de Toulouse, Météo-France, CNRS, 31057 Toulouse, France*
(3) *LSCE,* Gif-Sur-Yvette, *France*

CORSO
CO2MVS Research on Supplementary Observations

Funded by the European Union

Outlines

1. Introduction
2. Methodology
3. Active microwave (ASCAT) observation operator
4. Solar Induced Fluorescence (SIF) observation operator
5. Conclusion

Introduction

✓ CORSO project: Reducing the uncertainties in the land carbon budget

- large **uncertainties in Gross Primary Productivity (GPP)** predictions
- **constraint both water and carbon fluxes=> analyze both soil moisture and vegetation variables**

✓ Assimilate new type of land satellite observations in the Integrated Forecast System (IFS)

- **Level-1 active microwave observations**
  - sensitive to both vegetation structure (Petchiappan et al., 2021) and soil moisture (Wagner et al., 2013)
  - more accurate representation of uncertainties compared to retrievals

- **Solar Induced Fluorescence (SIF)**
  - emission of electromagnetic radiation in the red and far-red by '*chlorophyl a*' molecule under visible light
  - directly related to leaf physiological processes (photosynthesis)
  - correlation with both GPP and Leaf Area Index (LAI) (Guanter et al., 2014; He et al., 2017)

Introduction

✓ Observation operator

- Predict model-simulated counterpart of the satellite observation using the IFS fields as predictors
- Physically based observation operator: large uncertainties over land, complex and computationally expensive,
- ML alternative
  - Generic architectures can be applied to different types of EO
  - Computationally more efficient
  - Quickly test the assimilation of new types of observation

✓ Challenges

- Design simple and robust observation operator for their integration in the IFS at global scale
- Is the information content of the Earth System model fields sufficient to simulate the satellite signal ?
- How to ensure enough sensitivity to the input fields that we want to analyse (LAI, GPP)
- How to represent the uncertainties in the predictors and the output?
- Importance of localization : use latitude and longitude in the predictors ?

Methodology to design the ML-based observation operator

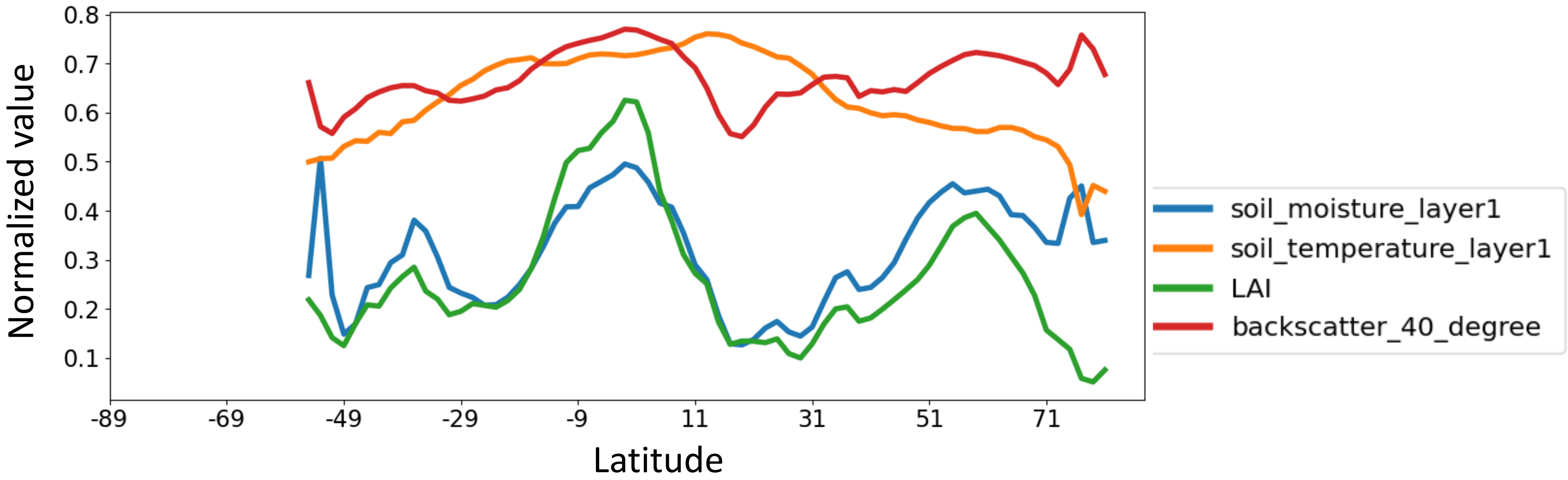
- **Training database**:
  - collocated observation and model fields in the observation space
  - quality control and filtering (snow, frozen soil, orographic surface…)
- **Feature selection**
  - process-based knowledge
  - XAI methods (e.g. SHAP)
- **ML model**:
  - Gradient boosted trees (XGBOOST, Chen et al., 2016) (XGB)
  - Feedforward neural network (NN)
- **Training and hyperparameter** tuning (training and validation set)
- **Evaluation on test set** (temporal profile, spatial distribution, gradient)
- **Implementation and test in IFS – data assimilation experiments**

ASCAT observation operator: Training database

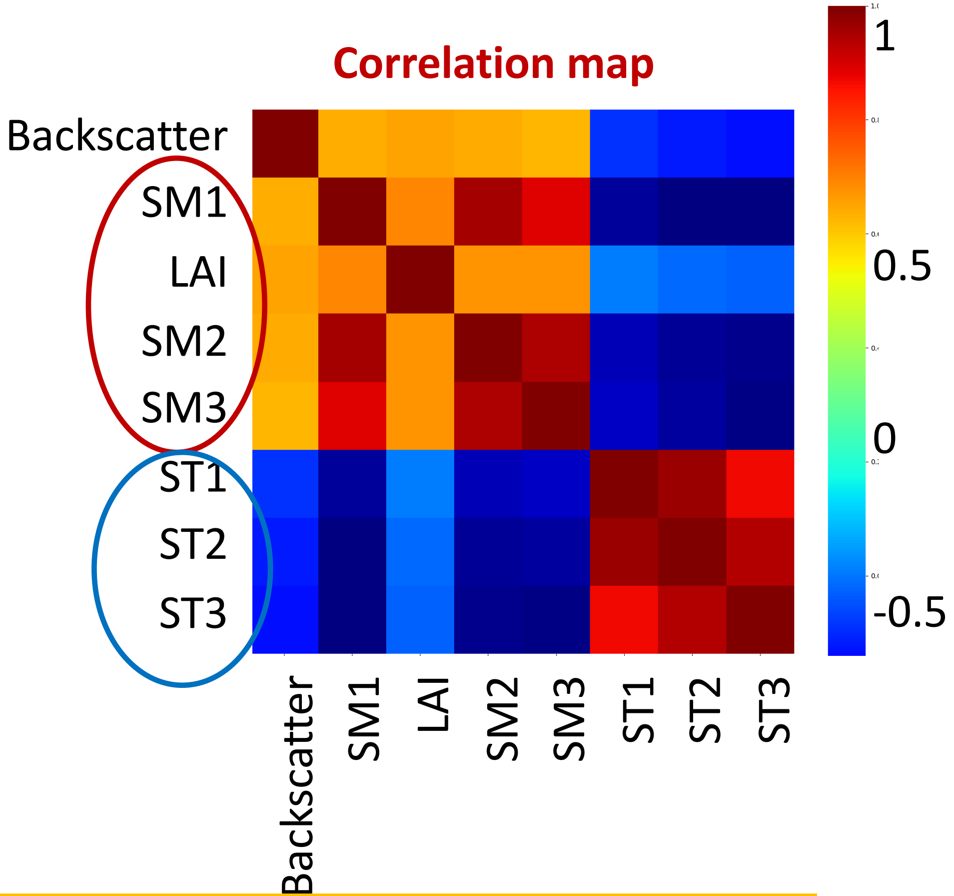
Training database (Aires, et al., QJRS 2021)

- target: ASCAT backscatter normalized at 40°
- model fields (features) from ERA-5: Leaf Area Index (LAI), soil moisture (SM) (3 layers), soil temperature (ST) (3 layers)
- localization : Latitude, longitude (sin/cos transform)
- period: 2016-2018 (training and validation), 2019 (testing)
- resolution: 0.25° grid.

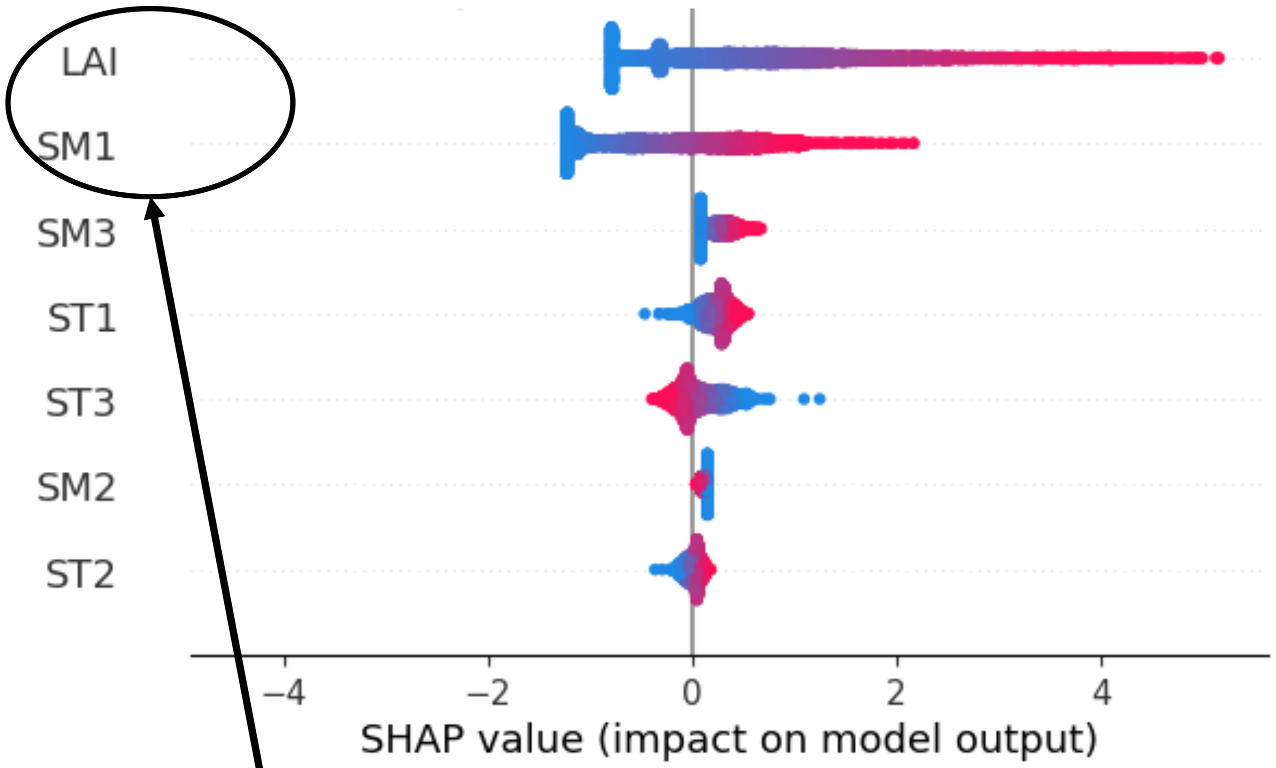
Latitude transect



ASCAT observation operator: Information content and explainability analysis



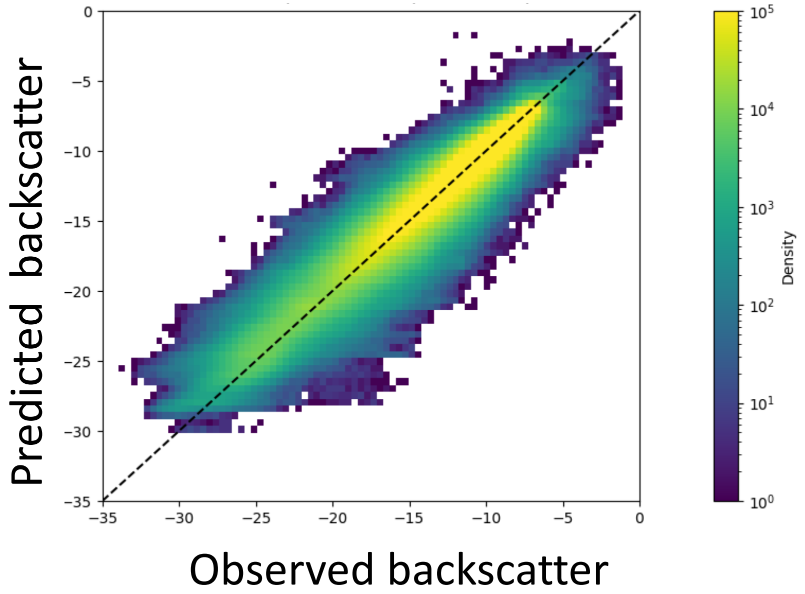
Contrasted correlation with backscatter:
- SM, LAI: positively correlated
- ST: negatively correlated



Vegetation (LAI) and surface soil moisture (SM1) are the most influent variables

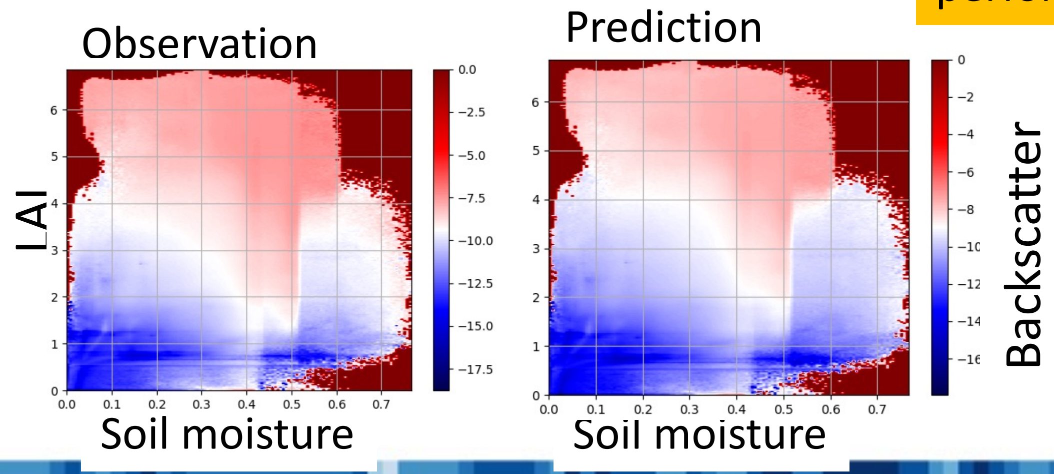
ASCAT observation operator: Performance evaluation

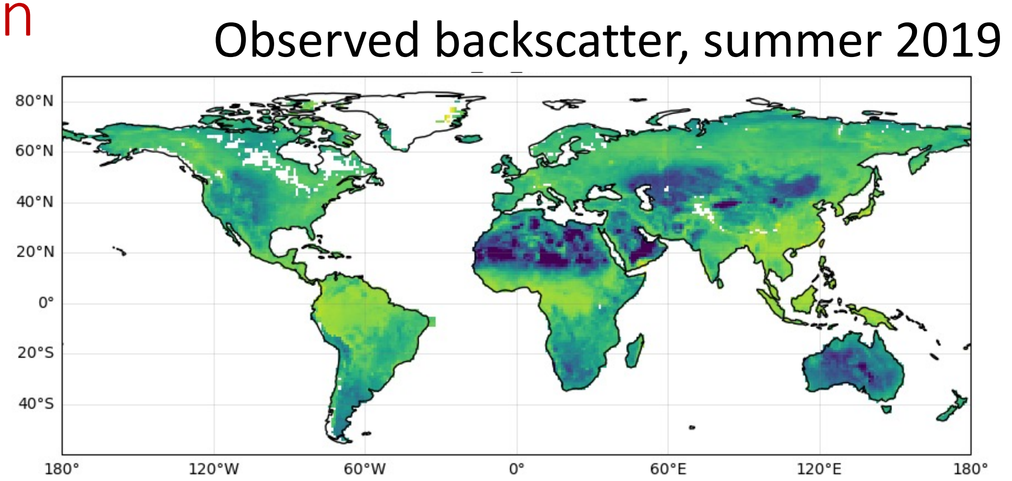
Test: R^2=0.93; RMSE=0.87;MAE=0.78;SD=0.87

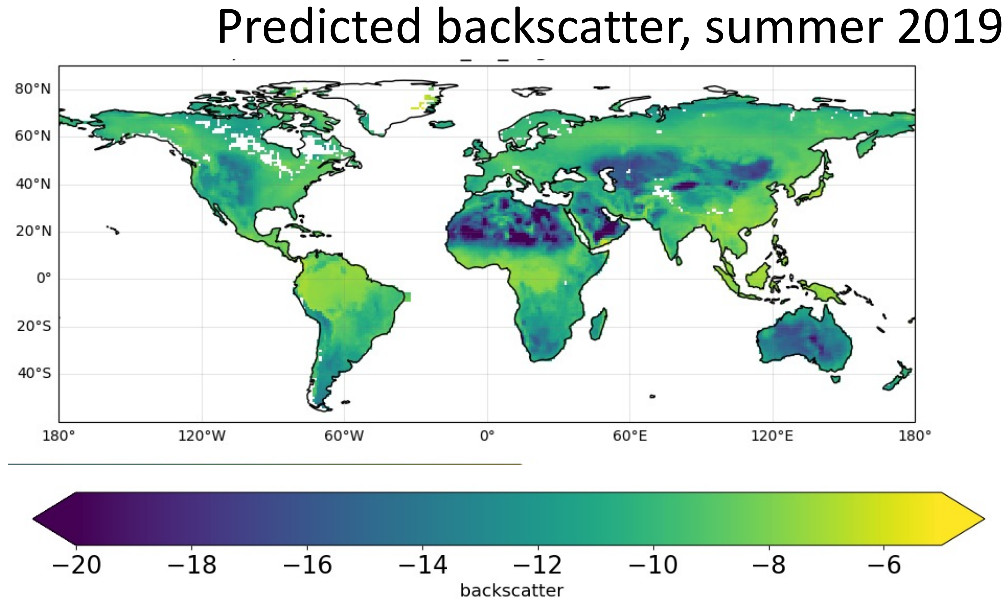


<u>NN model</u>:
3 years training,
4 hidden layers,
60 neurons,
global scale

Good training and generalization performances



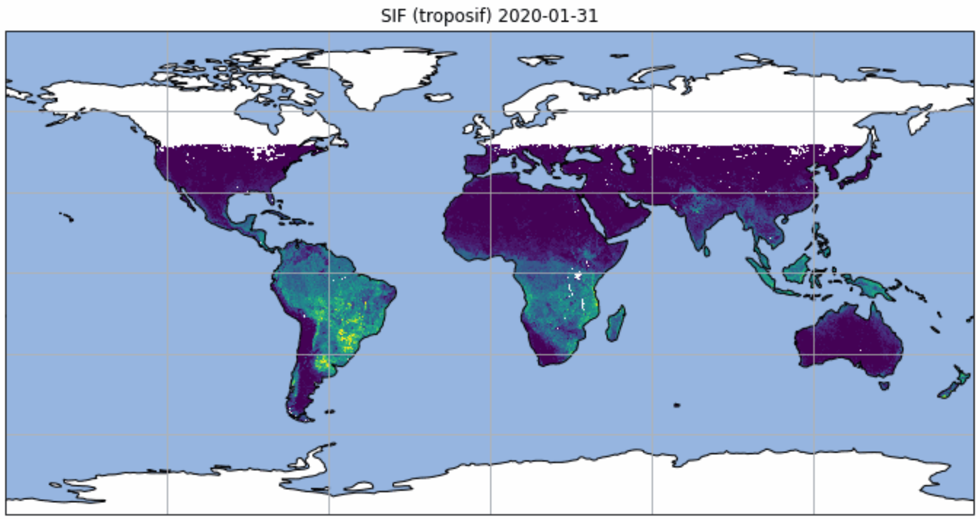




SIF observation operator: Training database

✓ Data

- Predictors: fields from ECLand land model offline simulations (IFS Cyc49r1)
- Target: SIF at 740nm satellite observations from TROPOMI/Sentinel-5p, Troposif dataset (Guanter et a., ESSD 2021)
- Resolution: 0.1° grid and at 8-day temporal frequency
- Filters: Large view and solar zenith angles, orography area, snow area, frozen soil
- Training: 2019-2020; Validation:2021; Test:2022



SIF observation operator: Feature selection

SIF canopy drivers

$$SIF_{canopy} = \underbrace{\underbrace{f_{esc} \times APAR}_{\text{Canopy structure (LAI)}} \times \phi_F}_{\text{Regulated by environmental factors}}$$

Canopy structure (LAI)

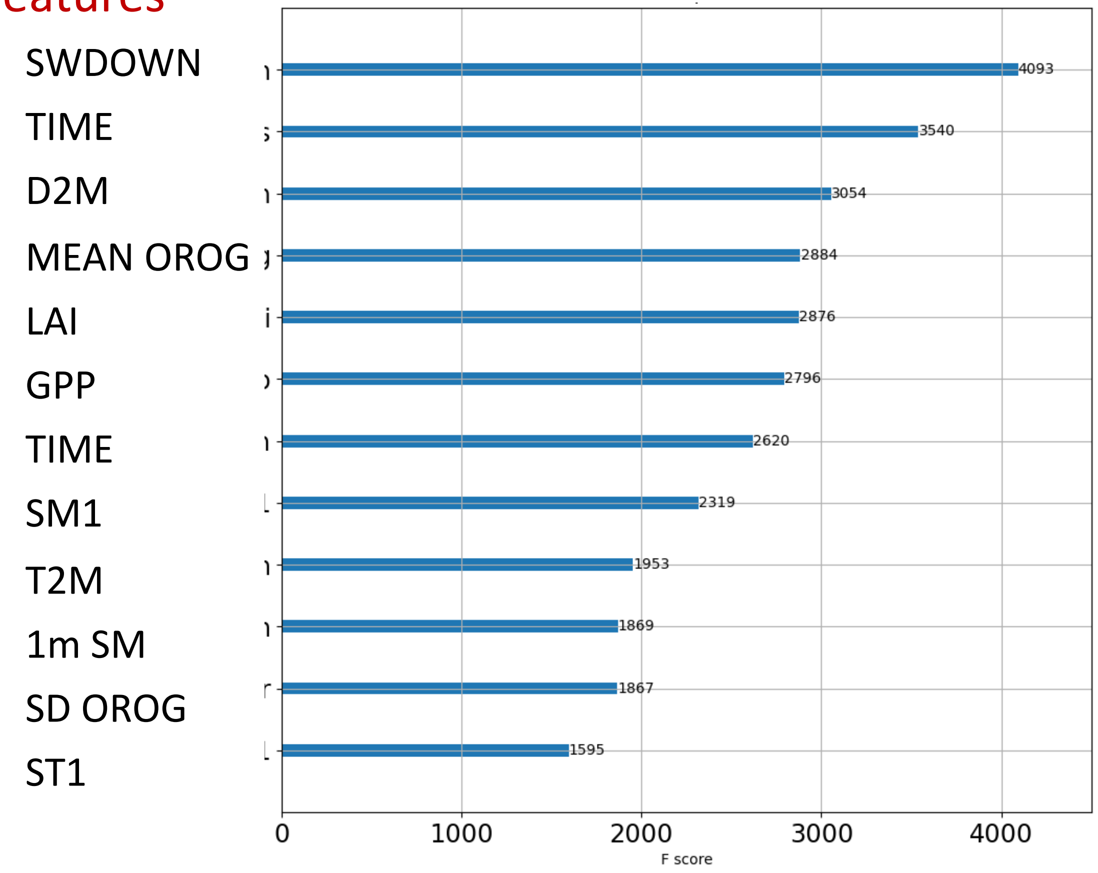
Leaf physiological characteristics (GPP)

Regulated by environmental factors:
soil moisture, solar radiation, 2m temperature and humidity

+ Temporal dependency: week of the year (cyclic transform)

Feature importance (xgboost)

features



SIF observation operator: ML model comparison

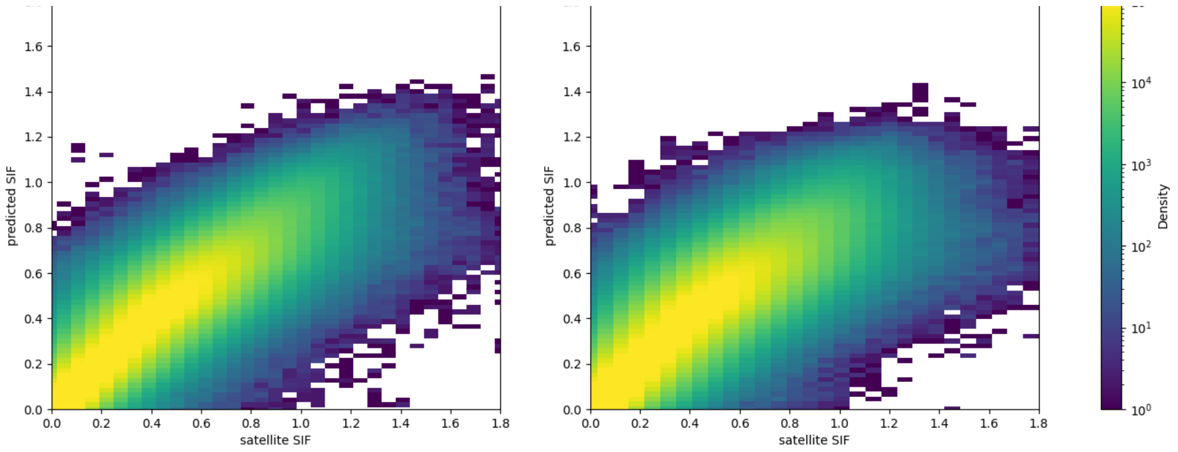
Training year=2019-2020, test=2022

Training

R^2=0.88, RMSE=0.09, MAE=0.25

Test

R^2=0.85, RMSE=0.1, MAE=0.26



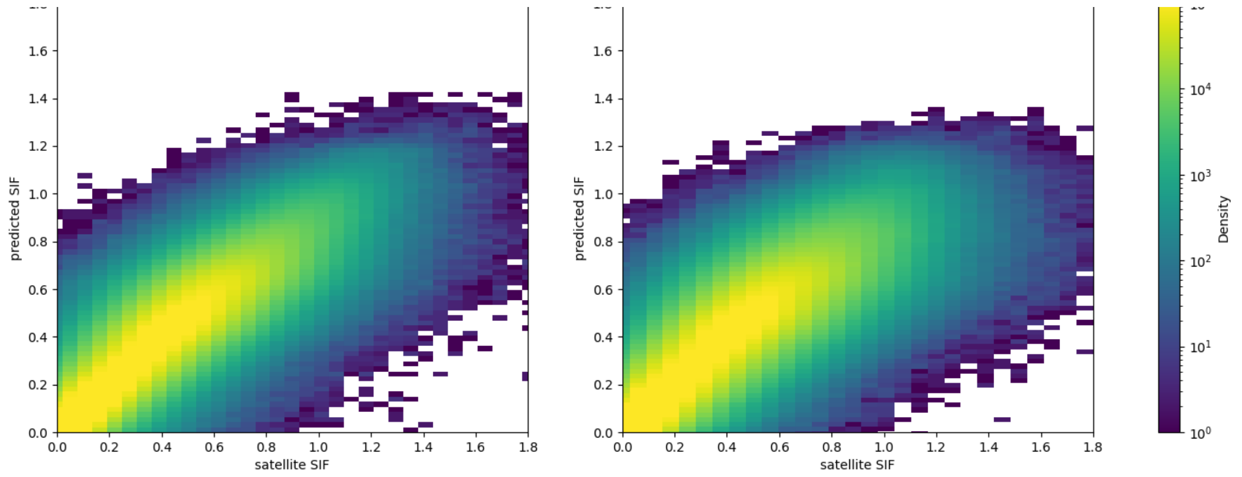
XGBOOST (ntrees=500, optimized hyperparameters)

Training

R^2=0.87, RMSE=0.09, MAE=0.25

Test

R^2=0.84, RMSE=0.1, MAE=0.27



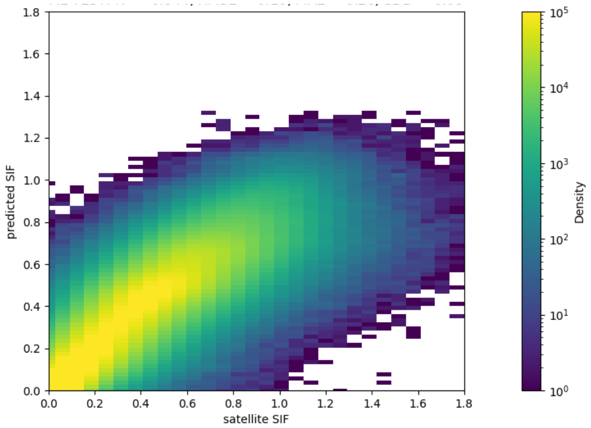
Feedforward NN (6 layers, 60 neurons, batch size=128, lr=0.001)

Equivalent performances between XGBOOST and NN

SIF observation operator: Global vs vegetation type ML model

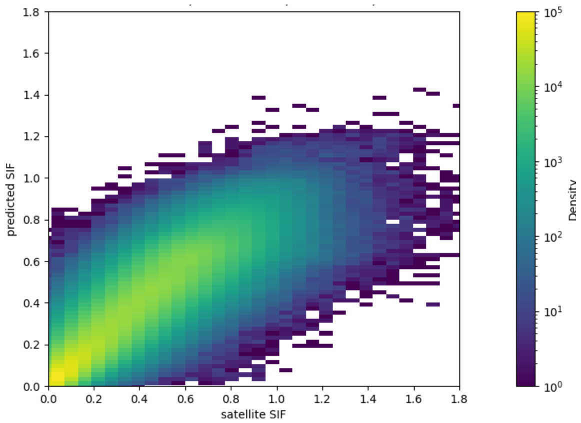
global

R^2=0.85, RMSE=0.1, MAE=0.26



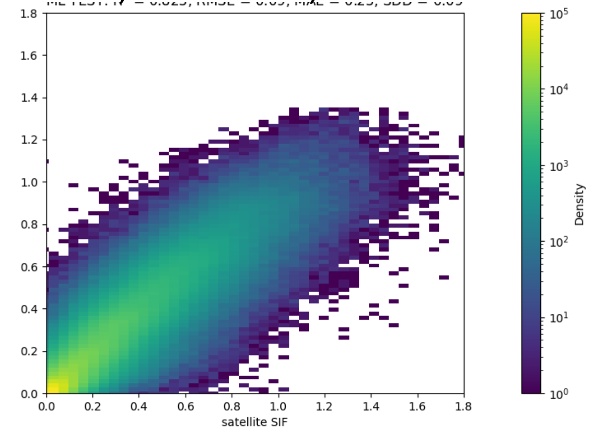
crop

R^2=0.82, RMSE=0.11, MAE=0.28



grassland

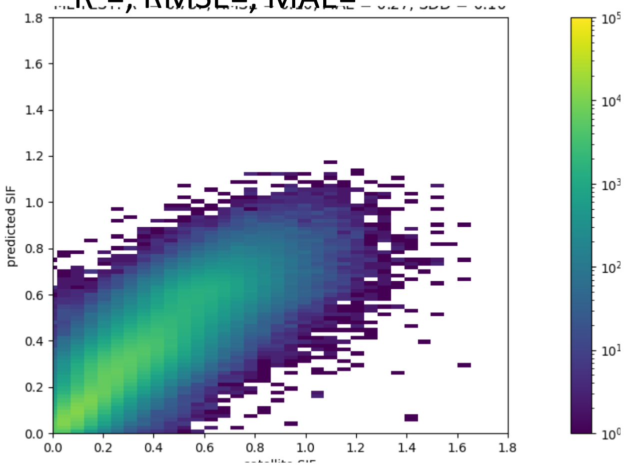
R^2=0.83, RMSE=0.09, MAE=0.25



Little benefit of training the model on distinct vegetation types

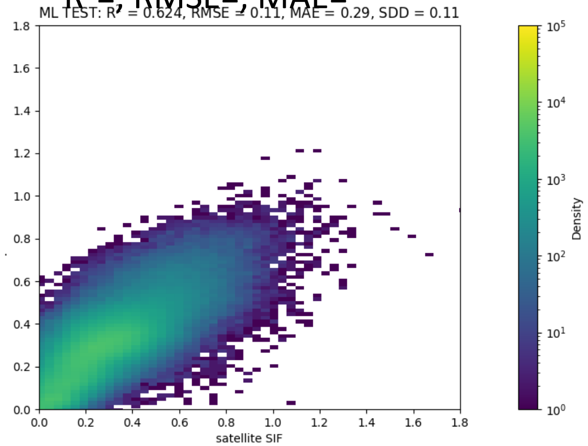
ENF

R^2=, RMSE=, MAE=



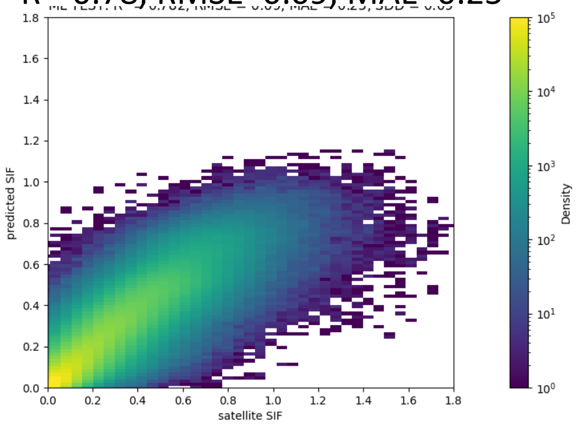
DNF

R^2=, RMSE=, MAE=



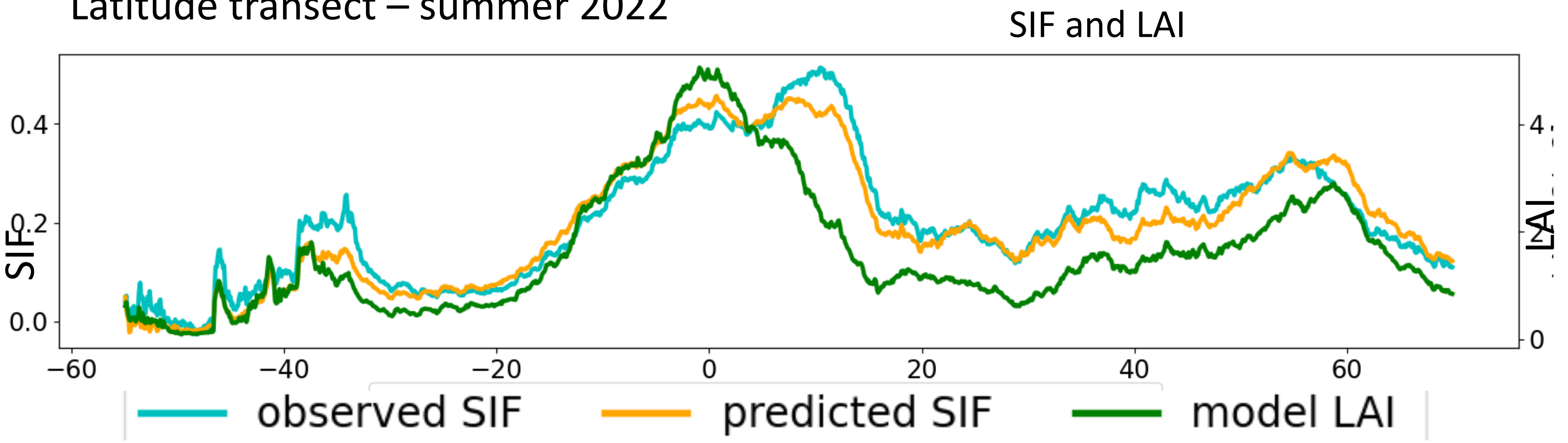
Shrubland

R^2=0.78, RMSE=0.09, MAE=0.25

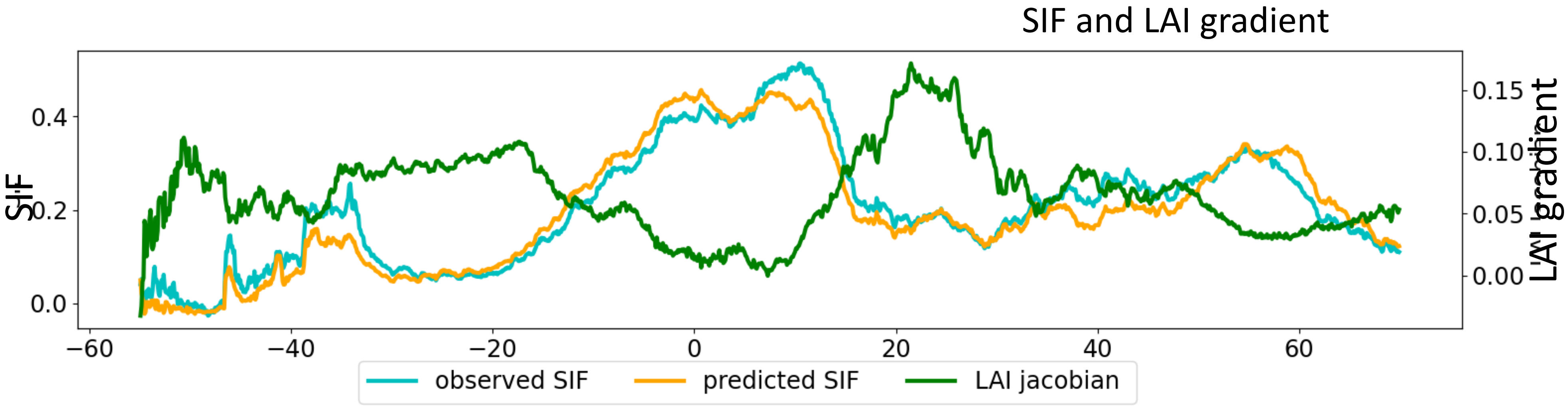


SIF observation operator: Evaluation

Latitude transect – summer 2022



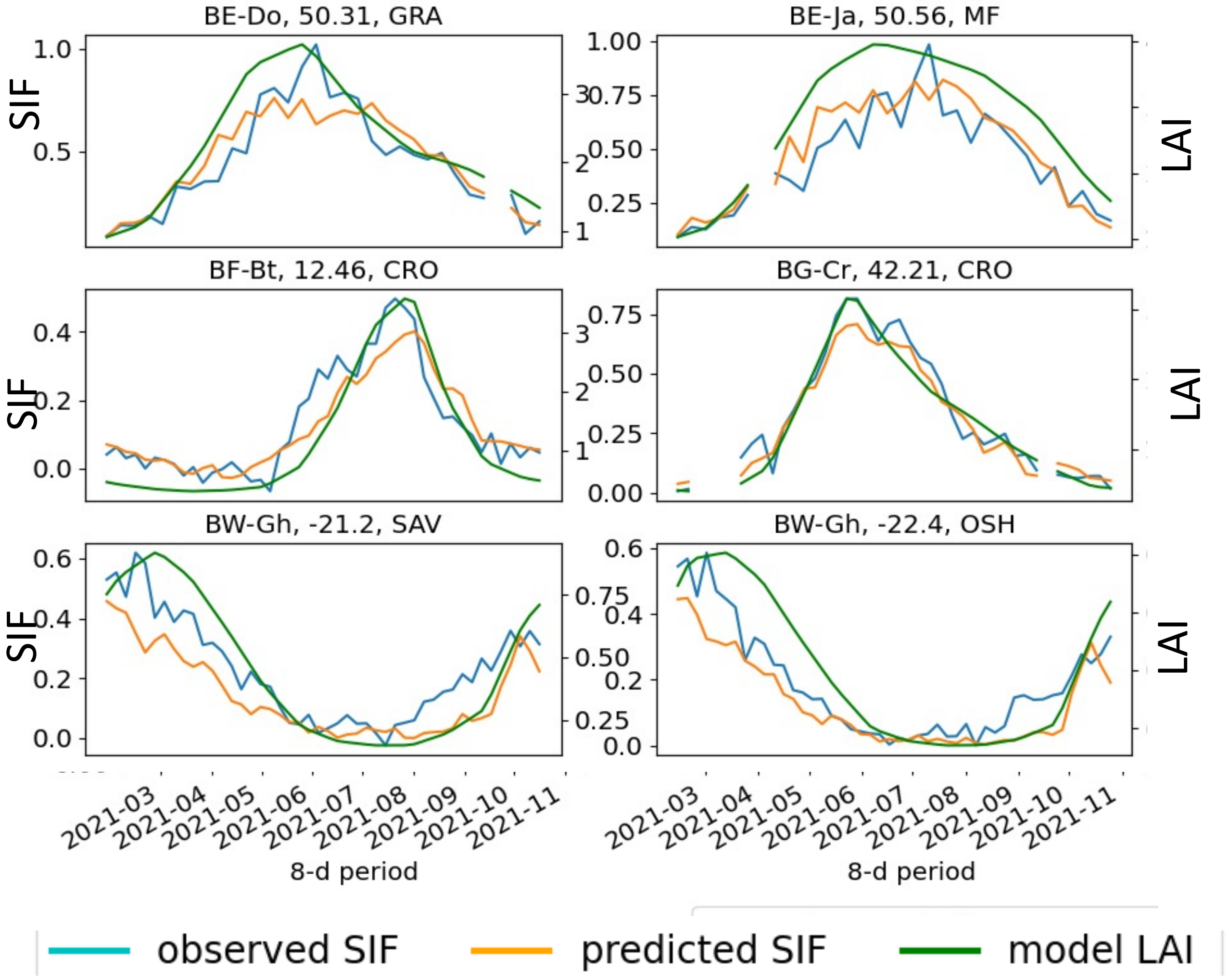
Accurate prediction of SIF spatial distribution



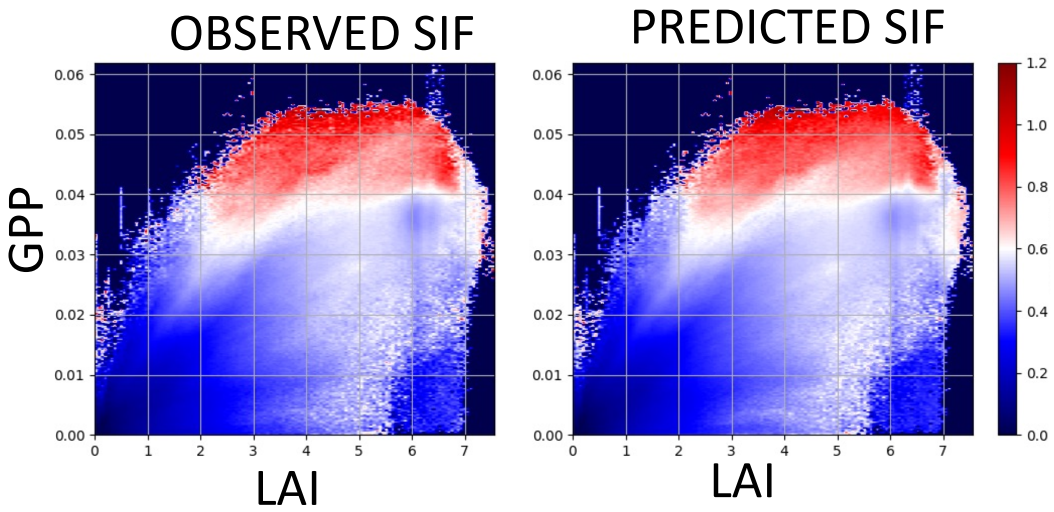
Good sensitivity to LAI

SIF observation operator: Evaluation

Seasonal evolution



GPP and LAI patterns



Accurate prediction of
- SIF seasonal evolution
- SIF patterns in GPP vs LAI spaces.

Conclusions

- **Simple feedforward NN** provides **accurate enough prediction of backscatter and SIF** satellite signals from the **ECMWF/IFS NWP model fields**

- Nex step : **test the assimilation in the IFS** and **evaluate the impact** on carbon fluxes, water fluxes and NWP near surface variables

- **ML-based observation operator** allows to **quickly test the assimilation** of **new types of observations, generic framework can be applied to other observations** (e.g. passive microwave observation)

- **Challenges and lesson learned**
  - Important to evaluate the sensitivity of the input fields that will be analyzed
  - Representation of uncertainties in both input features and satellite target
  - Risk of overfitting due to the use of latitude and longitude

Thanks for your attention

Acknowledgements

CORSO
CO2MVS Research on Supplementary Observations

Funded by the European Union

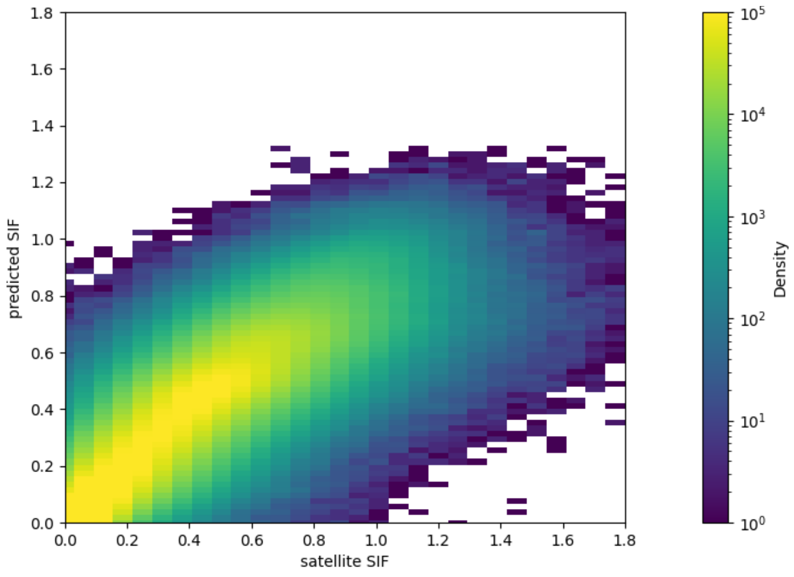
The CORSO project (grant agreement No101082194) is funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the Commission. Neither the European Union nor the granting authority can be held responsible for them

SIF observation operator: Impact of target variable

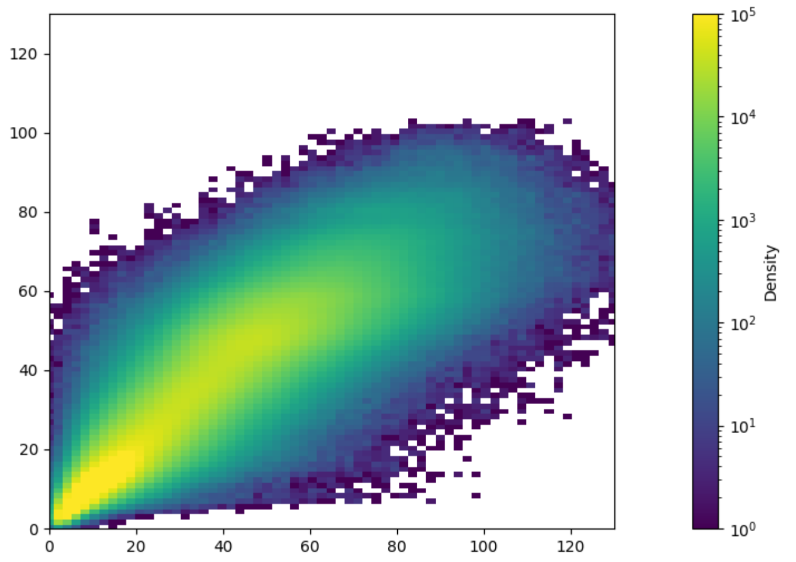
Target= SIF

Target= NIRVp (product of the near infrared reflectance of vegetation (NIR_V) over the NIR region and incoming PAR

R^2=0.85, RMSE=%, MAE=%



R^2=0.86, RMSE=%, MAE=%



SIF signal is more moisy than NVIRp
=> Reduced prediction performances