Machine Learning for Land Modelling

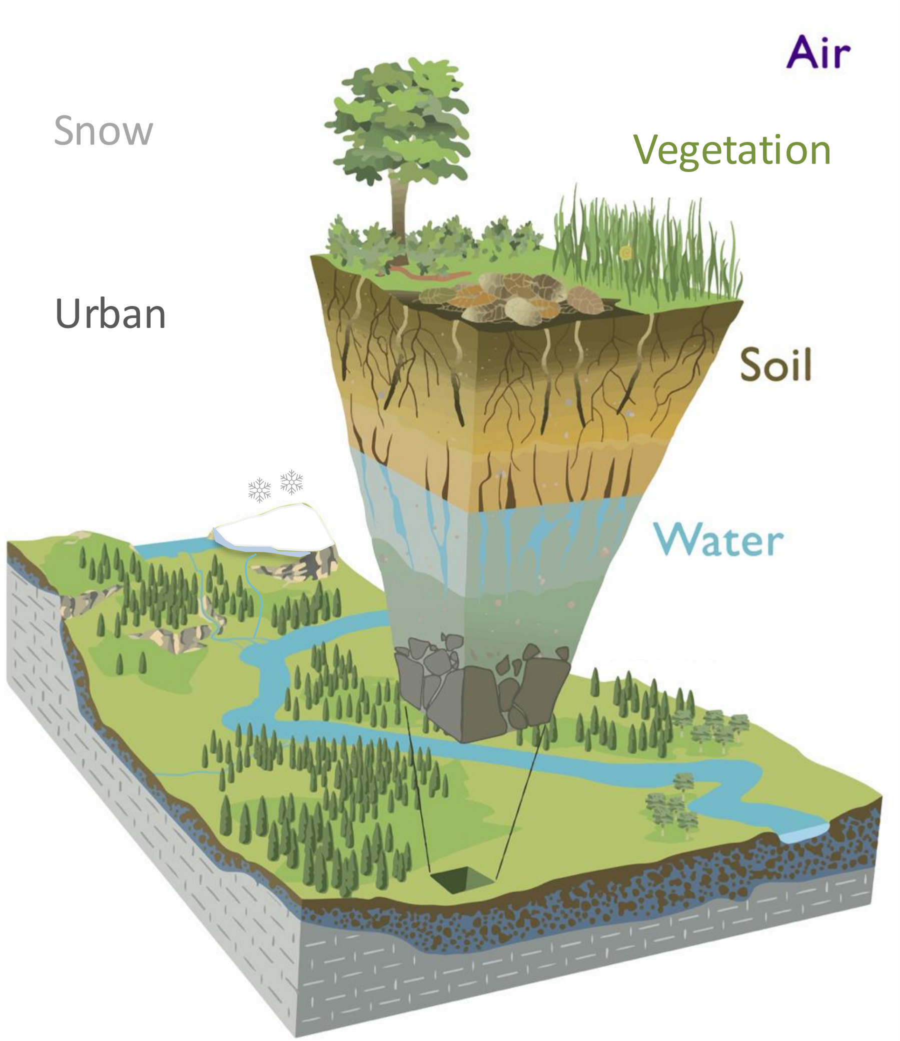
Christoph Rüdiger, Nina Raoult
Land Modelling, Research Department

ECMWF

Key processes of the land surface (model)

- Snow parameterisation/model
- Glaciers and sea-ice
- Sub-grid scale heterogeneity

- Urban processes (hydro, veg)
- Land cover
- Anthropogenic contributions



- Orography
- Coupling

- Updated climatologies
- Land cover and vegetation cover
- Parameterisations and phenology

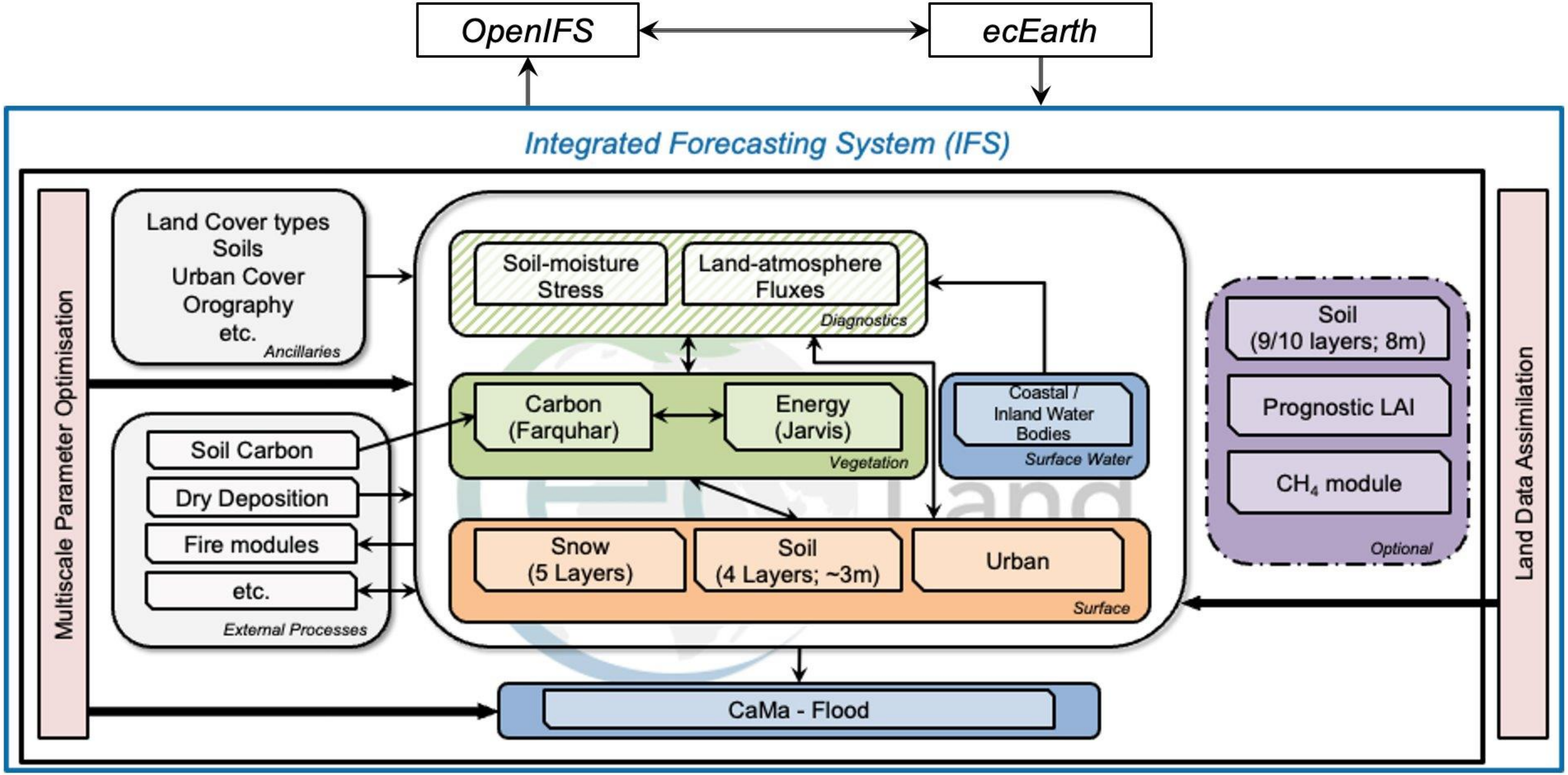
- Additional soil layers
- Soil maps and physics
- Parameterisations

- Runoff generation
- CaMa-Flood
- Irrigation/inundation
- Plant-water availability (soil dynamic range)
- Groundwater table representation

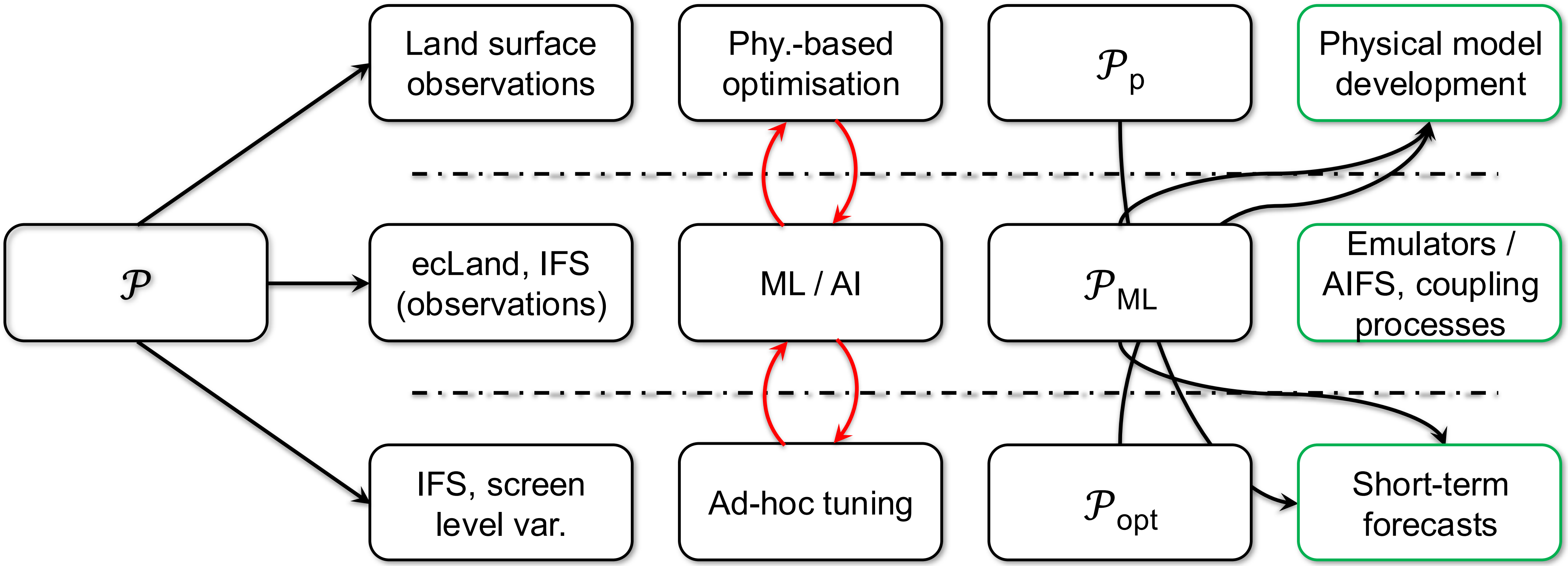
- Dynamic water bodies
- Coupling with ocean (2-ways)
- Lakes

Image from Chorover et al., 2007

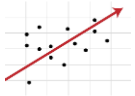
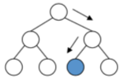
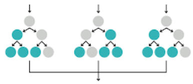
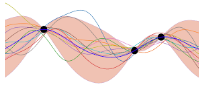
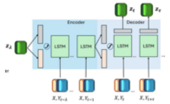
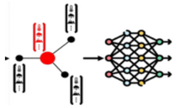
ecLand – ECMWF's land surface model



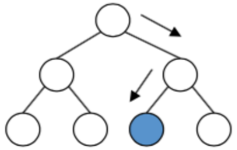
Bringing together physically- and data-driven parameters/processes



Machine Learning at a Glance

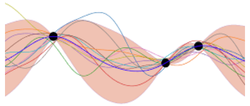
	Typical Methods	Common Uses in Land Surface Modelling
	Linear Regression, Logistic Regression	Simple empirical relationships, basic trend analysis
	K-Nearest Neighbours, Decision Trees	Simple classification (vegetation, soil type), exploratory modeling
	Random Forests, Support Vector Machines, Principal Component Analysis	Feature selection, process classification, parameter estimation
	Gradient Boosting, Gaussian Processes	Surrogate modeling, improved prediction accuracy, partial uncertainty analysis
	Deep Neural Networks (MLPs), LSTMs, Autoencoders	Temporal & spatial modelling, complex pattern recognition, surrogate modelling
	GANs, Reinforcement Learning, Graph Neural Networks	Generative modelling, adaptive control, coupled systems, cross-domain transfer
	Transfer Learning, Physics-informed ML, Multi-task / Multi-modal Learning, Large Language Models	Cross-domain transfer, hybrid models combining physical & ML models

Machine learning techniques



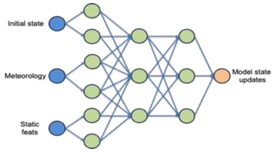
XGBoost

Pros: Fast, accurate for structured/tabular data.
Cons: Limited for spatiotemporal or complex dependencies.



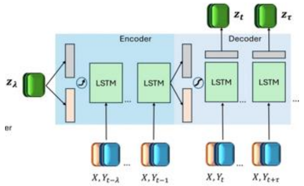
Gaussian Processes

Pros: Great uncertainty estimates; good for smooth interpolation.
Cons: Slow for big data; tricky for complex, high-dimensional patterns.



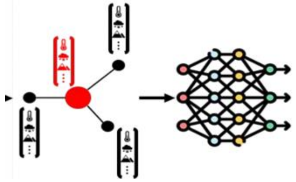
MLP

Pros: Simple, light-weight, works well with numerical inputs.
Cons: Struggles with spatial/temporal patterns.



LSTM

Pros: Captures temporal dependencies, good for time-series.
Cons: Slow training, sensitive to hyperparameters.

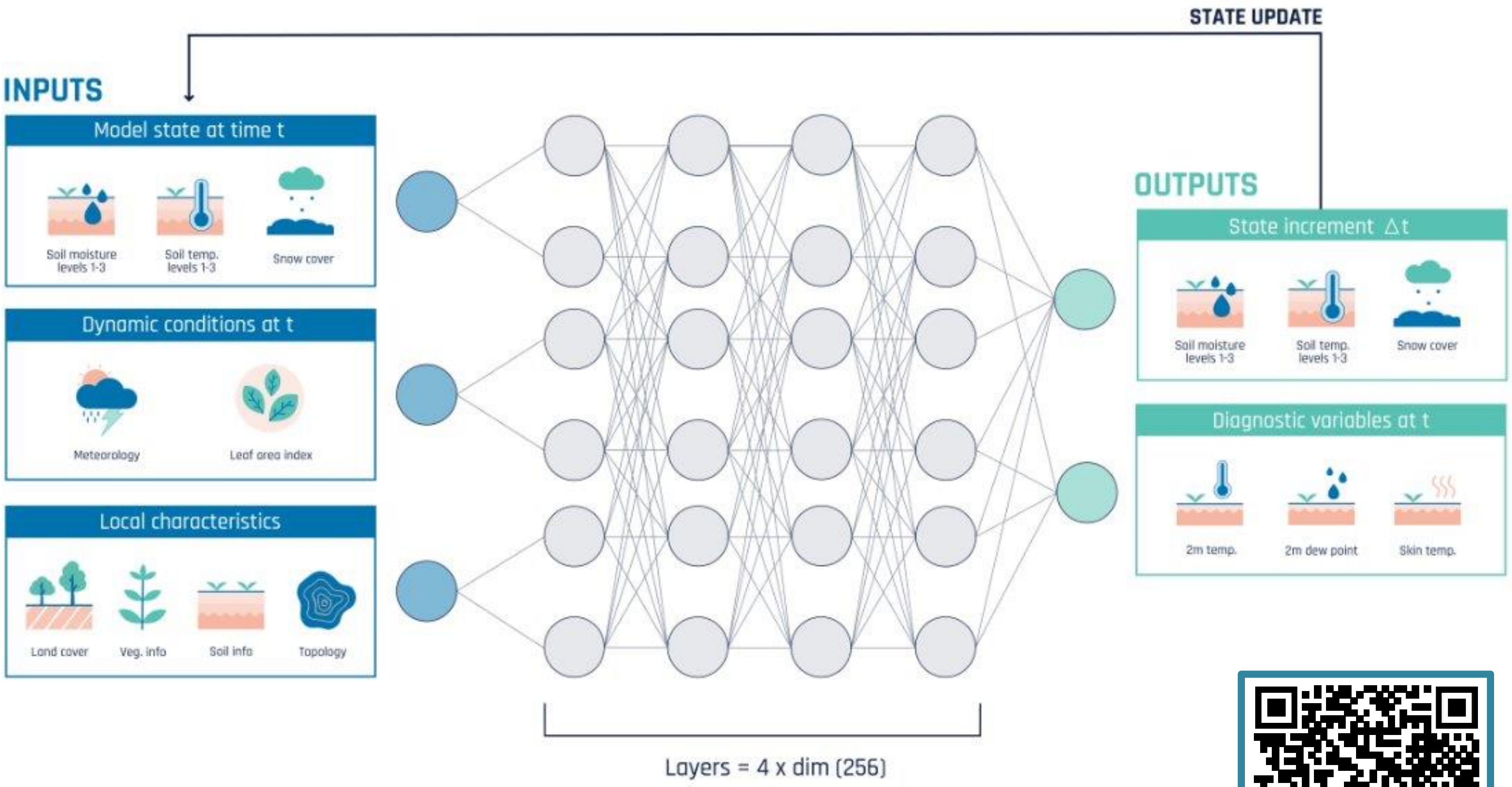


Graph NN

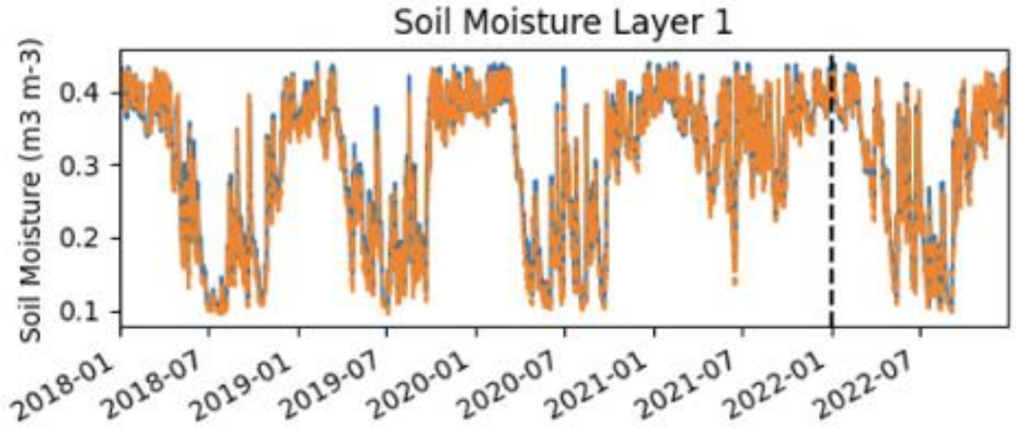
Pros: Models spatial relationships effectively.
Cons: High computational cost, complex implementation.

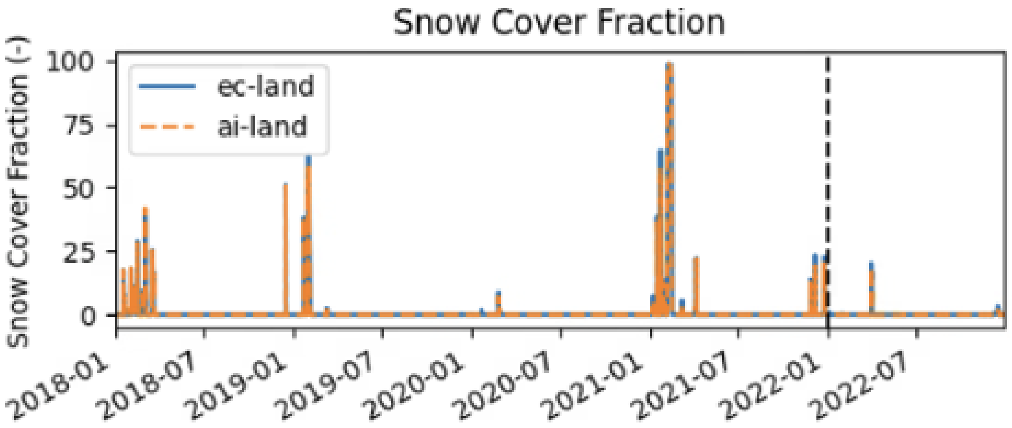
We use a Multi-layered Perceptron to emulate ecLand outputs

- **Fast & Flexible:**
  Quick to train & easy to adapt
- **Exploits Single-Column Structure of ecLand**
- **Resolution Agnostic:**
  Works across spatial & temporal scales.
- **Easily Transferable:**
  Can be reused or fine-tuned on new datasets.
- **Differentiable:**
  Supports gradient-based parameter estimation & data assimilation
- **Fine-Tuning on Observations:**
  Can integrate observational data to improve accuracy.



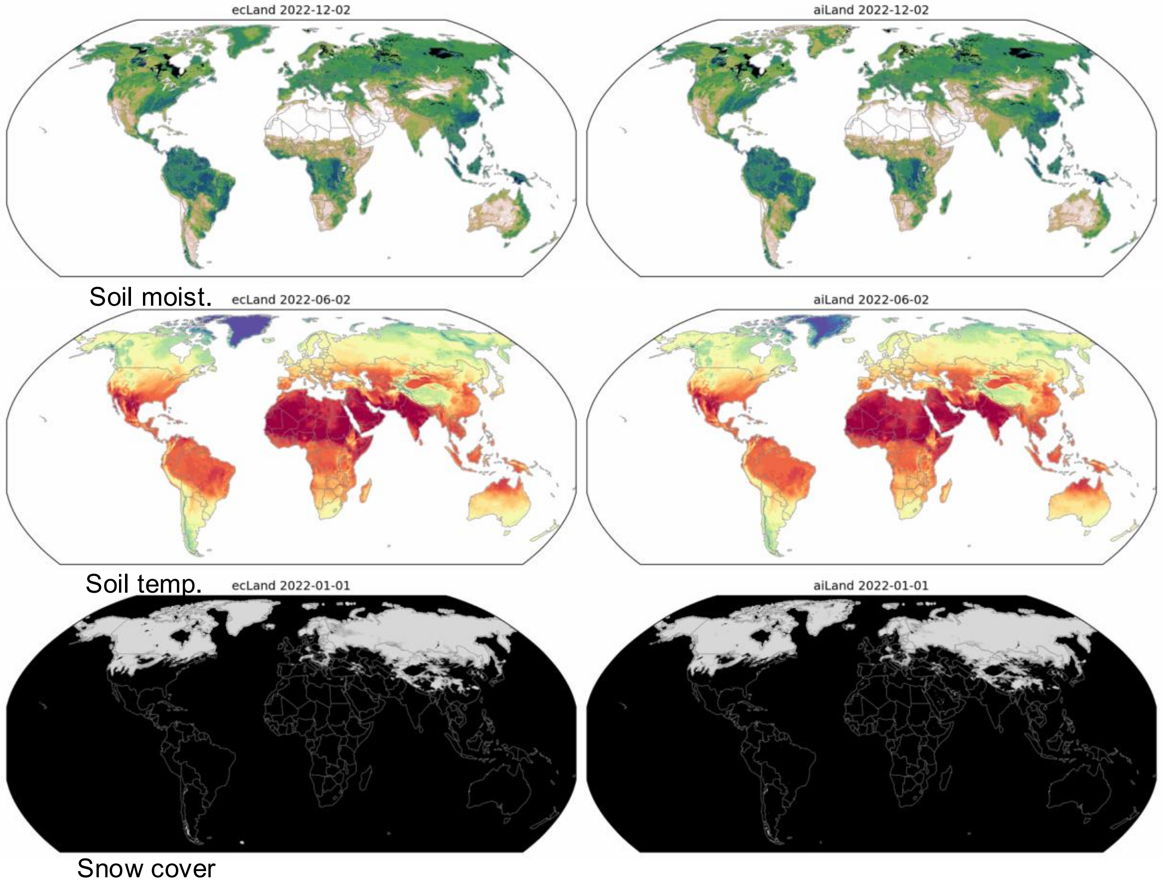
The AI model captures the behaviour of physical models in space & time





Timeseries for Bonn initialized only once at the beginning of the period, 2022 acts as a validation year (data not used in training)

ecLand: physical model **aiLand**: AI model

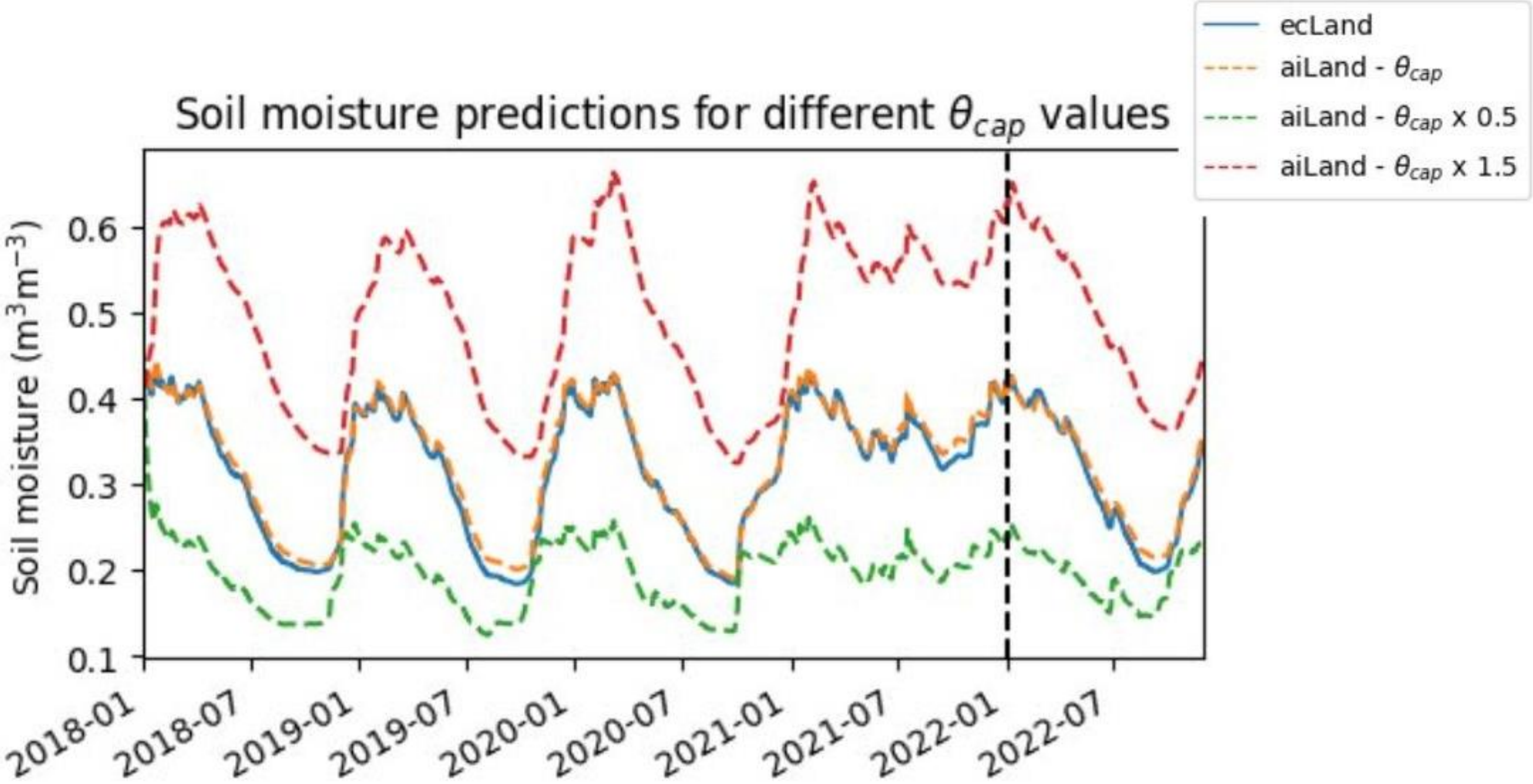


Learning the physics

Has the AI model captured physically consistent behaviour – and does it respond realistically when parameters are changed?

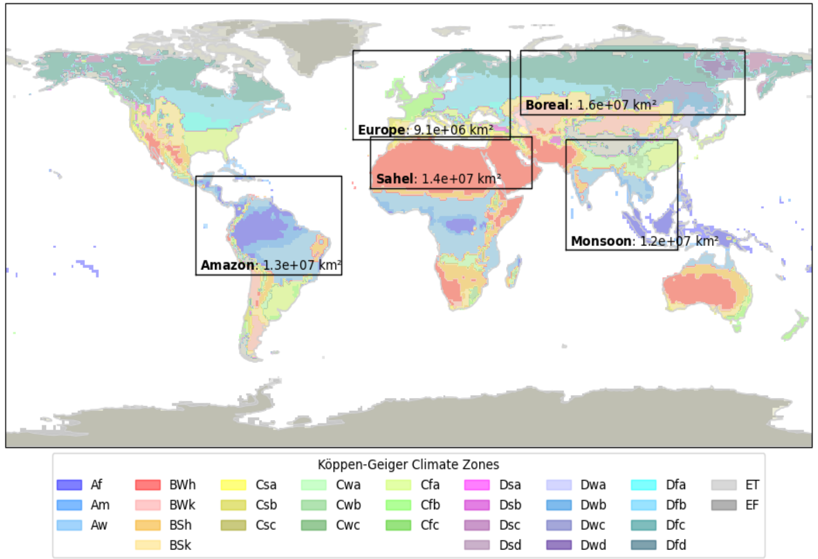
θ_{cap} – field capacity: controls maximum water soil can hold after drainage

Increasing θ_{cap} means the soil can **hold more water**, so we get **higher soil moisture** values

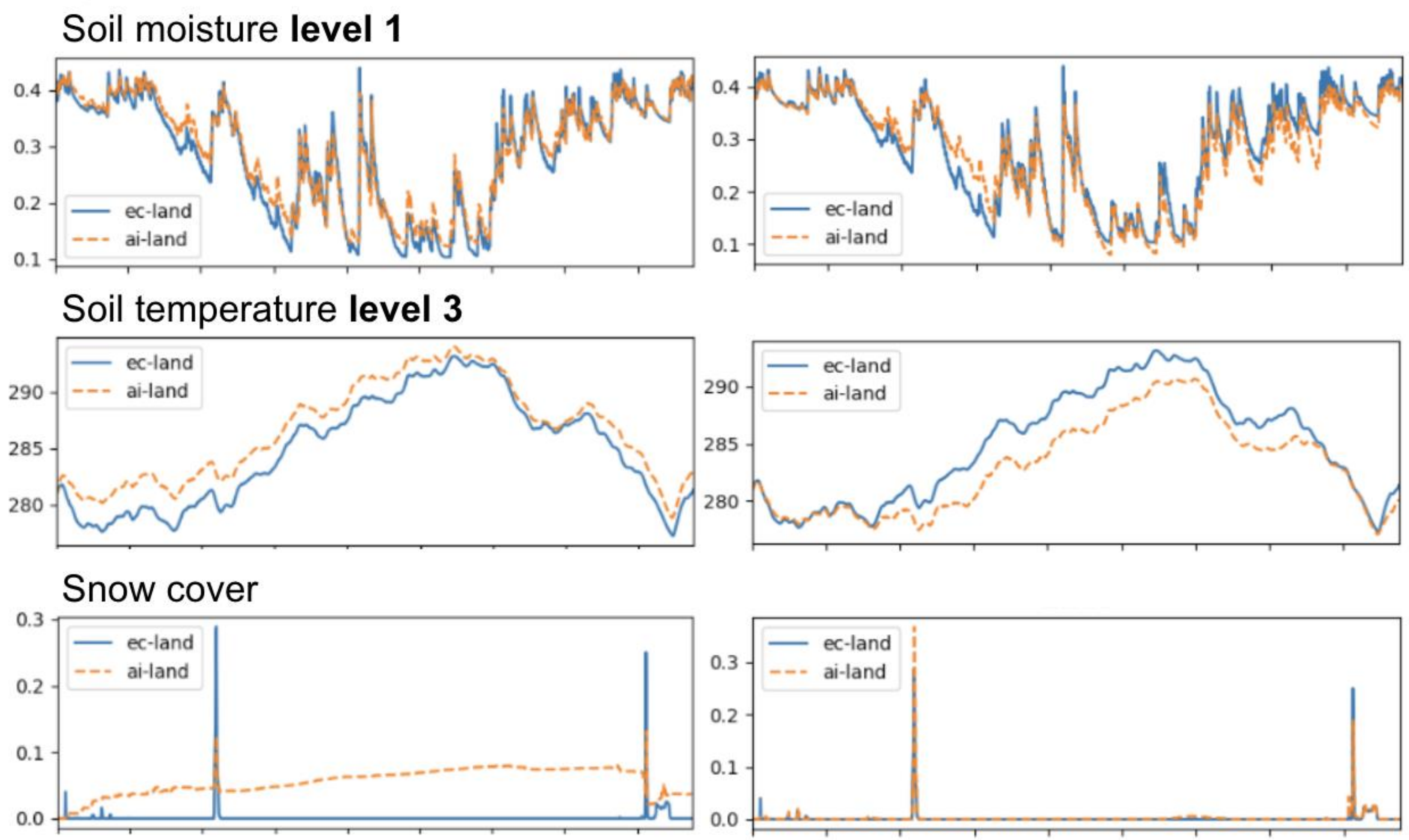


Promising results suggest we can use the AI model to improve the physical model through parameter estimation

Transferability



How does an AI model trained on **one region** perform when applied to areas with **different climates** and **conditions**?



aiLand trained on **Sahel** region

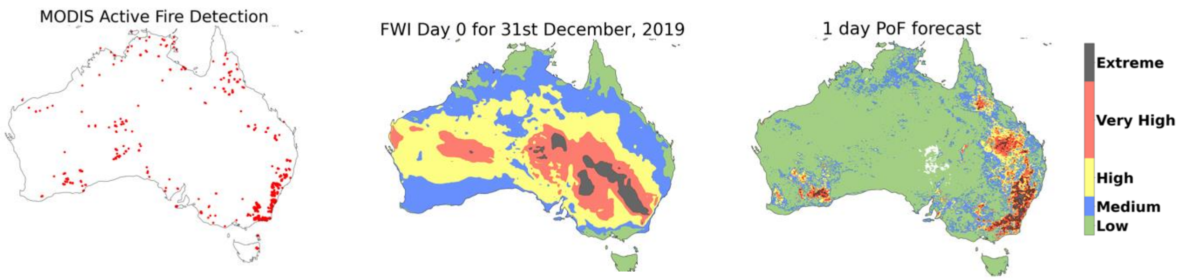
aiLand trained on **Boreal** region

Tested over point in **Europe**

Other applications of Machine Learning for Land Modelling at ECMWF

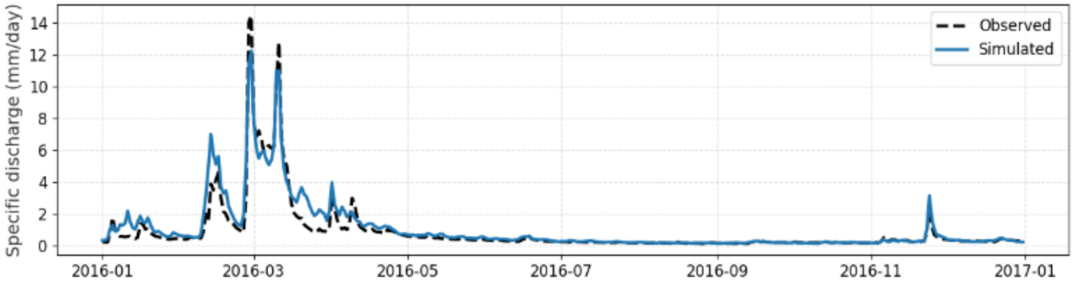
- **XGBoost** used for classification of fire likelihood

SPARKY
THE ECMWF FIRE MODEL



- **LSTM** used to predict daily streamflow at the catchment scale

AIFL

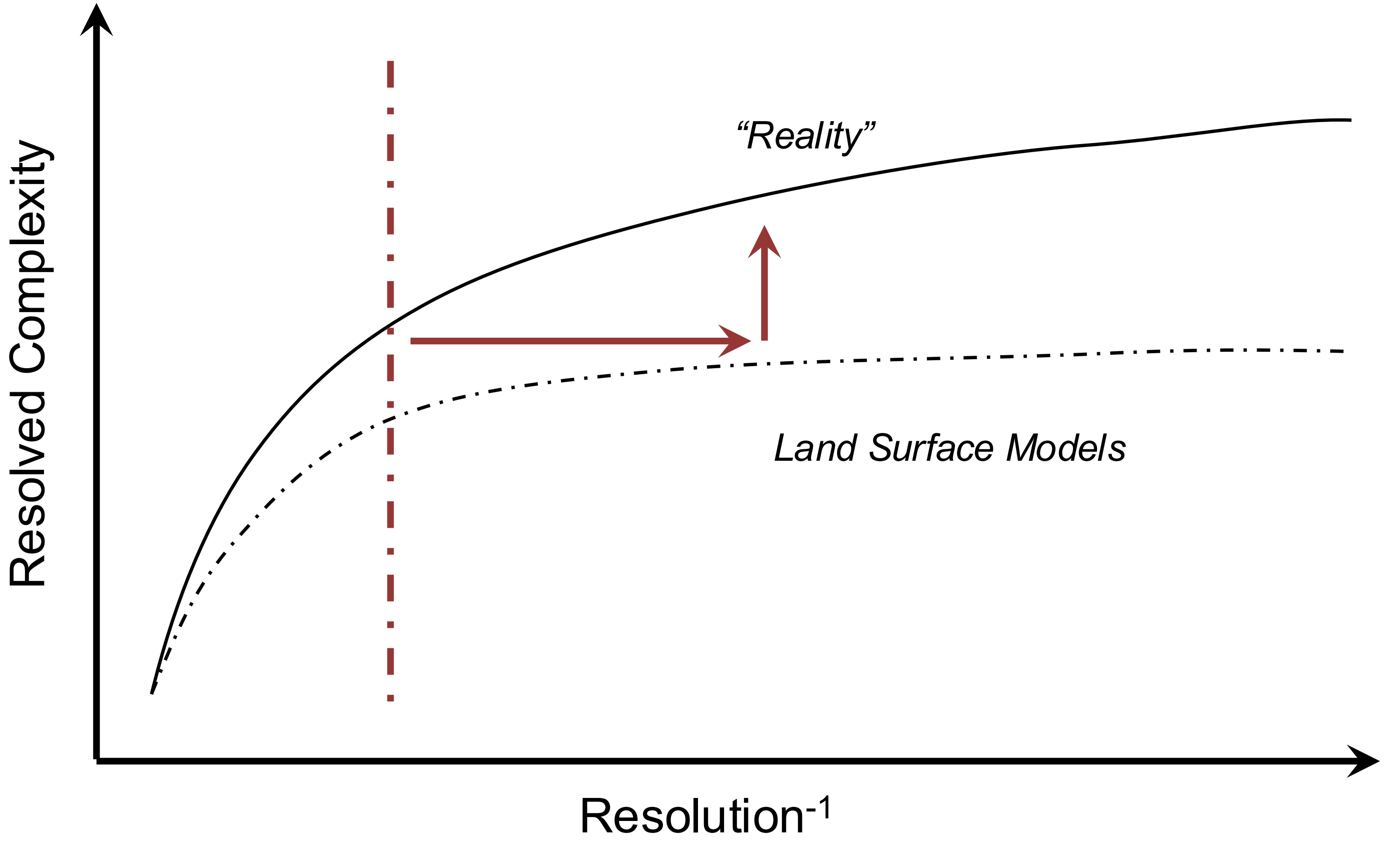


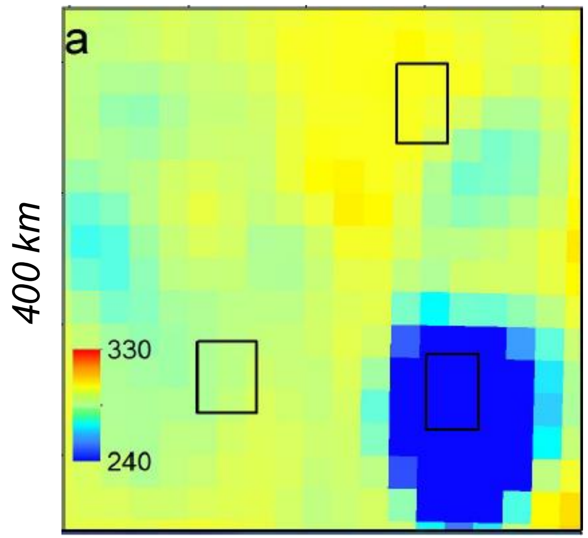
- ML used as a diagnostic tool:
  - **Gaussian process** emulators used to tune parameters through Bayesian optimization
  - **Neural network** used for quality control of surface physiographic fields using satellite Earth observations
  - **XGBoost** used as observation operators to assimilate land variable satellite retrievals

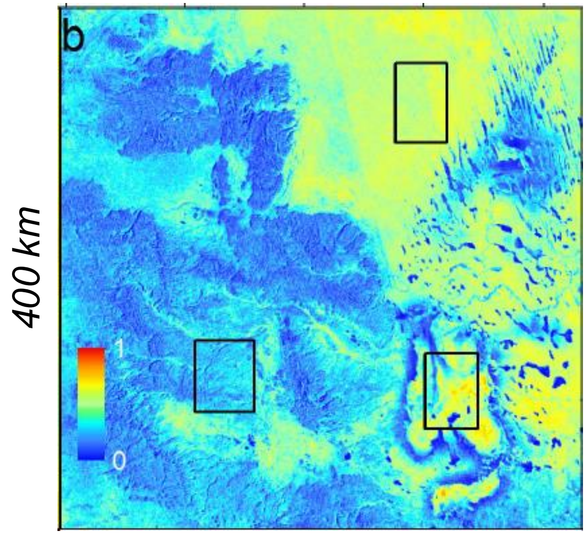
Day-to-day applications

- Streamflow forecasts
  - Use of meteorological input and poont observations
- Agricultural applications
  - Plant wetness for harvesting, optimisation of water and nutrient control, frost days etc.
- Wind and solar energy optimisation
  - Blade and panel angles, wind direction, panel efficiency, production management
- Downscaling
  - Combining high-resolution observations to downscale data
- Anomaly detection
  - On-the-fly assessment of model/observation quality
- Parameter tuning
  - Short-term model improvements

The Land Surface Model Paradigm







Open questions:

- How can we make full use of **kilometer-scale DestinE data** for land processes like soil moisture, vegetation, and water fluxes?
- How can **machine learning–based downscaling** capture local land features (topography, soil, vegetation) beyond what physics-based interpolation offers?
- What's the best way to **integrate ML with physical models** while keeping energy and water balances consistent?
- What's the optimal level of **physics awareness** to embed in learning algorithms?
- Can ML emulators trained on one Digital Twin or region be **transferred** to others?
- How do we assess **model robustness** under unseen or extreme climate scenarios?
- What does *"**trustworthy AI**"* mean for land surface predictions used in decision-making?