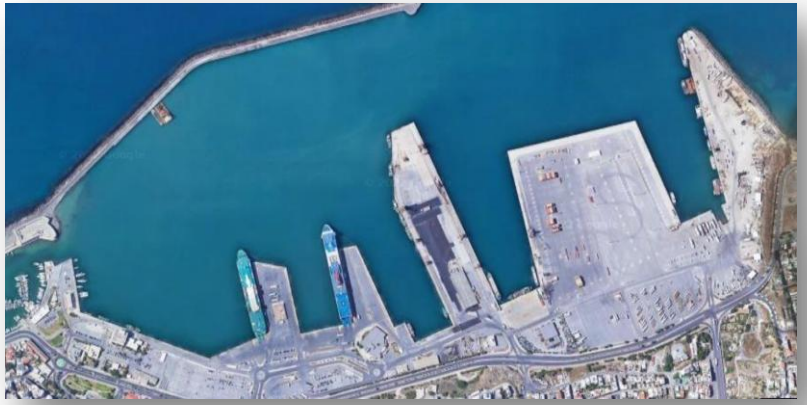
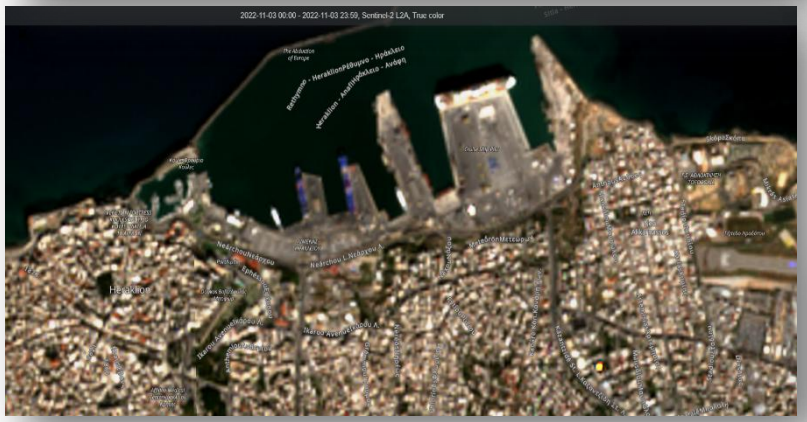
Modelling Uncertainty in Multi-Source Data Fusion

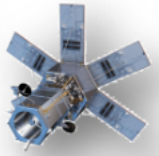
Greg Tsagkatakis

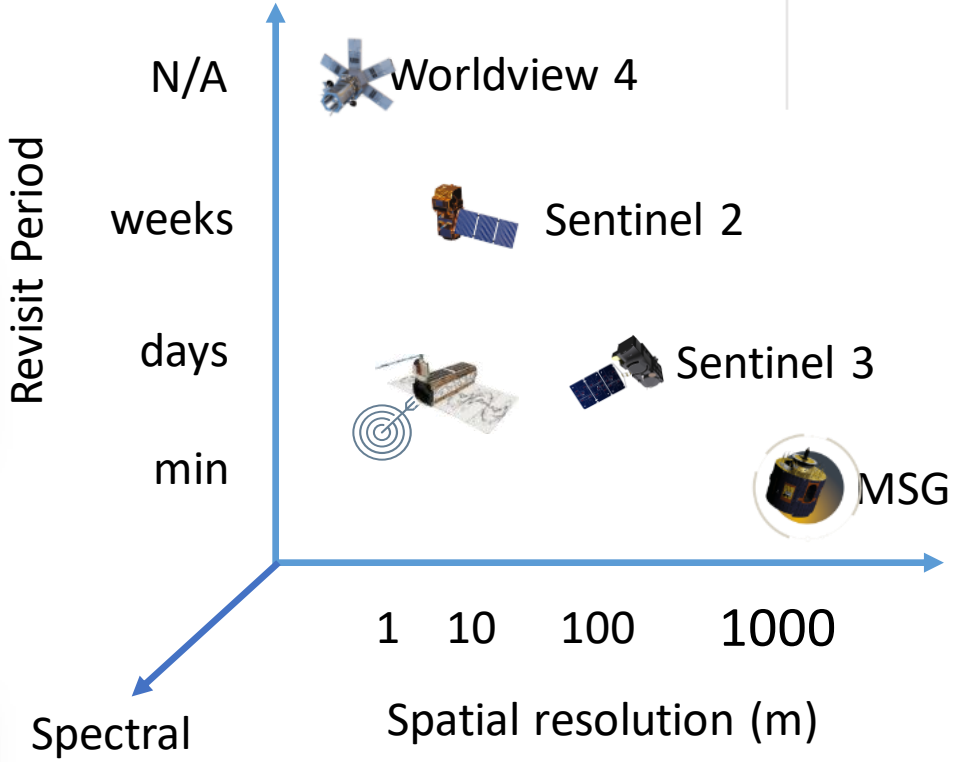
TITAN ARTIFICIAL INTELLIGENCE IN ASTROPHYSICS
FORTH INSTITUTE OF COMPUTER SCIENCE
Signal Processing Laboratory
CosmoStat
cea
Funded by the European Union

The quest for resolution

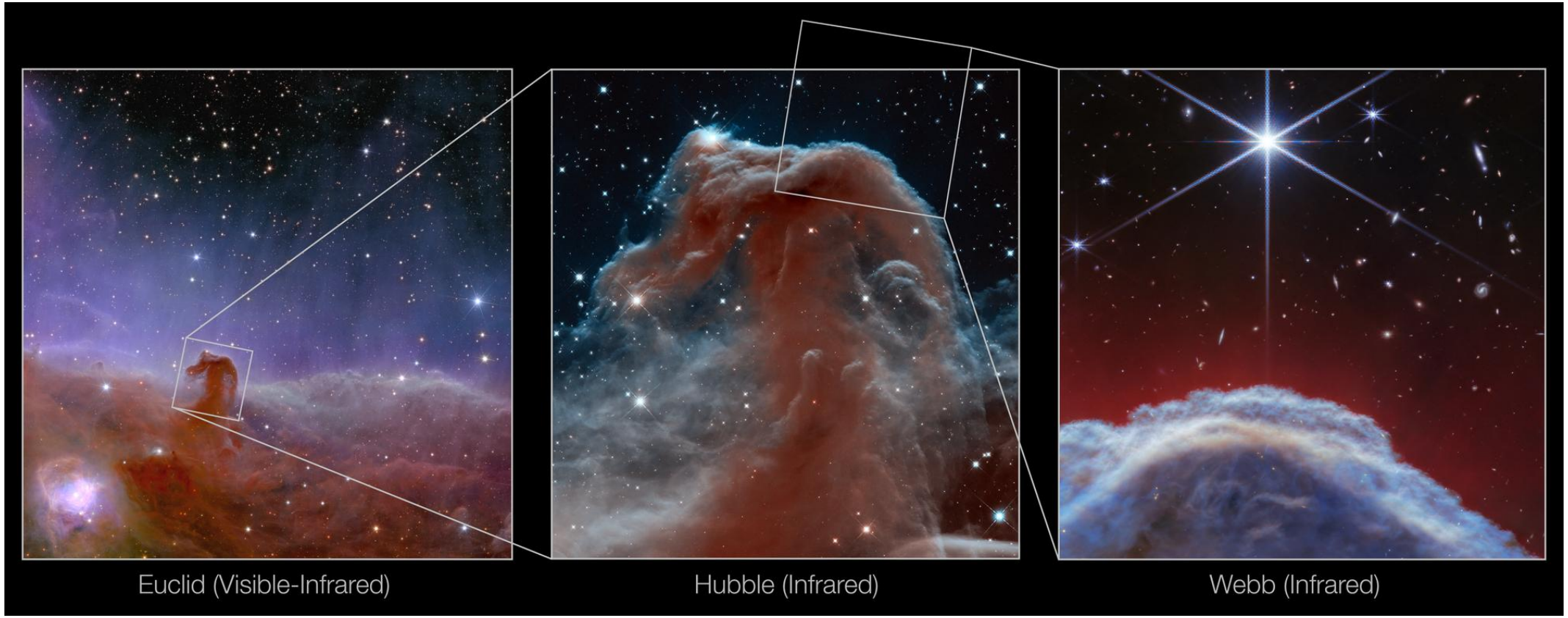








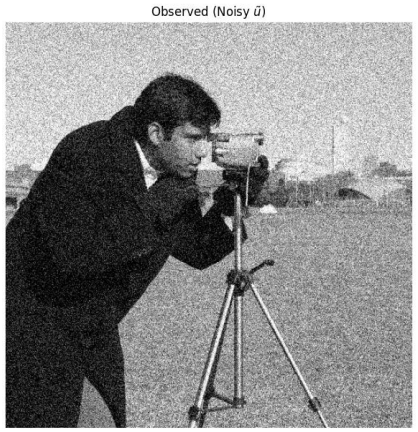
The quest for resolution



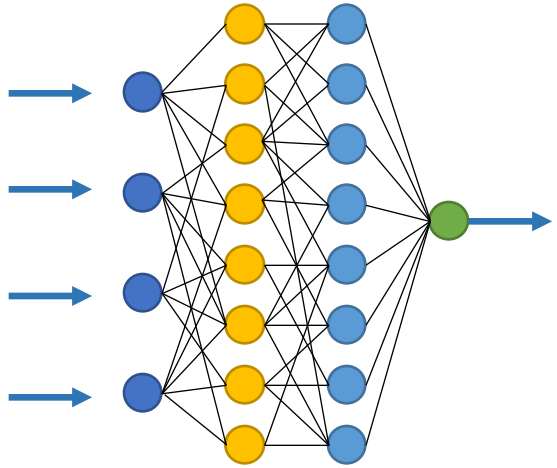
https://science.nasa.gov/asset/webb/horsehead-nebula-euclid-hubble-and-webb-images/
Image processing by J.-C. Cuillandre (CEA Paris-Saclay) and G. Anselmi

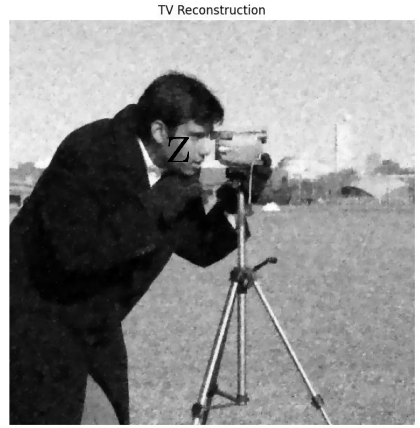
Learning based Super-Resolution

u



$\tilde{u} = Au + \varepsilon$





$\hat{u} = f(\tilde{u})$

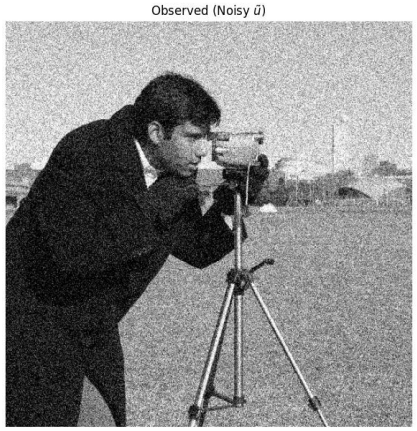
$\|\hat{u} - f(\tilde{u})\|_2$

f is a deep neural network or "any class" e.g.

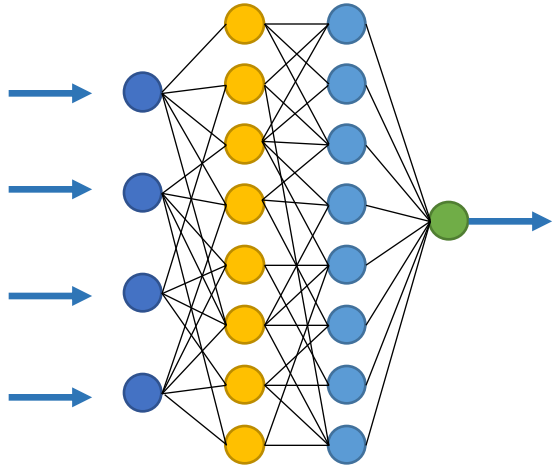
- convolutional
- autoregressive (transformer)
- diffusion

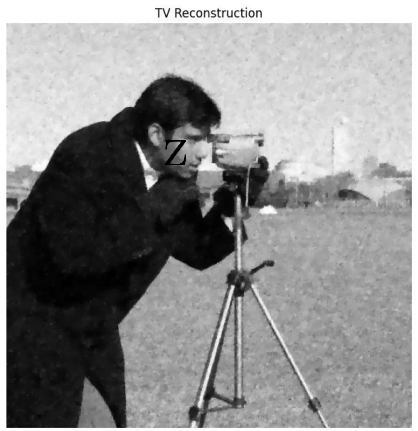
Non “trainable” regimes

u



$$\tilde{u} = Au + \varepsilon$$

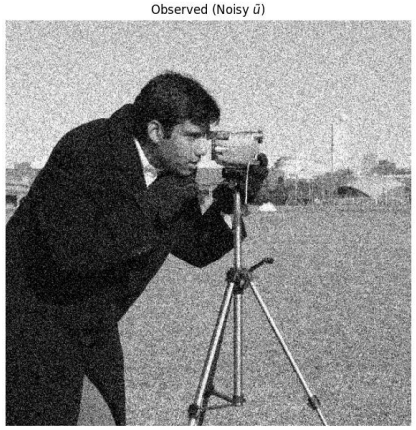


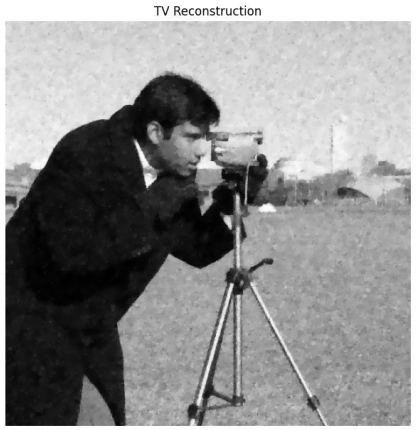


$$\hat{u} = f(\tilde{u})$$

ESA OPS-SAT
2022-2024

Classical minimization approach



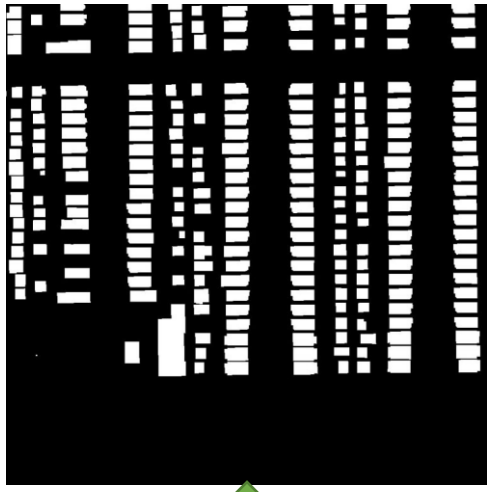


$$\tilde{u} = Au + \varepsilon$$

$$\text{Variational recovery } \min_u \; F(u) + R(u),$$

limitations

Motivation



Convex Variational Restoration with Global Feature Statistics via Probability Kernel (Measure) Lifting

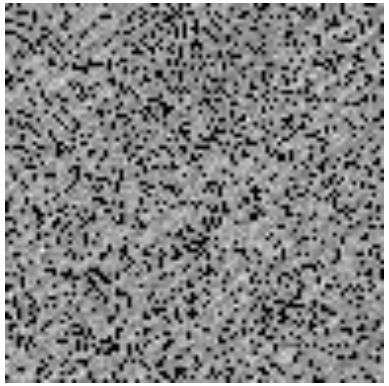
Nikos Arvanitakis

Jean-Luc Starck

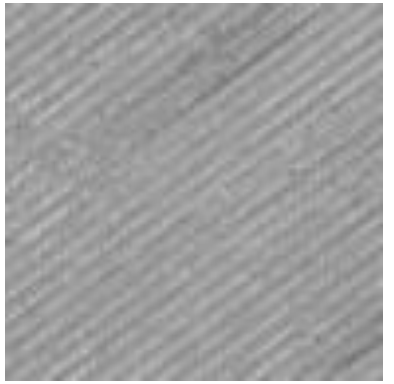
Jalal Fadili

TITAN ARTIFICIAL INTELLIGENCE IN ASTROPHYSICS
FORTH INSTITUTE OF COMPUTER SCIENCE
Signal Processing Laboratory
CosmoStat
cea
Funded by the European Union

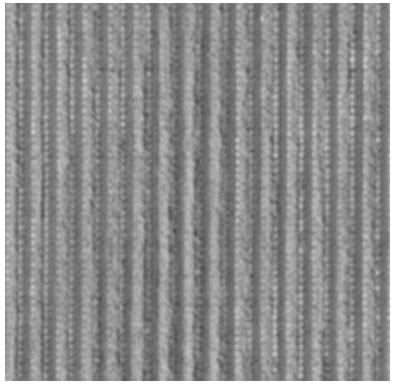
Concept

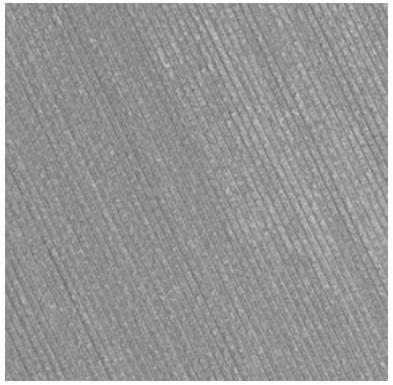


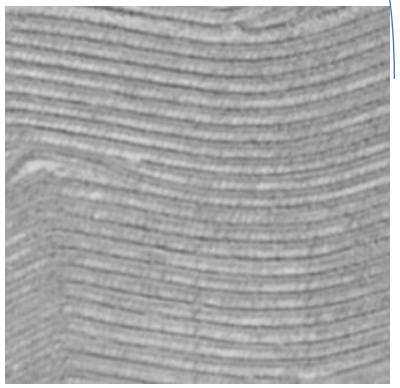
$$\min_u \; F(u) + R(u) + D(f(u), S)$$



Statistics S







"Training Set"

Measure Lifting and Mean-Relaxation

From Scalars to Distributions

Replace single-valued image $u(x)$ with a **kernel field** $\{\nu_x\}_{x\in\Omega}$, where ν_x is a probability distribution:

$$u(x) = \mathbb{E}_{\gamma\sim\nu_x}[\gamma] = \int_\Gamma \gamma \, d\nu_x(\gamma)$$

Feature Lifting via Operators B_k

For linear features $v_k = B_k u$ (e.g., gradients, filters), enforce the **mean-preserving constraint**:

$$\mathbb{E}_{z\sim\nu_y^{(k)}}[z] = (B_k u)(y)$$

Encoding Global Statistics

Linearizing Global Statistics

In this lifted space, the global feature histogram μ_k is the spatial average of local kernels:

$$\mu_k = \frac{1}{|\Omega_k|} \int_{\Omega_k} \nu_y^{(k)} \, dy$$

Because μ_k is now a **linear function** of the kernels $\nu^{(k)}$, the statistical penalty $E_k(\mu_k)$ becomes a convex term in the optimization problem.

The Wasserstein-TV (WTV) Regularizer

Limitations of Standard TV

Standard TV regularizes only the mean image, allowing local kernels to "spread" erratically if the averages remain smooth.

The Wasserstein-TV (WTV) Regularizer

WTV regularizes the kernel field $x \mapsto \nu_x^0$ directly by penalizing differences between neighboring distributions measured by a Wasserstein distance.

$$WTV(\nu^0) \approx \sum_{(i,j)\in\mathcal{N}} W(\nu_i^0, \nu_j^0)$$

The Mean-Relaxed Restoration Problem

The final restoration problem is a large-scale convex program:

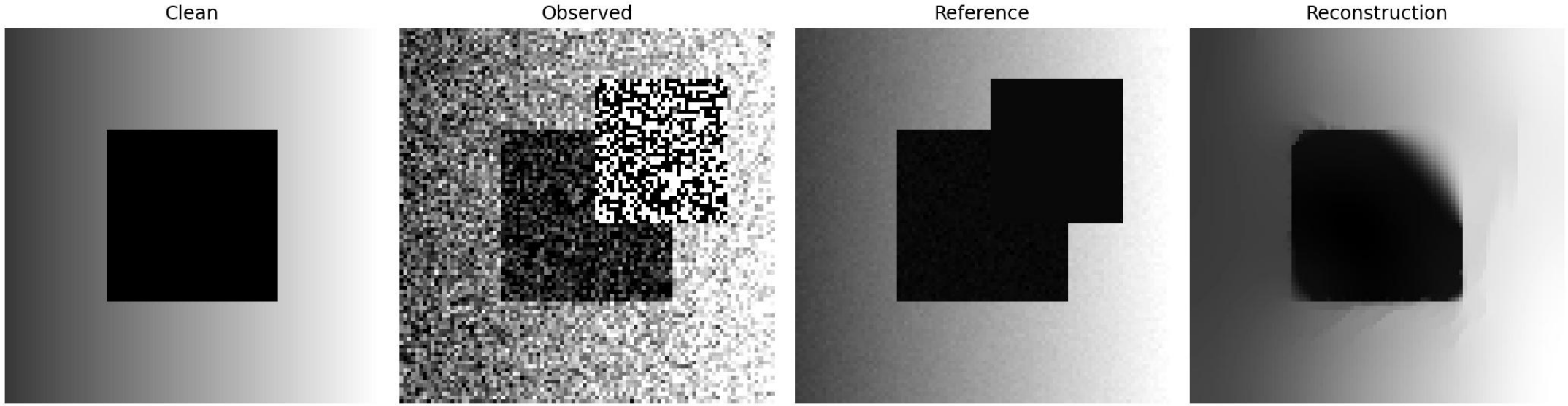
$$\overline{J}(u) = \int_{\Omega} \rho(x, u(x))\, dx + TV(u) + \sum_{k=0}^{K} \overline{E}_k(B_k u)$$

Key Technical Advantages

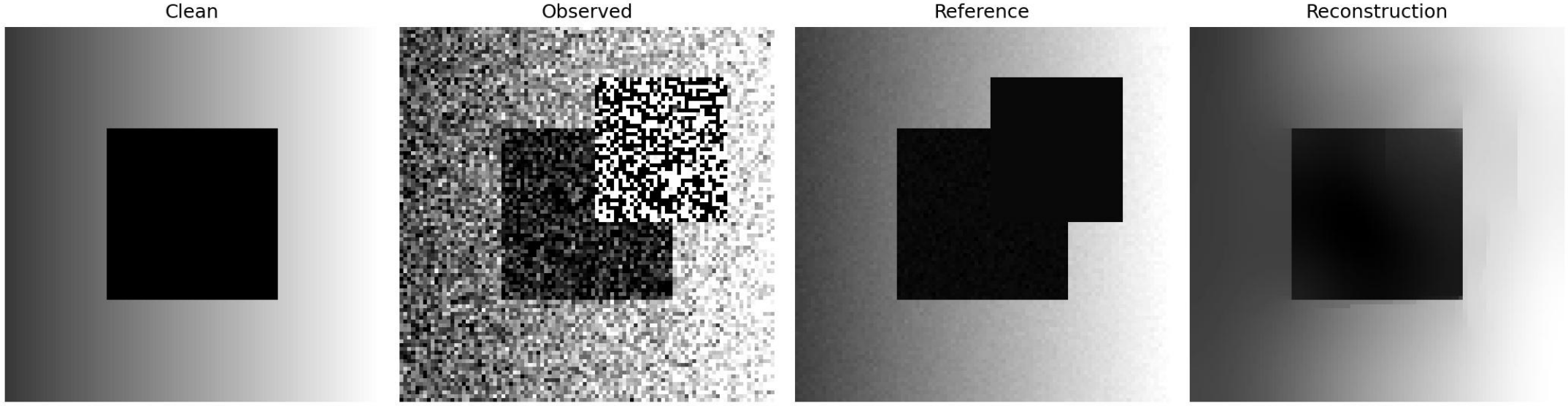
- **Global Optimality:** Convexity ensures convergence to a global minimum.
- **Arbitrary Features:** Can enforce statistics for intensities, filter responses, or directional derivatives.

Indicative results

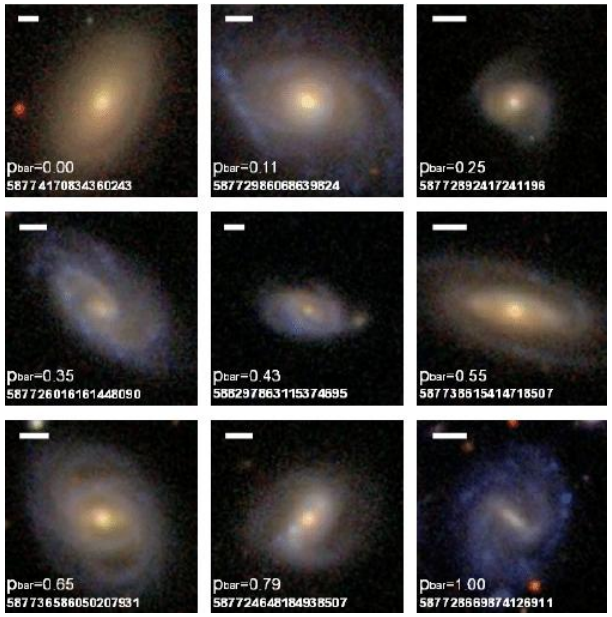
TV regularization

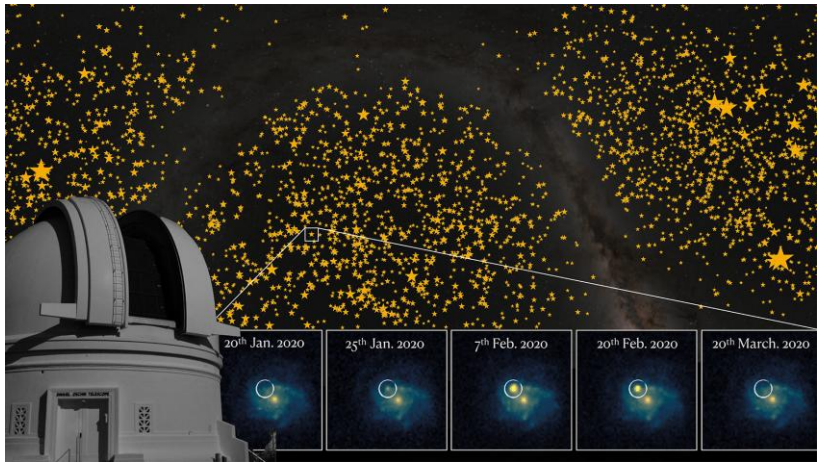


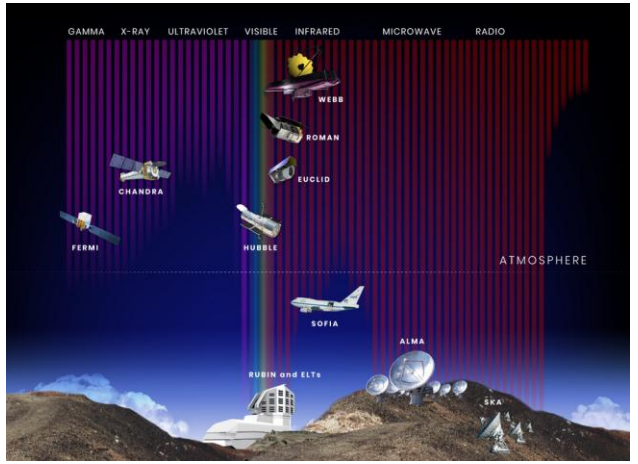
WTV regularization



Multi-source data fusion: Astronomy







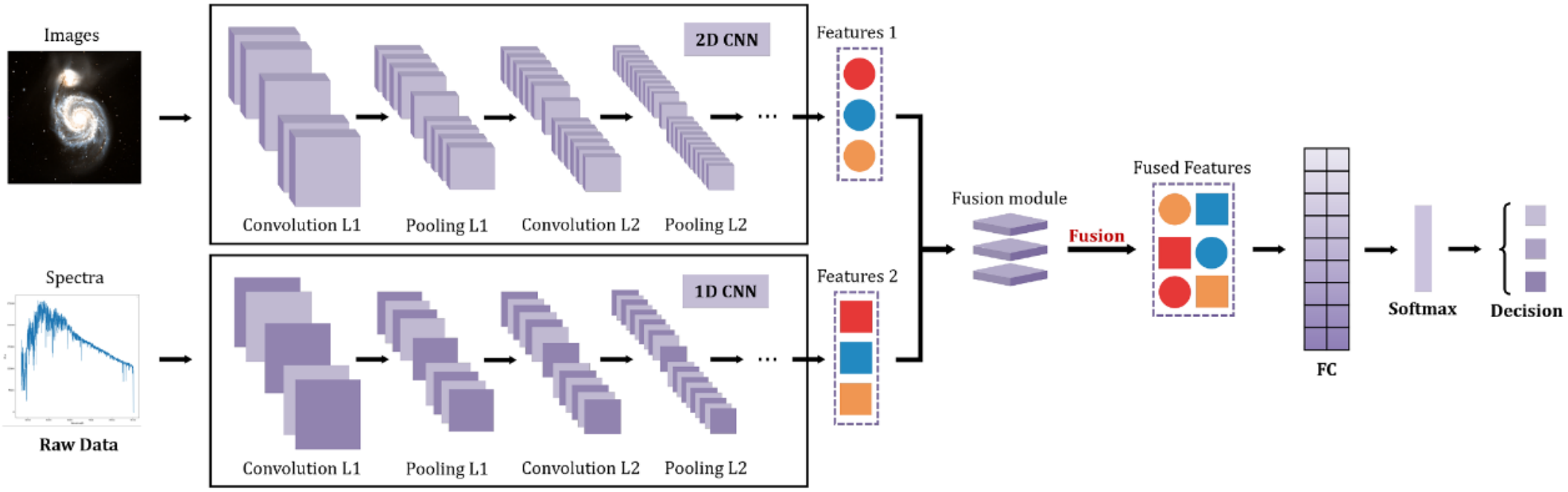
SDSS

- Optical (u, g, r, i, z bands)
- Moderate spatial resolution
- Massive catalog with spectroscopy

ZTF

- High-cadence optical
- Light curves

Exemplary fusion pipeline



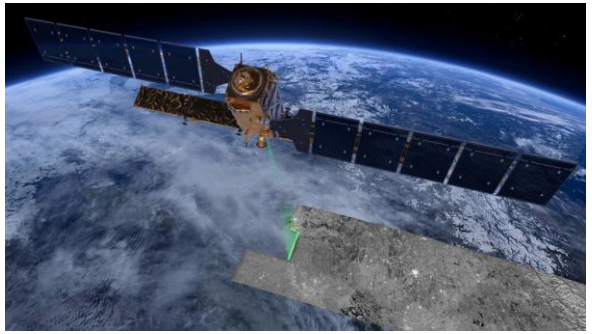
Multi-source data fusion: Earth Observation

In-situ

+ High quality

+ High temporal resolution

- Localized



Remote Sensing

+ Global coverage

+ Moderate temporal resolution

- Coarse-resolution

Numerical models

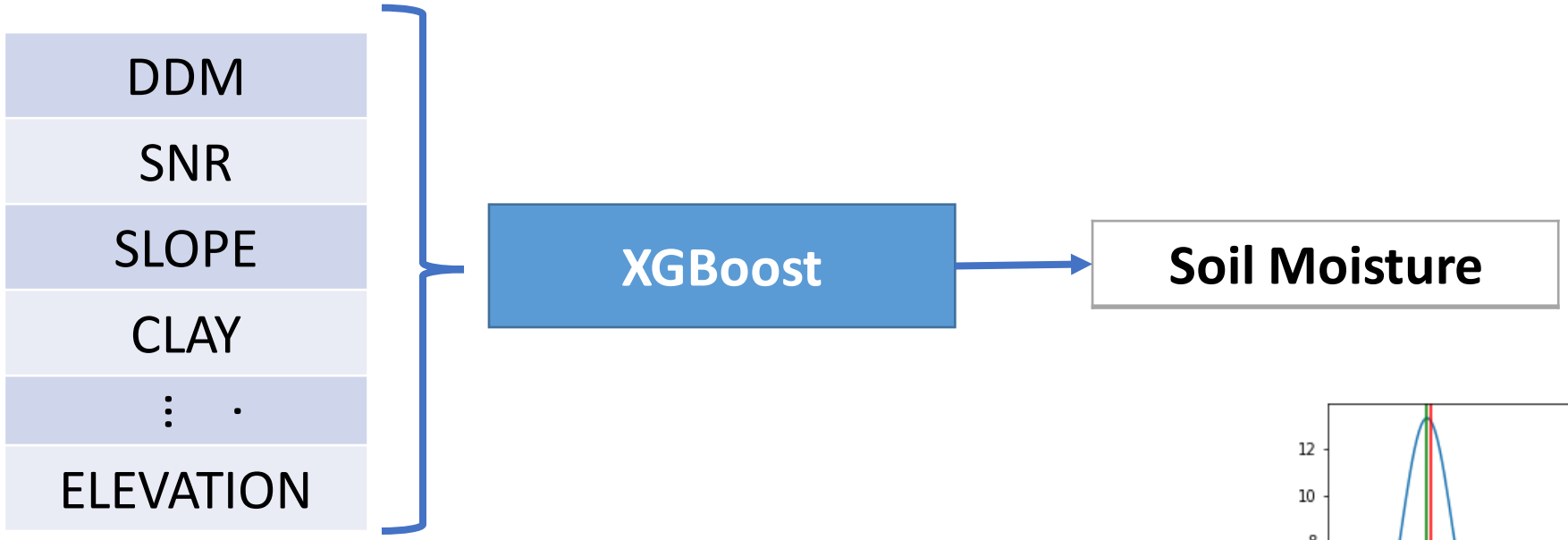
+ Capture physical laws

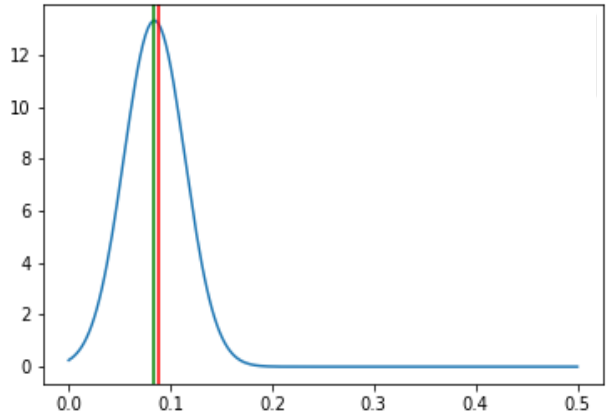
+ Flexible spatial/temporal resolution

- Expensive

ML-based retrieval

- XGBoost: regularized (L1 & L2) Gradient Boosting

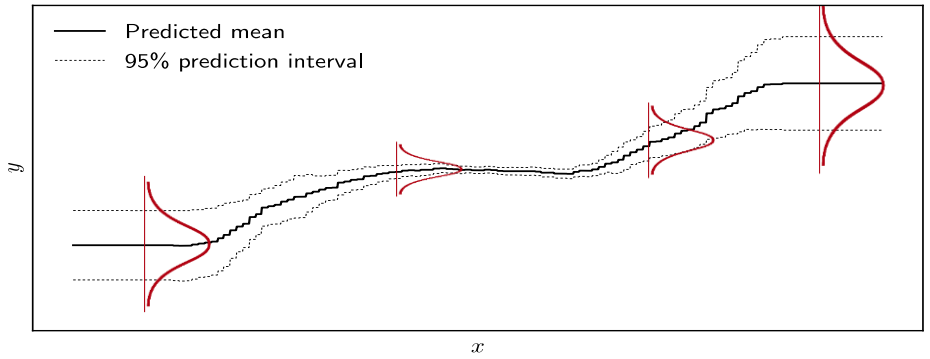




Question	ML answer
What is the Soil Moisture	0.02 cm^3/cm^3
How certain are you	??

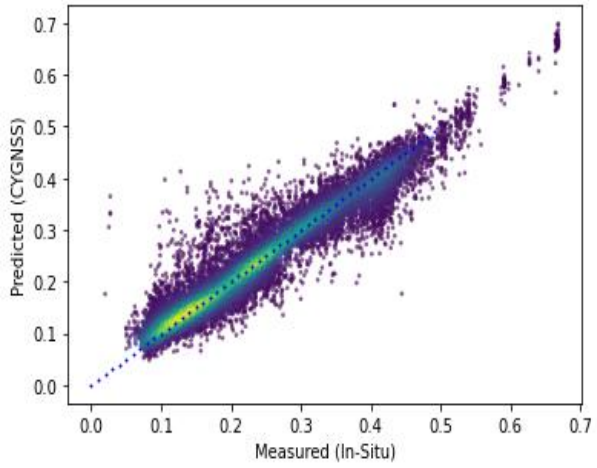
Probabilistic regression

- NGBoost: Natural gradient boosting for probabilistic prediction

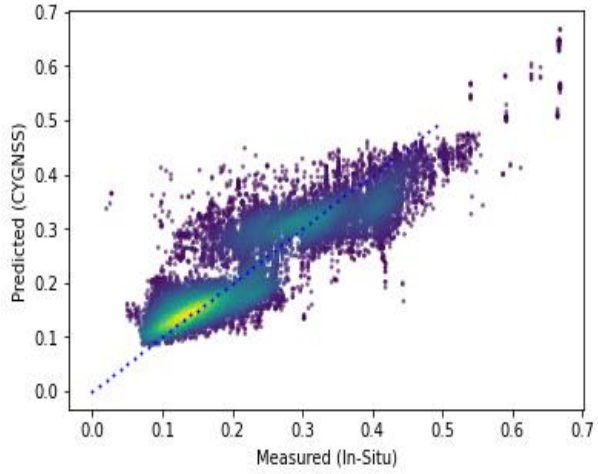


	RMSE $[m^3/m^3]$	ubRMSE $[m^3/m^3]$
SMAP (baseline)	0.103 (0.005)	0.151 (0.007)
CYGNSS (XGBoost)	0.055 (0.002)	0.059 (0.001)
CYGNSS (NGBoost)	0.058 (0.003)	0.060 (0.001)

XGBoost



NGBoost

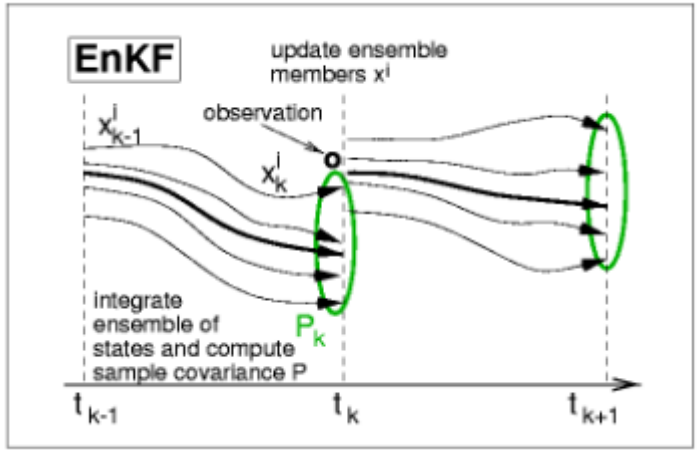


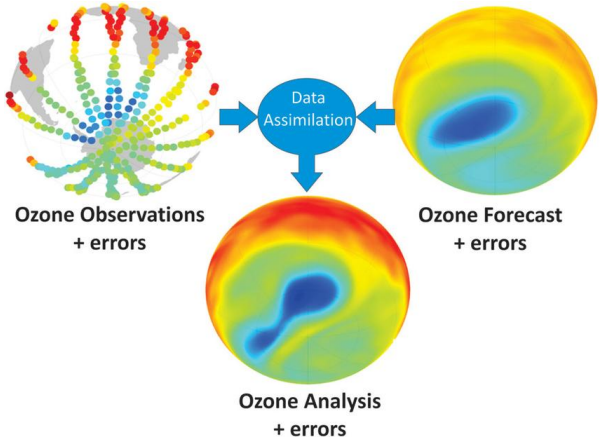
Data Assimilation (DA)

Data assimilation determines how much model states should be updated based on the observed model output.

$$\mathbf{y}_t = \mathbf{H}_t \mathbf{x}_t + \mathbf{v}_t, \qquad \mathbf{v}_t \sim \mathcal{N}_{m_t}(\mathbf{0}, \mathbf{R}_t),$$
$$\mathbf{x}_t = \mathbf{M}_t \mathbf{x}_{t-1} + \mathbf{w}_t, \qquad \mathbf{w}_t \sim \mathcal{N}_n(\mathbf{0}, \mathbf{Q}_t),$$

$\mathbf{y}_t$ is the observed data vector, $\mathbf{x}_t$ is the unobserved state vector





Multi-Source Uncertainty Aware Fusion for Soil Moisture Estimation

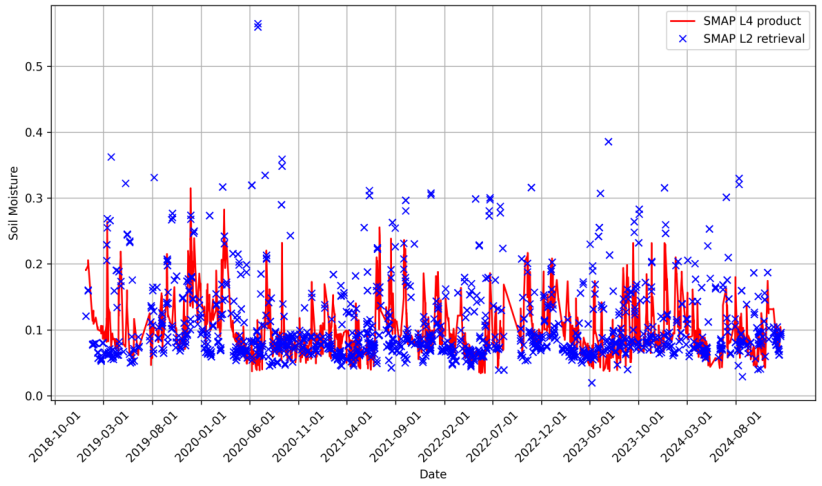
L. Polychronakis

M. Moghaddam (USC)

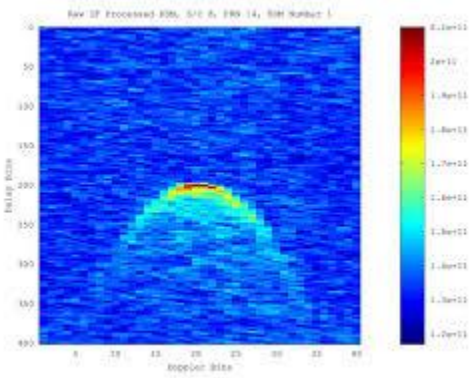
TITAN ARTIFICIAL INTELLIGENCE IN ASTROPHYSICS
FORTH INSTITUTE OF COMPUTER SCIENCE
Signal Processing Laboratory
CosmoStat
cea
Funded by the European Union

Proposed approach

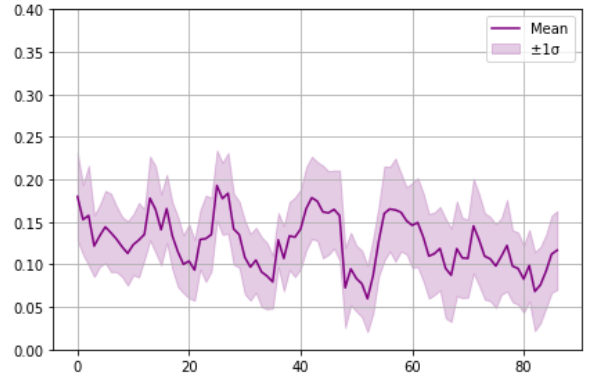
Data Assimilation



Observations



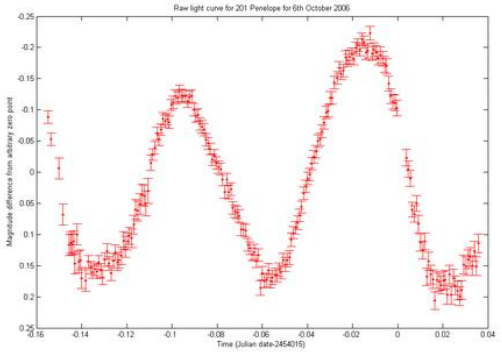
$$p(y)p(x \mid y) \propto p(y \mid x)$$



Estimation

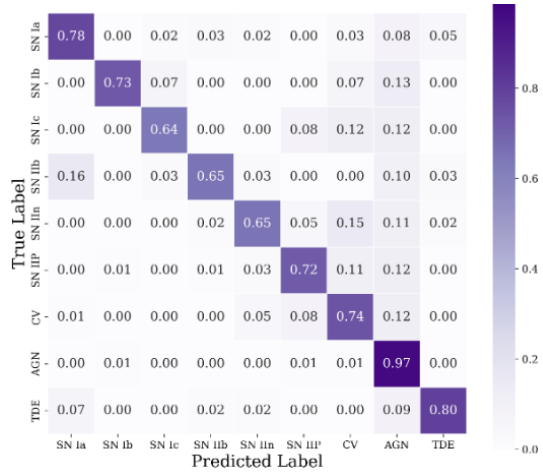
Applications in astrophysics

Light curves



Images

$$p(y)p(x \mid y) \propto p(y \mid x)$$



Estimation

Gaussian fusion

Retrieval

- $\mathbf{x}_t^{(i)} \in \mathbb{R}^{d_x}$: remote sensing observations
- $\mathbf{z}_t^{(i)} \in \mathbb{R}^{d_z}$: ancillary observations
- $y_t^{(i)} \in \mathbb{R}$: surface soil moisture value at time t.

Forecasting

- Input $y_{t-\tau+1:t}^{(i)}$: SM over a window of size τ.
- Output $y_{t+\Delta t}^{(i)} \in \mathbb{R}$: SM at $t + \Delta t$

$$\mathcal{L}_{\text{retrieval}} = \sum_{i=1}^{N} \left(\frac{1}{2\sigma^2} (y^{(i)} - \mu)^2 + \log \sigma \right)$$

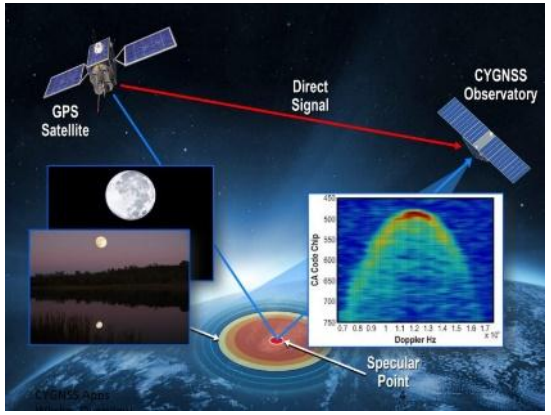
$$\mathcal{L}_{\text{forecasting}} = \sum_{i=1}^{N} \left(\frac{1}{2\sigma^2} (y_{t+\Delta t}^{(i)} - \mu)^2 + \log \sigma \right).$$

$$p(y \mid \mathbf{x}, \mathbf{z}) \sim \mathcal{N}(\mu_p, \sigma_p^2)$$

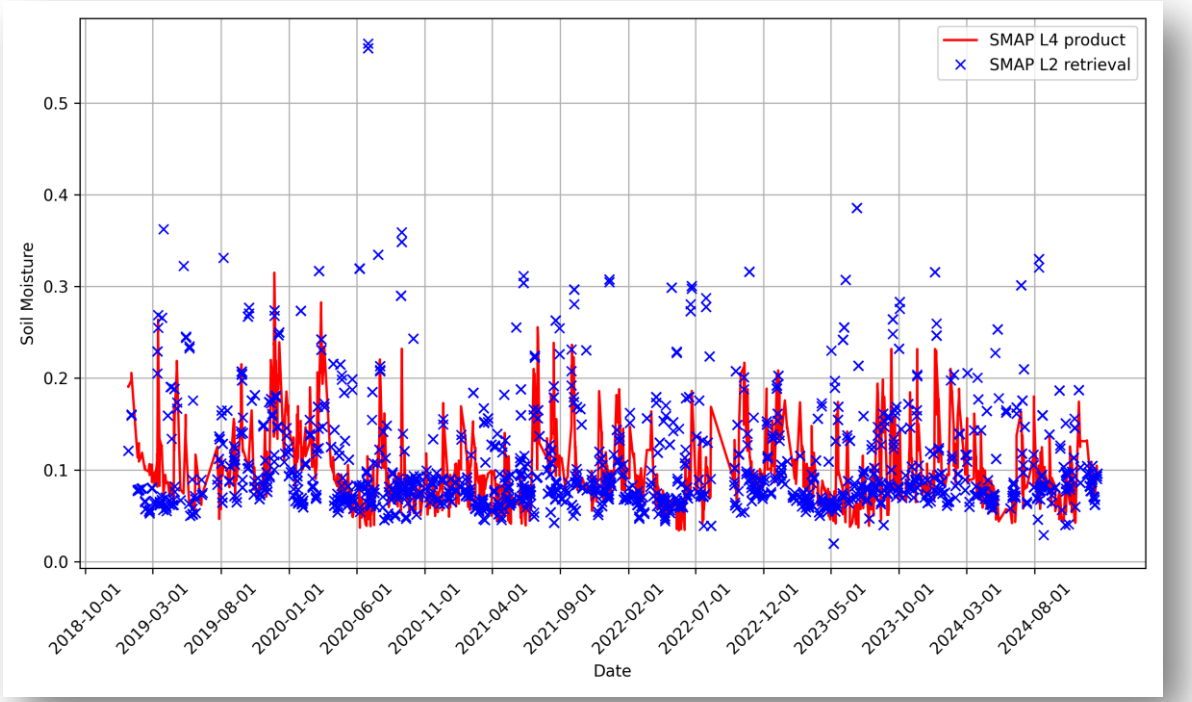
$$\mu_p = \sigma_p^2 \left(\frac{\mu_f}{\sigma_f^2} + \frac{\mu_r}{\sigma_r^2} \right)$$

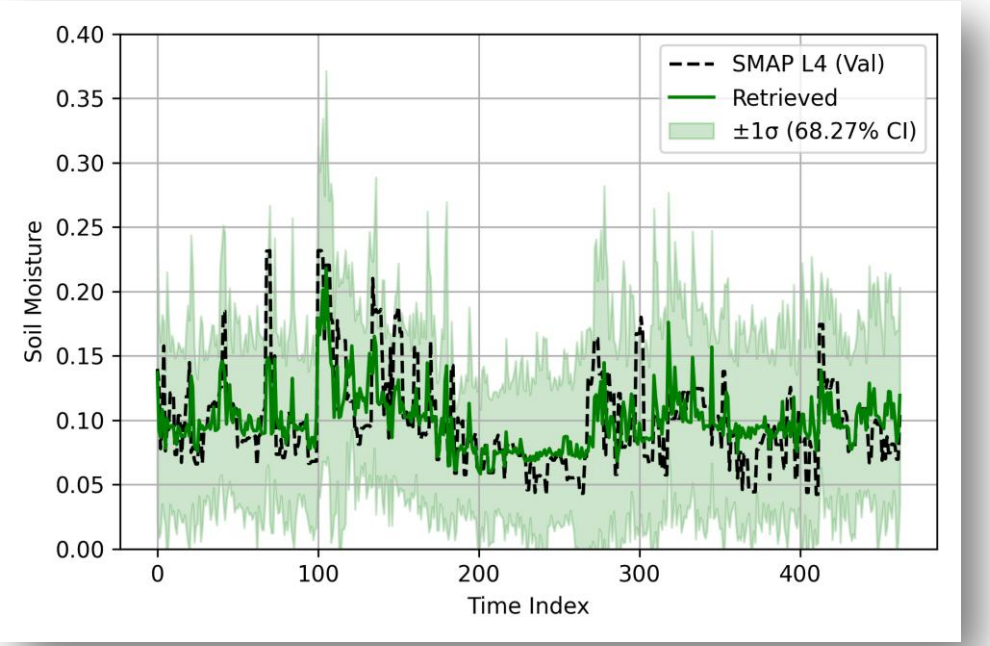
$$\sigma_p^2 = \left(\frac{1}{\sigma_f^2} + \frac{1}{\sigma_r^2} \right)^{-1}$$

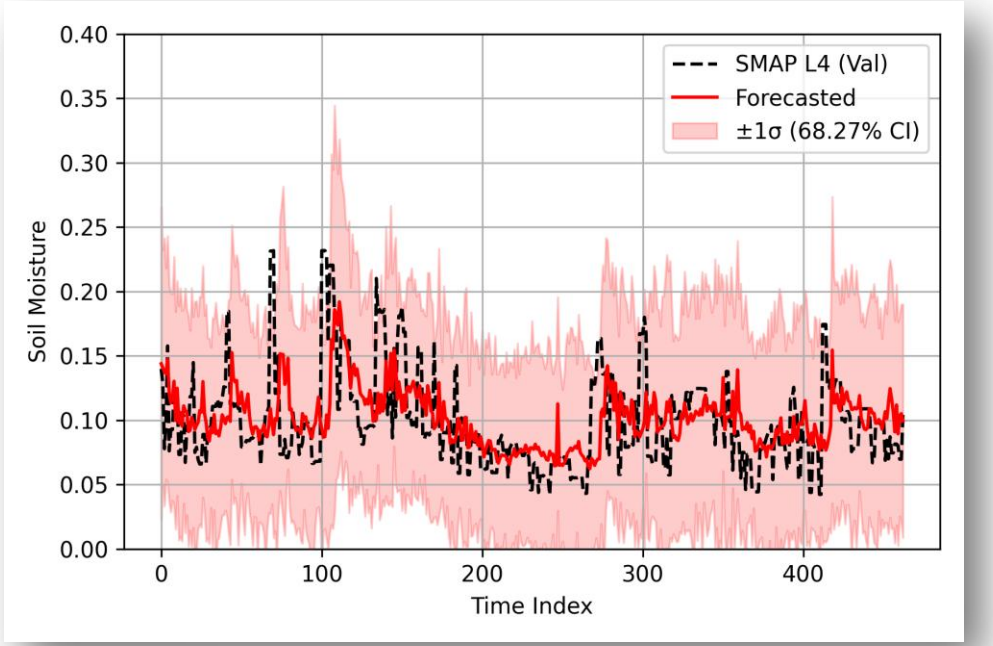
Case study: soil moisture

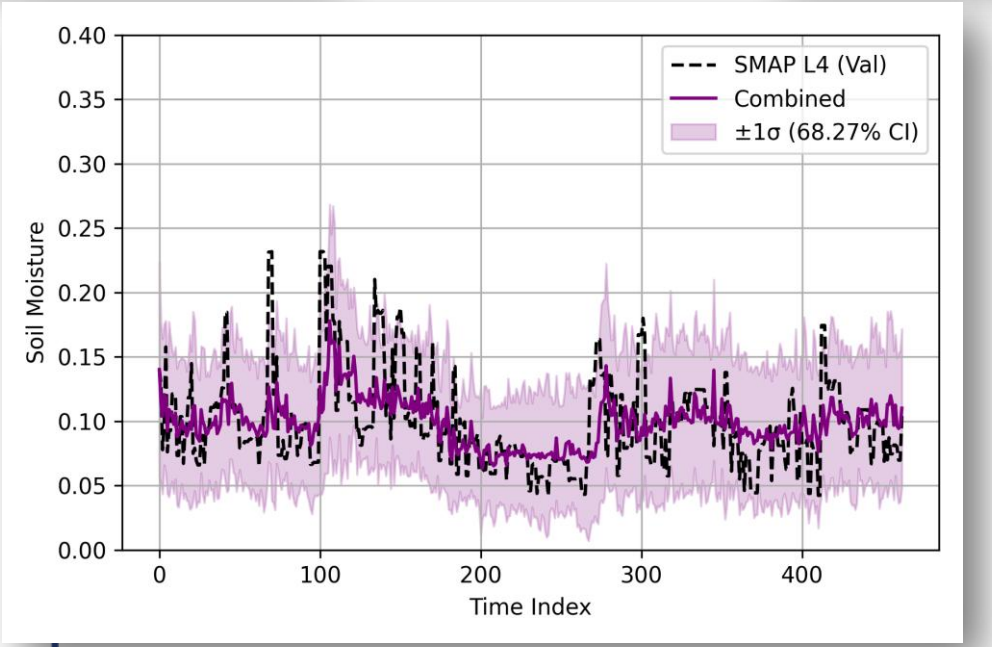


JR-1, JR-2, JR-3 (NM),
Kendall & Lucky Hills (AZ)
Z1 and Z4 in CO.




SMAP L4 (Val)
Retrieved
±1σ (68.27% CI)
Soil Moisture
Time Index
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0
100
200
300
400


SMAP L4 (Val)
Forecasted
±1σ (68.27% CI)
Soil Moisture
Time Index
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0
100
200
300
400

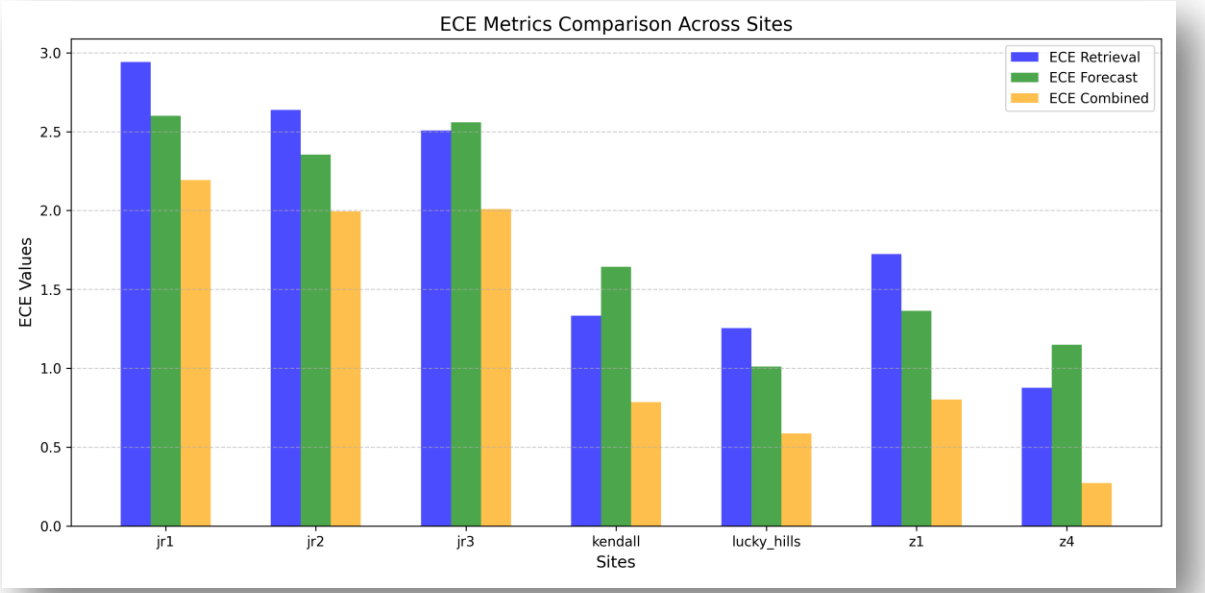

SMAP L4 (Val)
Combined
±1σ (68.27% CI)
Soil Moisture
Time Index
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0
100
200
300
400

Performance metrics

ubRMSE

Site	Retrieval	Forecast	Combined
JR-1	**0.0281**	0.0318	0.0282
JR-2	**0.0197**	0.0242	0.0203
JR-3	0.0248	0.0310	**0.0247**
Kendall	0.0411	0.0462	**0.0387**
Lucky Hills	0.0357	0.0450	**0.0356**
Z1	**0.0442**	0.0608	0.0522
Z4	0.0322	0.0404	**0.0309**

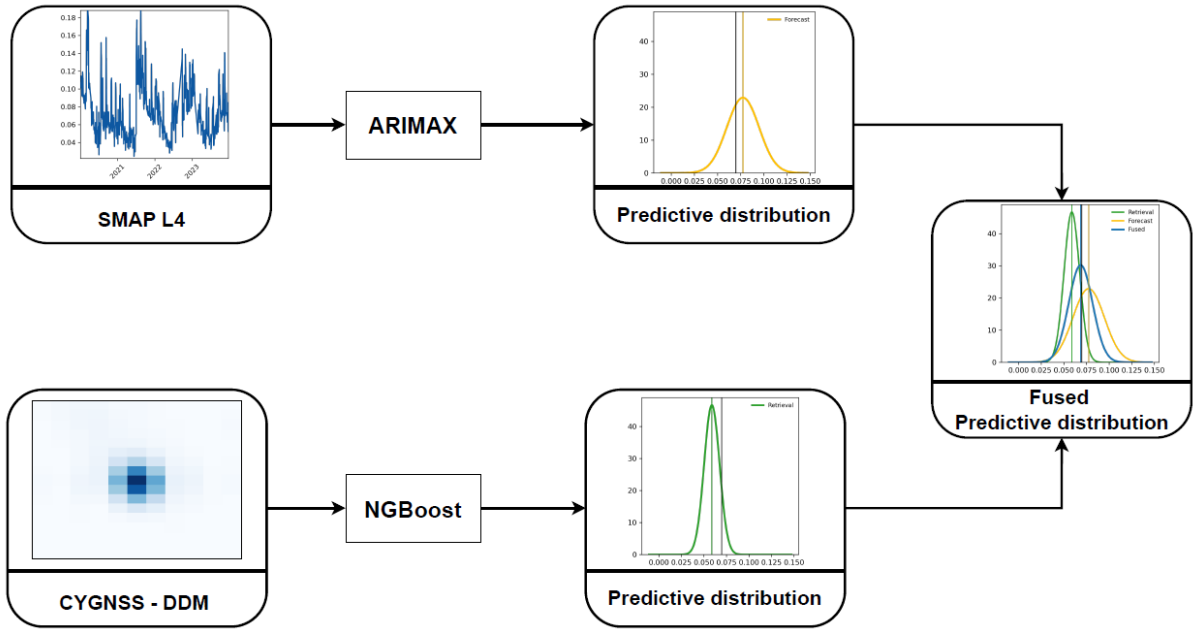
Expected Calibration Error



Limitations

- **Gaussianity:** each expert is reduced to (μ, σ^2); skew/heavy-tailed errors are not represented.
- **Variance-sensitive:** Inverse-variance fusion always reduces σ; if assumptions fail, the reduction is too aggressive $\rightarrow$ overconfidence.
- **No quality gating:** intermittent/low-quality retrievals can dominate if they report small σ^2.

Fusing predictive distributions



- We combine $p_{\text{ret}}(y_t)$ and $p_{\text{fcst}}(y_t)$ into a single $p_{\text{fused}}(y_t)$.
- **Wasserstein-2 barycenter**: minimizes the (weighted) sum of squared Wasserstein distances to the experts.

Distances between distributions

- **Standard L_2 distance (between densities):**

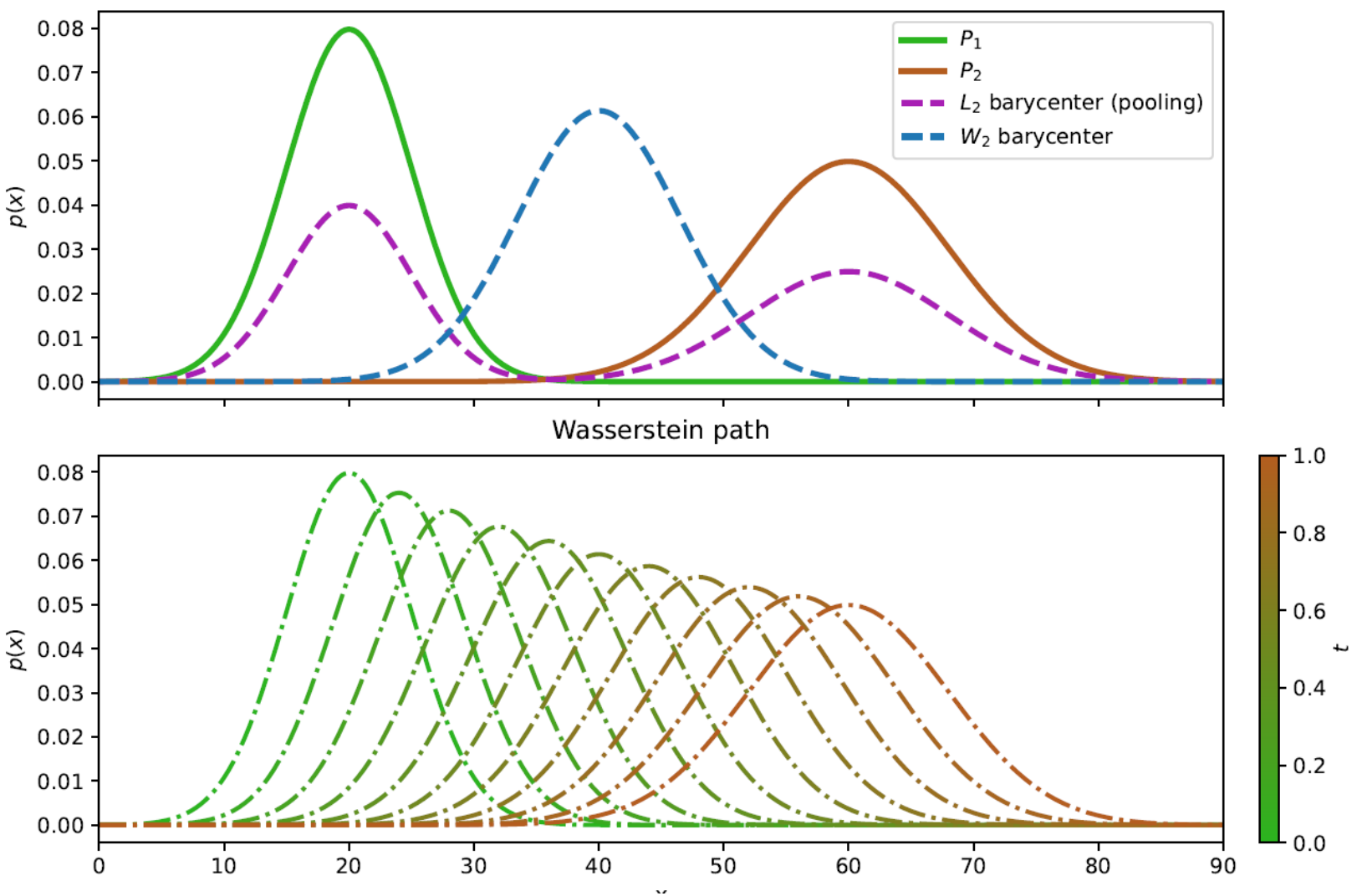
$$\|p - q\|_2 = \left(\int_{\mathbb{R}} (p(x) - q(x))^2 \, dx \right)^{1/2}$$

- **Wasserstein-2 distance ("earth mover"):**

$$W_2^2(p, q) = \inf_{\pi \in \Pi(p,q)} \int_{\mathbb{R} \times \mathbb{R}} (x - y)^2 \, d\pi(x, y)$$

where $\Pi(p, q)$ is the set of couplings with marginals p and q.

Wasserstein path



Wasserstein Barycenter fusion (1D)

- We fuse retrieval and forecast by minimizing W_2^2 distances:

$$p_{\text{fused}} = \arg\min_p \; w\, W_2^2(p, p_{\text{ret}}) + (1-w)\, W_2^2(p, p_{\text{fcst}})$$

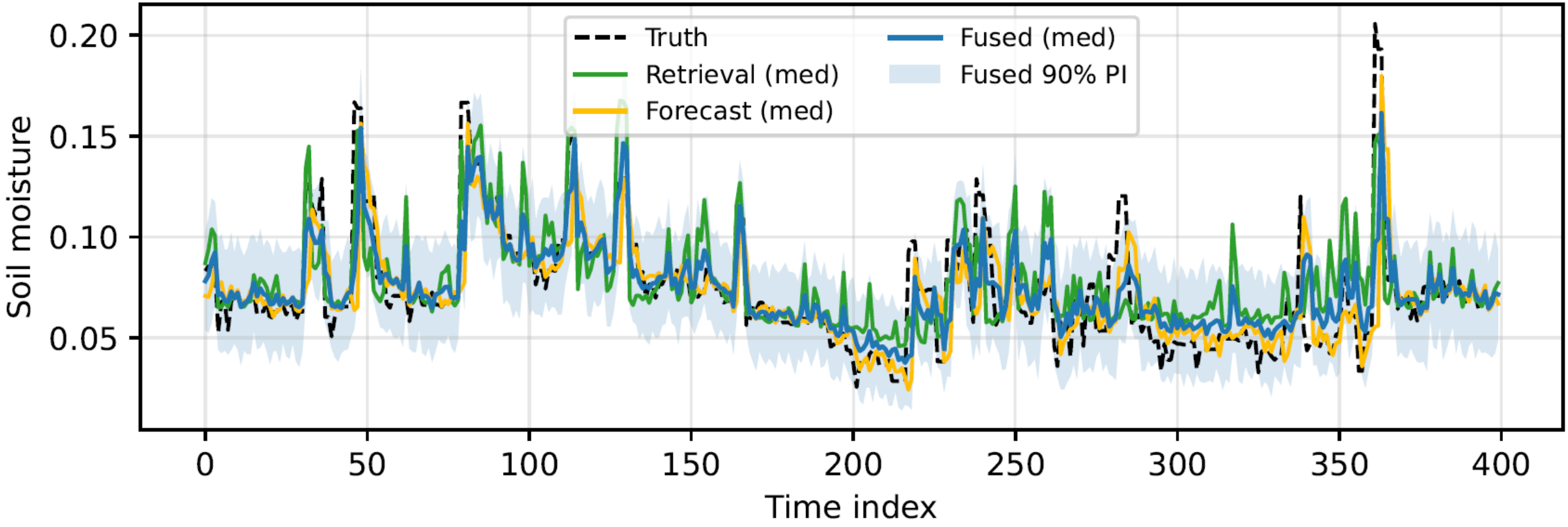
- In **1D**, W_2 has a closed form via quantiles:

$$W_2^2(p,q) = \int_0^1 \left(Q_p(\tau) - Q_q(\tau)\right)^2 d\tau$$

- And the barycenter is **quantile averaging**:

$$Q_{\text{fused}}(\tau) = w\, Q_{\text{ret}}(\tau) + (1-w)\, Q_{\text{fcst}}(\tau)$$

Indicative behavior



State-space models

The Forecast Backbone: State-Space ARIMA

$$x_t = Ax_{t-1} + w_t, \quad w_t \sim \mathcal{N}(0, Q)$$

$$y_t = Hx_t + v_t, \quad v_t \sim \mathcal{N}(0, R)$$

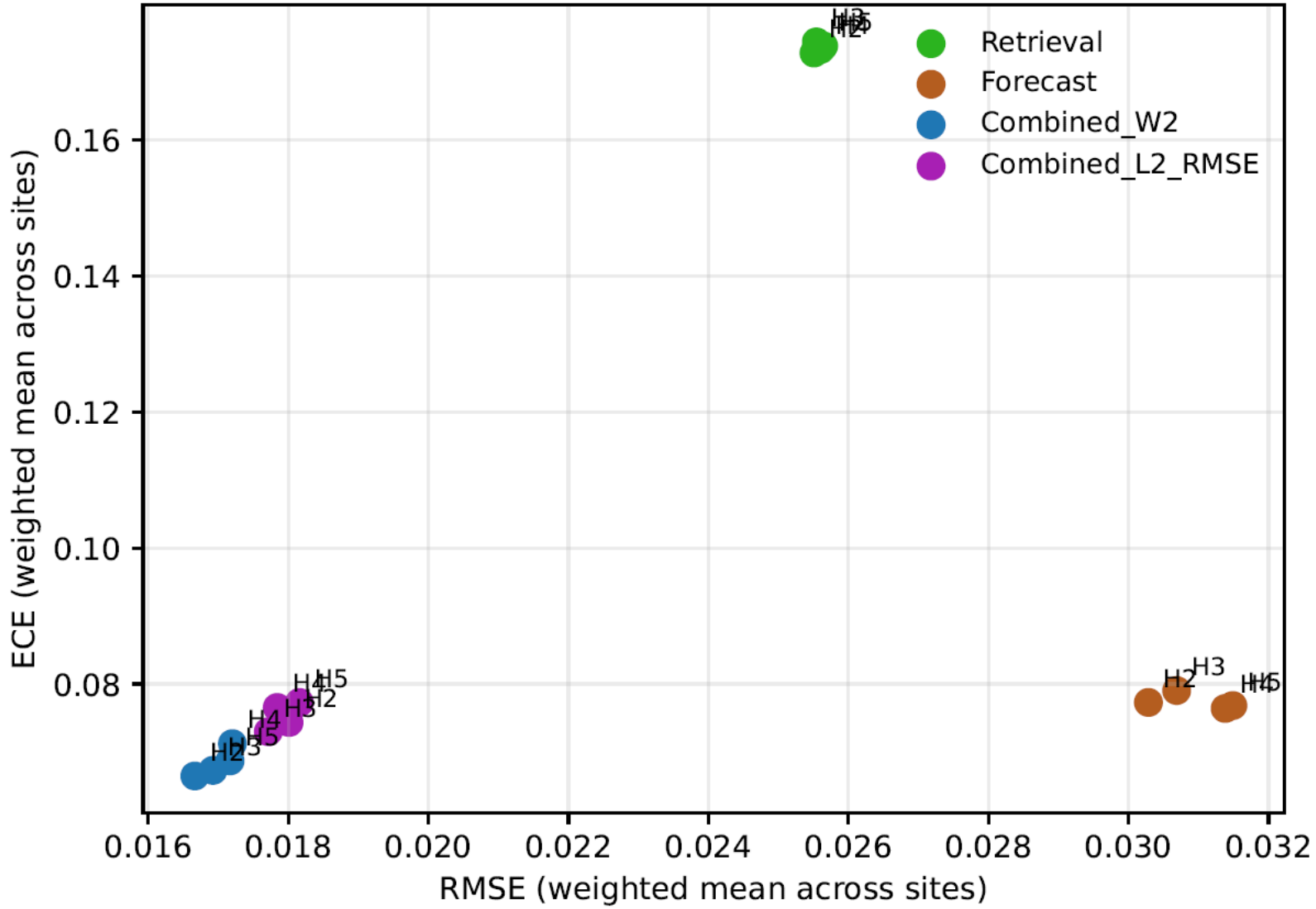
Kalman Filtering for Gaps and Latency: If data is missing or delayed, the "Update" step is skipped, but the "Predict" step still valid:

$$\hat{x}_{t|t-1} = A\hat{x}_{t-1|t-1}, \quad P_{t|t-1} = AP_{t-1|t-1}A^T + Q$$

Adaptive Weighting with Forgetting

$$\min_{w \in [0,1]} \sum_{k \leq t} \gamma^{t-k} (y_k - (wm_r(k) + (1-w)m_f(k)))^2$$

Results


Retrieval
Forecast
Combined_W2
Combined_L2_RMSE
ECE (weighted mean across sites)
0.16
0.14
0.12
0.10
0.08
RMSE (weighted mean across sites)
0.016
0.018
0.020
0.022
0.024
0.026
0.028
0.030
0.032

FORTH
INSTITUTE OF COMPUTER SCIENCE

SIGNAL PROCESSING LAB

https://spl.ics.forth.gr

TITAN ARTIFICIAL INTELLIGENCE IN ASTROPHYSICS

FORTH INSTITUTE OF COMPUTER SCIENCE

Signal Processing Laboratory

CosmoStat

cea

Funded by the European Union