

GEWEX ML4LM Webinar | 16 November 2026

Physics-informed Machine Learning for Land Data Assimilation

Xin Li

National Tibetan Plateau Data Center

State Key Laboratory of Tibetan Plateau Earth System, Environment and Resources

Institute of Tibetan Plateau Research

Chinese Academy of Sciences, Beijing, China







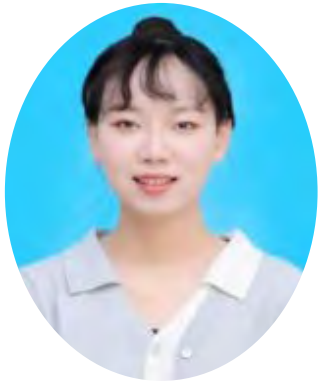














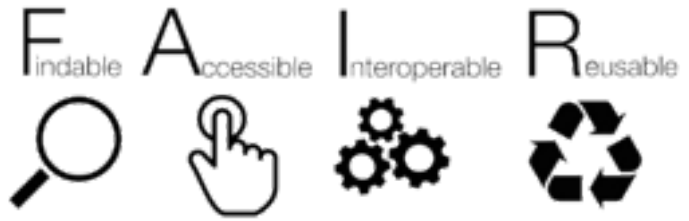


TPDC: National Tibetan Plateau Science Data Center

- A **national data center** dedicated to promoting scientific data for the **Tibetan Plateau** and its surrounding regions, established in **2019**
- Users **> 155,000** Datasets **> 8,800** Size: **>880 TB** Views: **>33 million**



Li et al., 2021, Nature Geoscience; Pan et al., 2021, BAMS





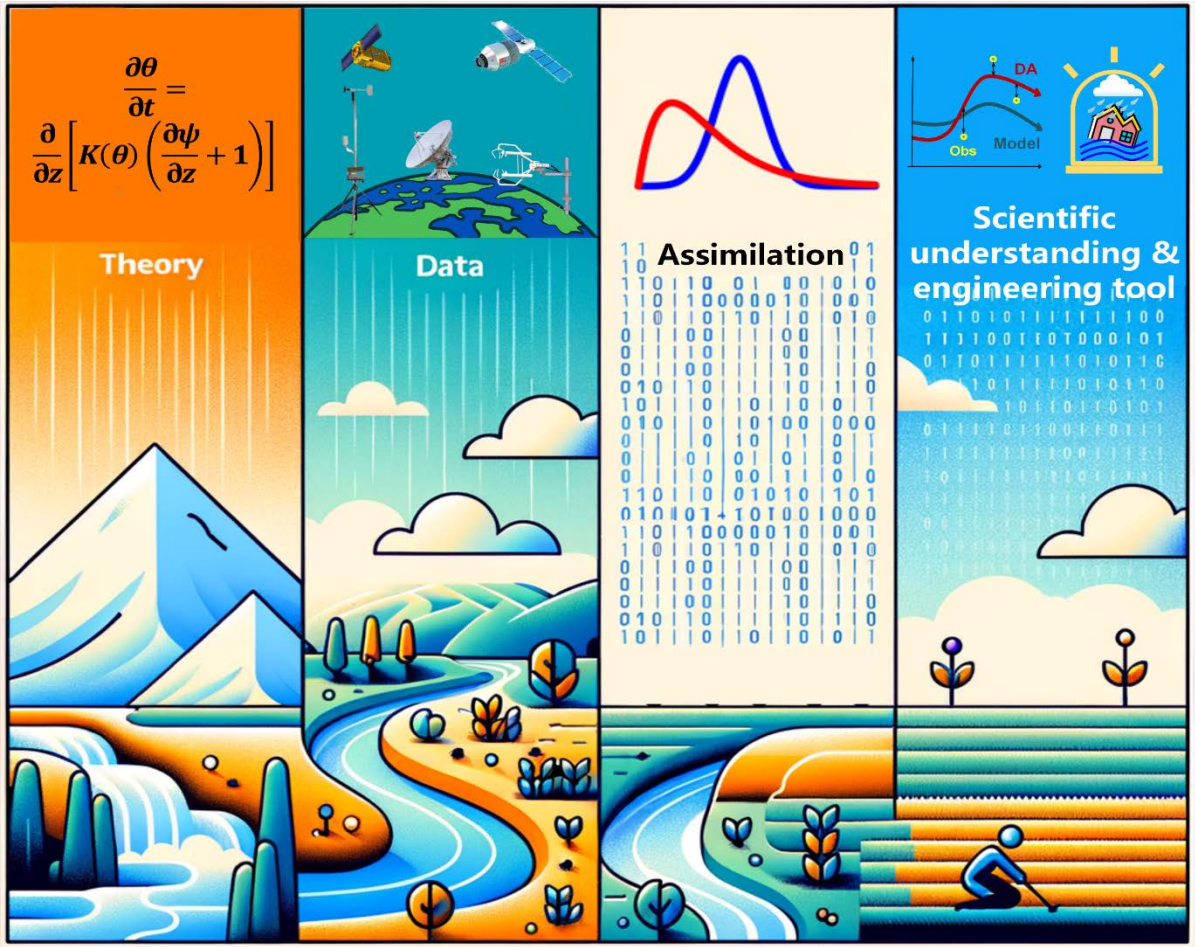
Contents

1. Land data assimilation: Fundamentals and evolution
2. Machine learning: Advances and Earth system applications
3. Data assimilation and machine learning: Bayesian framework
4. Integration of machine learning and land data assimilation: Recent advances
5. Key challenges and future research directions

1. Land data assimilation:

Fundamentals and evolution

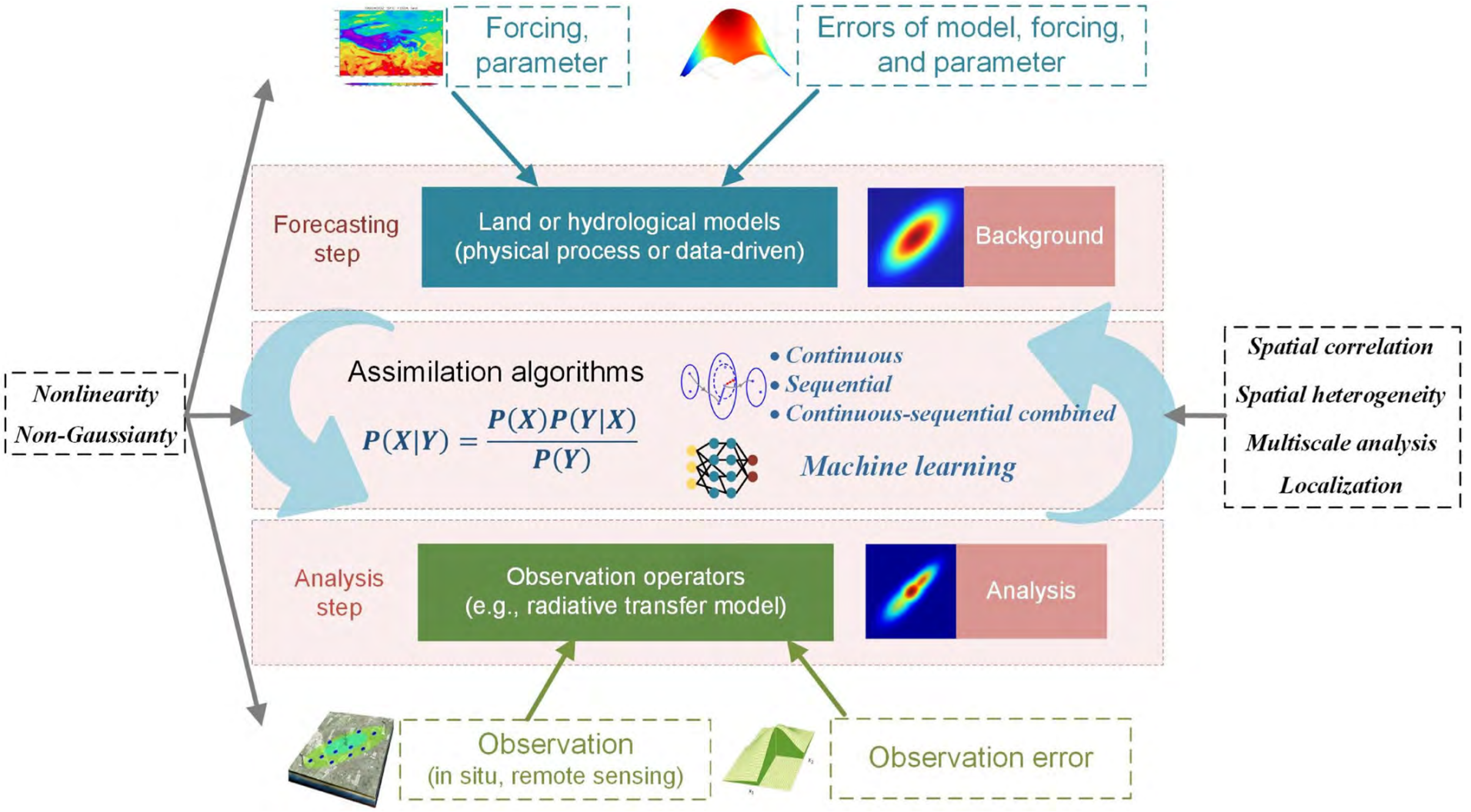
Data assimilation: Harmonizing theory and data



DA advances scientific understanding and serves as an engineering tool for Earth system sciences, reflecting the trend of **harmonizing theory and data** in the era of big data and artificial intelligence.

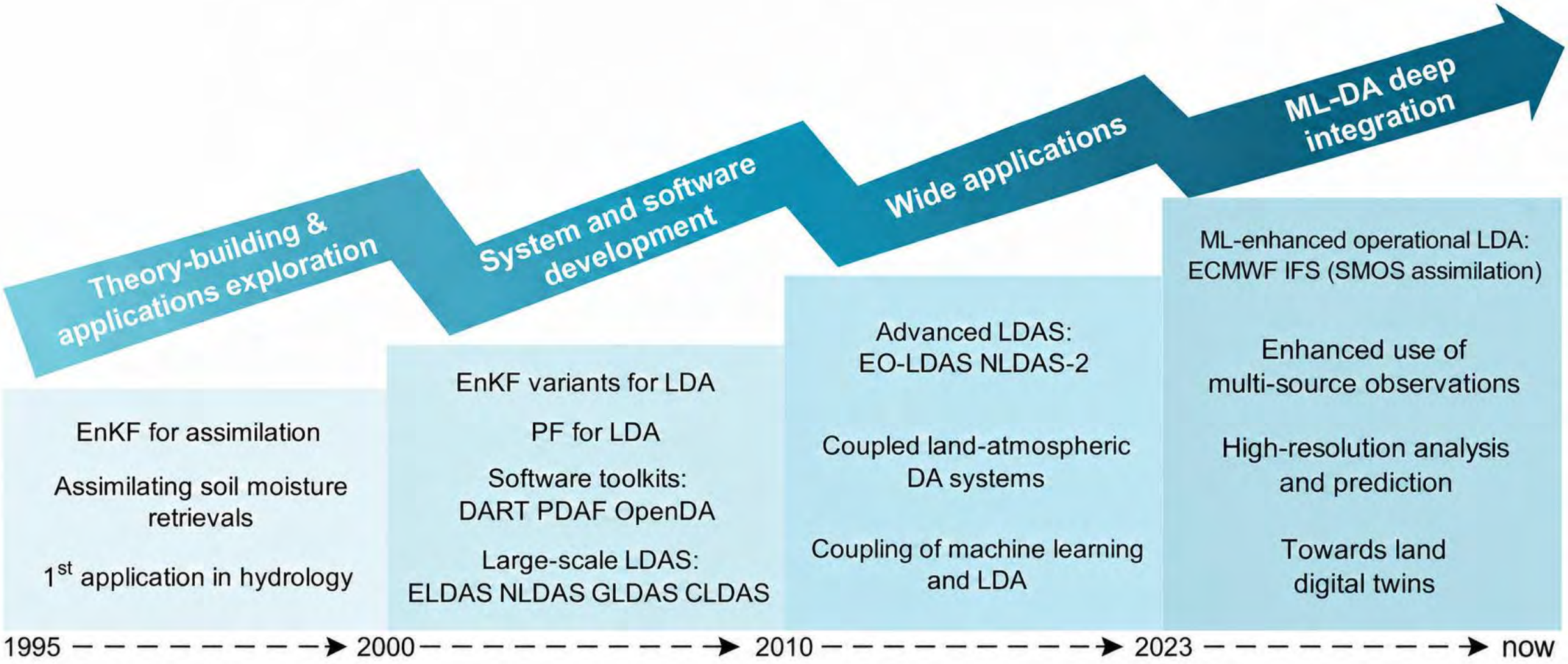
Li, X., Liu, F., Ma, CF, Hou, JL, Zheng, DH, Ma, HQ, Bai YL, Han XJ, Vereecken H, Yang K, Duan QY, Huang CL. Land data assimilation: Harmonizing theory and data in land surface process studies. *Reviews of Geophysics*, 2024, 62, e2022RG000801. https://doi.org/10.1029/2022RG000801

What is land data assimilation?



Li et al., 2024, https://doi.org/10.1029/2022RG000801

Evolution of land data assimilation



Modified from Li et al., 2024, https://doi.org/10.1029/2022RG000801

What makes land data assimilation different?

1. High nonlinearity and multiscale dependence

States, fluxes and parameters are strongly nonlinear, scale-dependent and spatially correlated, unlike the smoother chaotic dynamics of the atmosphere.

2. Non-Gaussian errors

Both model and observation errors are non-Gaussian; land states are bounded and truncated, for example soil moisture limits, with many thresholds.

3. Forcing outweighs initial conditions

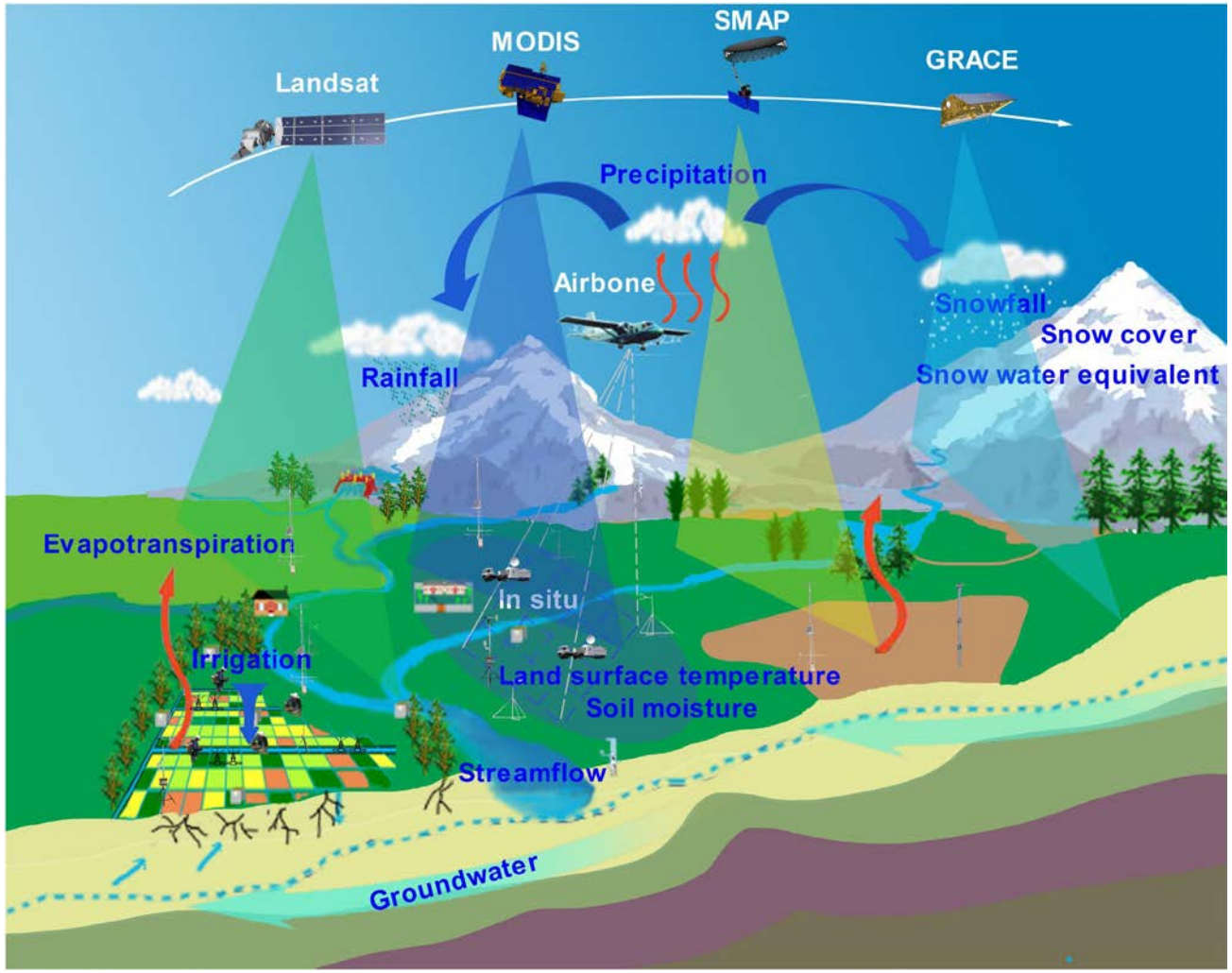
External forcing such as precipitation and radiation drives land predictability more than the initial state, the reverse of NWP, where initial conditions dominate.

4. High parameter sensitivity

Land predictions are highly sensitive to uncertain soil and vegetation parameters, so parameter estimation is an essential part of assimilation.

Li et al., 2024, https://doi.org/10.1029/2022RG000801

What variables does land data assimilation estimate?



Multisource, multiplatform observations (satellite, airborne, ground) are assimilated to estimate:

- Soil moisture
- Snow water equivalent / snow depth
- Groundwater
- Evapotranspiration
- Land surface temperature
- Other variables

Li et al., 2024, https://doi.org/10.1029/2022RG000801

Major land data assimilation systems (LDAS)

These LDASs are already operational in global and regional systems producing long-term datasets.

System	Extent	Resolution	DA method
GLDAS	Global	0.25 degree, 3-hourly	4DVar, EnKF
NLDAS	North America	0.125 degree, hourly	4DVar, EnKF
ELDAS	Europe	0.2 degree, 3-hourly	4DVar, OI
CaLDAS	Canada	6-hourly	OI, EnKF
CLDAS	China	0.1 degree, 3-hourly	EnKF

Li et al., 2024, https://doi.org/10.1029/2022RG000801

LDA workshops and training courses in China

Four domestic land data assimilation training courses



1st course, 2007, Lanzhou



2nd summer school, 2010, Beijing



3rd course, 2013, Changchun



4th course, 2023, Tianjin

LDA workshops and training courses in China

CAS land data assimilation symposia



1st CAS symposium, 2011, Xi'an



2nd CAS symposium, 2013, Beijing



CAS symposium, 2016, Lanzhou

International workshops on catchment hydrological modeling and data assimilation (CAHMDA)



CAHMDA-IV, 2010, Lhasa



CAHMDA-VII, 2017, Xi'an

2. Machine learning:

Advances and Earth system applications

Data assimilation in the big data era

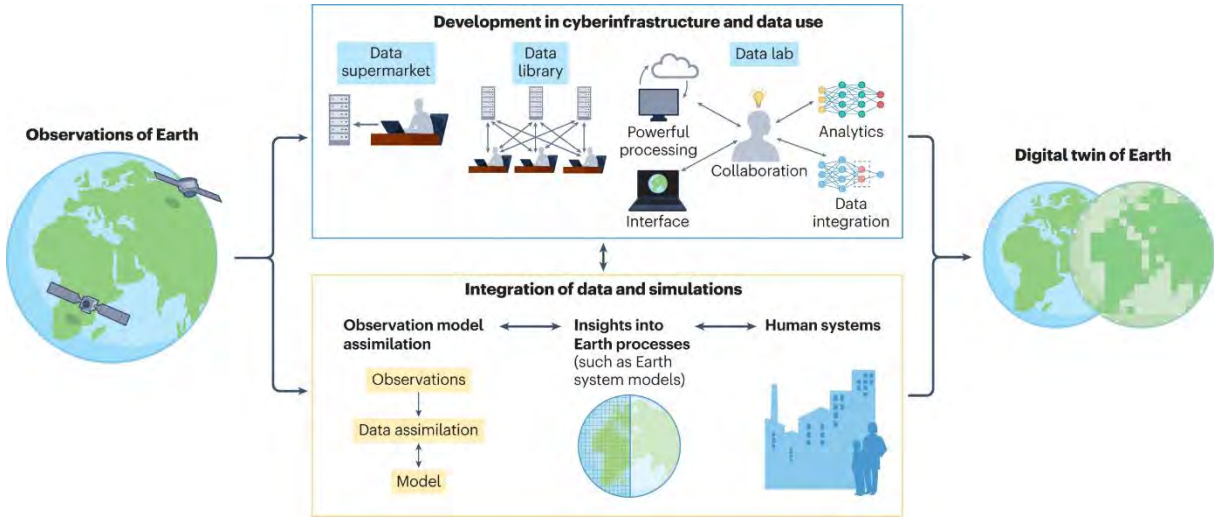
nature reviews earth & environment
https://doi.org/10.1038/s43017-023-00409-w

Review article

Check for updates

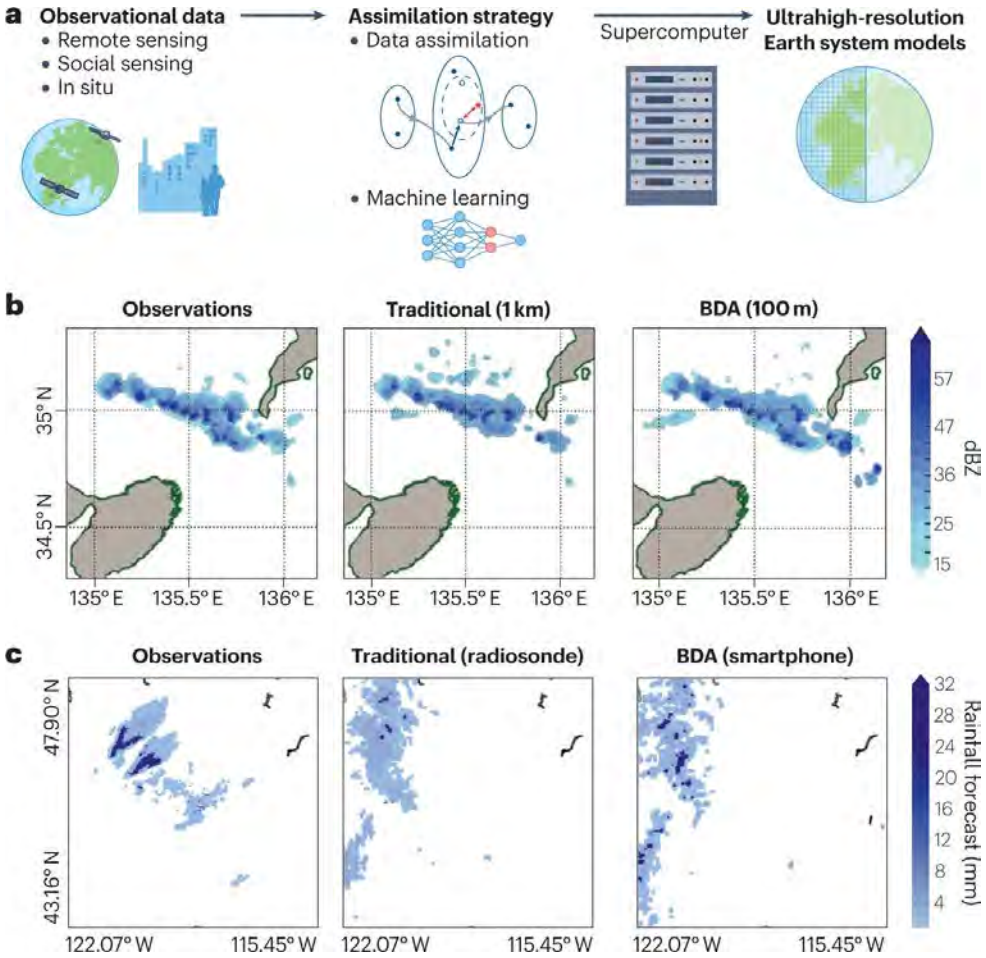
Big Data in Earth system science and progress towards a digital twin

Xin Li[1,2,7], Min Feng[1,2,7], Youhua Ran[2,3], Yang Su[2,3], Feng Liu[3], Chunlin Huang[3], Huanfeng Shen[4], Qing Xiao[2,5], Jianbin Su[1], Shiwei Yuan[1] & Huadong Guo[2,5,6]



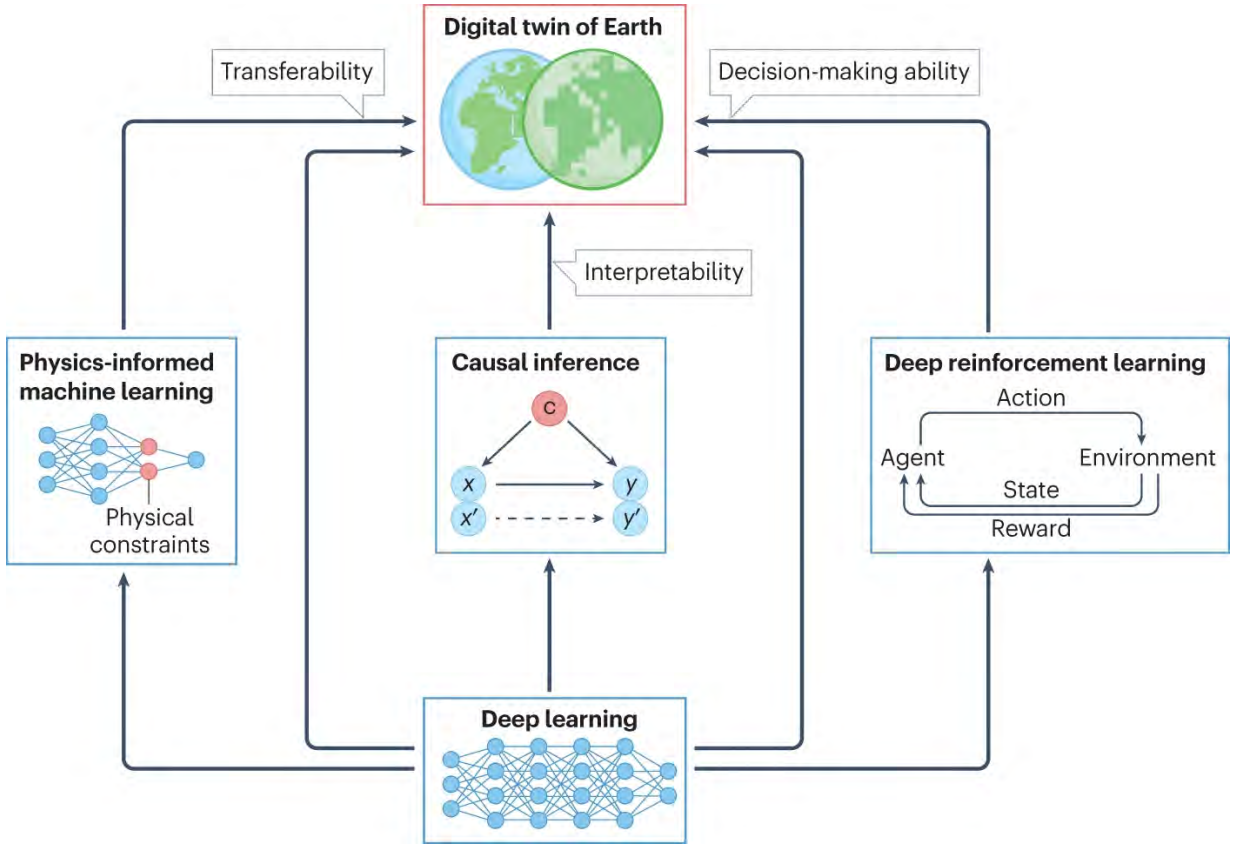
Transition of data use in Earth system science

Big Data assimilation



Li et al., 2023, Nature Reviews Earth & Environment. https://doi.org/10.1038/s43017-023-00409-w

AI-enabled advances in Earth system science



Representation learning for high-dimensional data

Extract complex spatiotemporal patterns and features through hierarchical nonlinear transformations.

Physics-informed machine learning

Integrate physical knowledge into data, model architectures, or loss functions to improve physical consistency and generalizability.

Causal inference for process understanding

Identify causal relationships and intervention effects to support process interpretation and scientific understanding.

Reinforcement learning for decision-making

Learn optimal action policies through environmental interactions and feedback to support decision-making in complex systems.

Li et al., 2023, https://doi.org/10.1038/s43017-023-00409-w

AI enhances forecasting skill and computational efficiency

nature

Explore content ⌄ About the journal ⌄ Publish with us ⌄

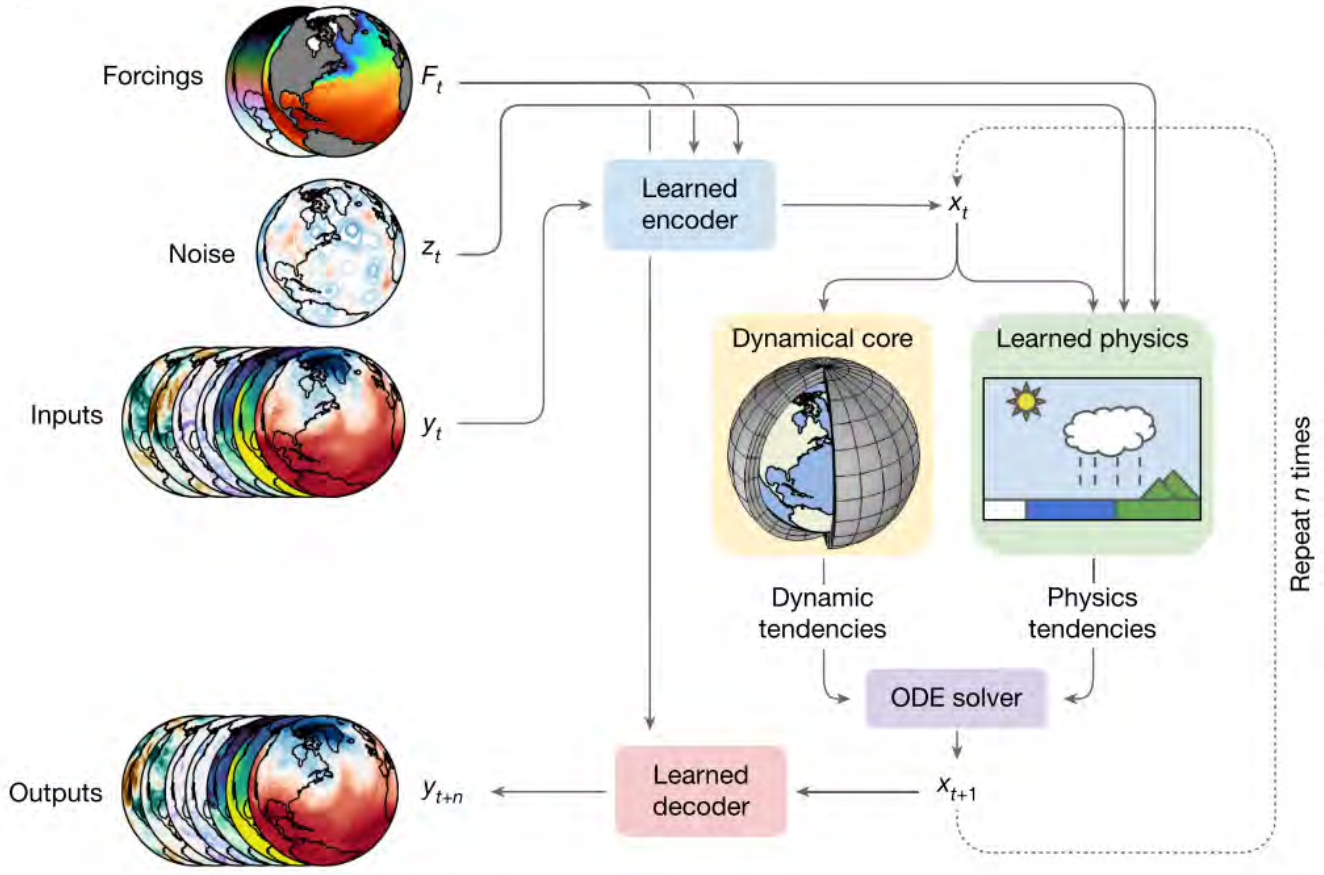
nature › articles › article

Article | Open access | Published: 22 July 2024

Neural general circulation models for weather and climate

Dmitrii Kochkov, Janni Yuval, Ian Langmore, Peter Norgaard, Jamie Smith, Griffin Mooers, Milan Klöwer, James Lottes, Stephan Rasp, Peter Düben, Sam Hatfield, Peter Battaglia, Alvaro Sanchez-Gonzalez, Matthew Willson, Michael P. Brenner & Stephan Hoyer

- NeuralGCM achieves forecast skill comparable to that of existing AI models for 1–10-day predictions.
- NeuralGCM achieves forecast skill comparable to the ECMWF Ensemble Prediction System for 1–15-day forecasts.
- NeuralGCM reduces computational costs by 3–5 orders of magnitude compared with traditional GCMs.



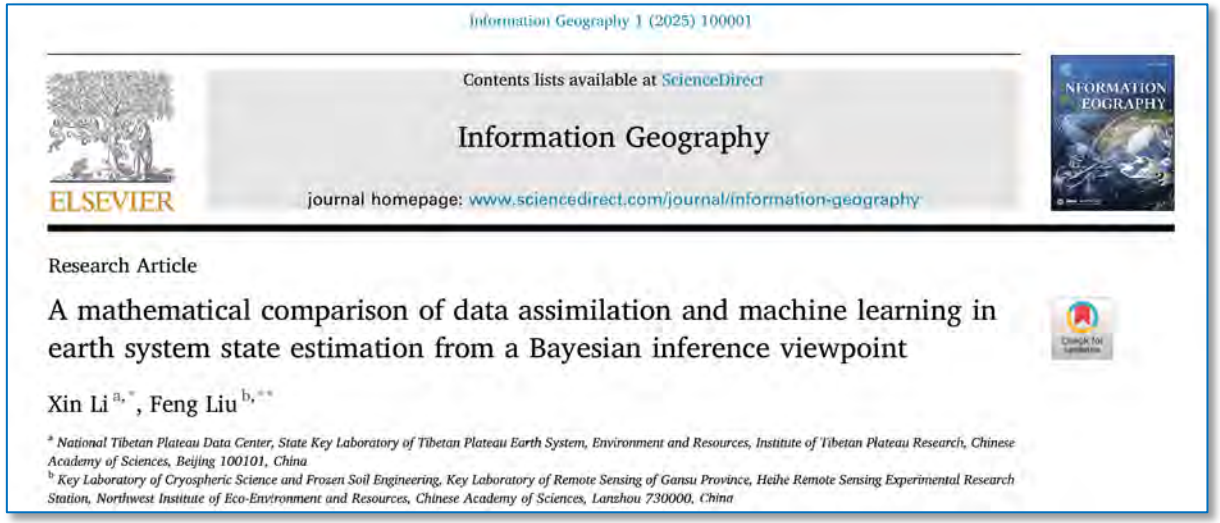
Neural general circulation model (NeuralGCM) architecture

Kochkov et al., 2024, Nature. https://doi.org/10.1038/s41586-024-07744-y

3. Data assimilation and machine learning:

Bayesian framework

A common state-estimation problem



Information Geography 1 (2025) 100001

Contents lists available at ScienceDirect

Information Geography

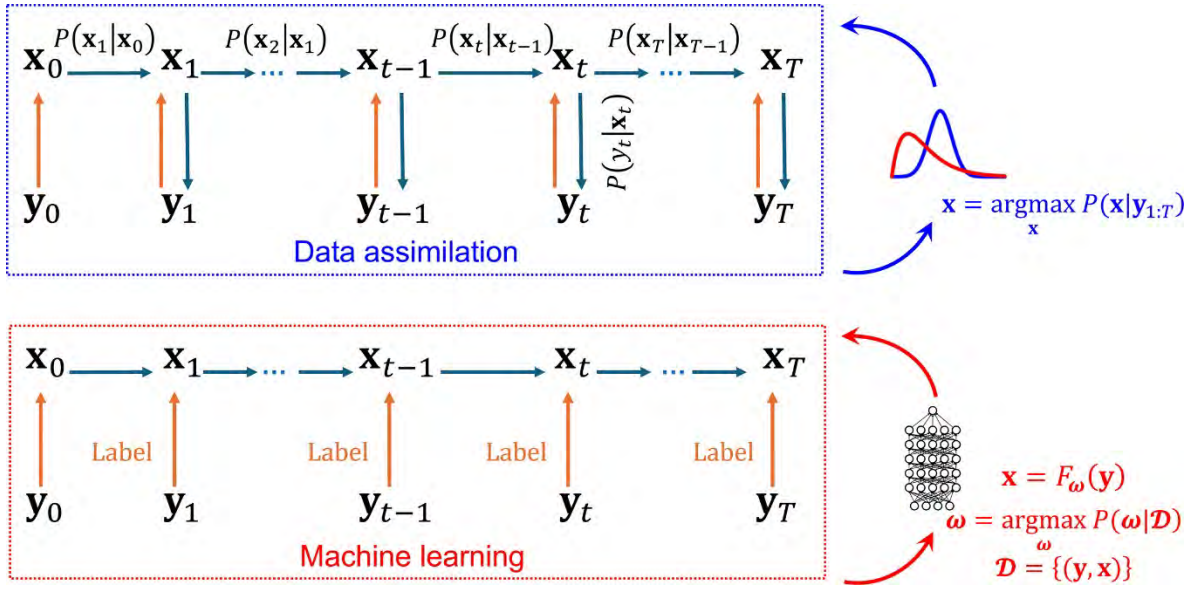
journal homepage: www.sciencedirect.com/journal/information-geography

Research Article

A mathematical comparison of data assimilation and machine learning in earth system state estimation from a Bayesian inference viewpoint

Xin Li [a,*], Feng Liu [b,**]

[a] National Tibetan Plateau Data Center, State Key Laboratory of Tibetan Plateau Earth System, Environment and Resources, Institute of Tibetan Plateau Research, Chinese Academy of Sciences, Beijing 100101, China
[b] Key Laboratory of Cryospheric Science and Frozen Soil Engineering, Key Laboratory of Remote Sensing of Gansu Province, Heihe Remote Sensing Experimental Research Station, Northwest Institute of Eco-Environment and Resources, Chinese Academy of Sciences, Lanzhou 730000, China



- DA: Observations constrain states evolving under a dynamical model.
- ML: Input–target pairs train a parametric model.

Li & Liu, 2025, Information Geography. https://doi.org/10.1016/j.infgeo.2025.100001

Formulation of state estimation in DA

$$\arg\max_{\mathbf{x}} P(\mathbf{x}|\mathbf{y}_{1:T}) \quad \Rightarrow \quad \operatorname{argmax}_{\mathbf{x}} P(\mathbf{x}_0) \prod_{t=1}^{T} P(\mathbf{y}_t|\mathbf{x}_t) P(\mathbf{x}_t|\mathbf{x}_{t-1})$$

Maximum a posteriori — *Bayesian inferences*

⇓ *Logarithm transformation*

$$\begin{aligned}
&\operatorname{argmax}_{\mathbf{x}} ln\, P(\mathbf{x}|\mathbf{y}_{1:T}) \\
&= \operatorname{argmax}_{\mathbf{x}} \left[ln\, P(\mathbf{x}_0) + \sum_{t=1}^{T} ln\, P(\mathbf{y}_t|\mathbf{x}_t) + \sum_{t=1}^{T} ln\, P(\mathbf{x}_t|\mathbf{x}_{t-1}) \right] \\
&= \operatorname{argmin}_{\mathbf{x}} - \left[ln\, P(\mathbf{x}_0) + \sum_{t=1}^{T} ln\, P(\mathbf{y}_t|\mathbf{x}_t) + \sum_{t=1}^{T} ln\, P(\mathbf{x}_t|\mathbf{x}_{t-1}) \right]
\end{aligned}$$

Cost function of data assimilation

$$\underset{\mathbf{x}}{\operatorname{argmin}} - \left[ln\, P\,(\mathbf{x}_0) + \sum_{t=1}^{T} ln\, P(\mathbf{y}_t|\mathbf{x}_t) + \sum_{t=1}^{T} ln\, P(\mathbf{x}_t|\mathbf{x}_{t-1}) \right]$$

Non-Gaussian MCMC

$$\mathbf{x}_0 \sim \mathcal{N}(\mathbf{x}_b, \mathbf{B})$$

$$\mathbf{y}_t|\mathbf{x}_t \sim \mathcal{N}(H(\mathbf{x}_t), \mathbf{R}_t)$$

$$\mathbf{x}_t|\mathbf{x}_{t-1} \sim \mathcal{N}(M(\mathbf{x}_{t-1}), \mathbf{Q}_t)$$

Gaussian

$$\underset{\mathbf{x}}{\operatorname{argmin}} \frac{1}{2} \left[(\mathbf{x}_0 - \mathbf{x}_b)^T \mathbf{B}^{-1} (\mathbf{x}_0 - \mathbf{x}_b) + \sum_{t=1}^{T} (\mathbf{y}_t - H_t(\mathbf{x}_t))^T \mathbf{R}_t^{-1} (\mathbf{y}_t - H_t(\mathbf{x}_t)) + \sum_{t=1}^{T} (\mathbf{x}_t - M(\mathbf{x}_{t-1}))^T \mathbf{Q}_t^{-1} (\mathbf{x}_t - M(\mathbf{x}_{t-1})) \right]$$

Cost function

Model operator: dynamical model

Observation Observation operator

Theory

Data Theory

$$J(\mathbf{s}) = \sum_{t=1}^{T} \|\mathbf{s}_t - M(\boldsymbol{s}_{t-1})\|^2_{\mathbf{Q}_t^{-1}} + \sum_{k=1}^{K} \|\mathbf{y}_t - H(\boldsymbol{s}_t)\|^2_{\mathbf{R}_t^{-1}} + \|\boldsymbol{s}_0 - \boldsymbol{s}_b\|^2_{\mathbf{B}^{-1}}$$

Model error

Observation error

Xin Li, Feng Liu. 2025, https://doi.org/10.1016/j.infgeo.2025.100001

Formulation of the state estimation and prediction problem in ML

Training data: $\mathcal{D} = \{(\mathbf{y}_0, \mathbf{x}_0), \ \ldots, (\mathbf{y}_t, \mathbf{x}_t)\}$.

- $\mathbf{y}$ represents the input data, $\mathbf{x}$ represents the targeted/labeled output data in supervised learning.
- $\mathcal{D}$ can be a subset of $\{(\mathbf{y}_0, \mathbf{x}_0), \ \ldots, (\mathbf{y}_t, \mathbf{x}_t)\}$, and temporally asynchronous

ML model parameter set: $\boldsymbol{\omega} \in \mathbb{R}^{1\times L}$ is an L-dimensional parameter set.

- L could be very large in many cases such as deep learning.
- $\boldsymbol{\omega}$ is the weights of a neural network, and is assumed to be time invariant.

Parametric model: $F_{\boldsymbol{\omega}}: \mathbf{y} \rightarrow \mathbf{x}$ that best fits training data using $\boldsymbol{\omega}$.

- $F_{\boldsymbol{\omega}}$ is usually an ML model.

Cost function of machine learning

$$P(\boldsymbol{\omega}|\boldsymbol{\mathcal{D}}) \propto P(\boldsymbol{\mathcal{D}}|\boldsymbol{\omega})P(\boldsymbol{\omega}) = P(\mathbf{x}|\mathbf{y},\boldsymbol{\omega})P(\boldsymbol{\omega})$$

$$\underset{\boldsymbol{\omega}}{\operatorname{argmax}}\, P(\boldsymbol{\omega}|\boldsymbol{\mathcal{D}}) = \underset{\boldsymbol{\omega}}{\operatorname{argmax}} P(\mathbf{x}|\boldsymbol{y},\boldsymbol{\omega})P(\boldsymbol{\omega}) = \underset{\boldsymbol{\omega}}{\operatorname{argmax}}\, \prod_{t=0}^{T} P(\mathbf{x}_t|\mathbf{y}_t,\boldsymbol{\omega})\, P(\boldsymbol{\omega})$$



Gaussian

$$\underset{\boldsymbol{\omega}}{\operatorname{argmin}}\, \frac{1}{2}\left[\sum_{t=0}^{T}\left(\mathbf{x}_t - F_{\boldsymbol{\omega}}(\mathbf{y}_t)\right)^T\left(\mathbf{x}_t - F_{\boldsymbol{\omega}}(\mathbf{y}_t)\right) + \sum_{t=0}^{T}(\boldsymbol{\omega}-\tilde{\boldsymbol{\omega}})^T{\boldsymbol{\sigma}_{\omega}}^{-1}(\boldsymbol{\omega}-\tilde{\boldsymbol{\omega}})\right]$$

$$L(\boldsymbol{\omega}) = \sum_{t=0}^{T} \| \mathbf{x}_t - F_{\boldsymbol{\omega}}(\mathbf{y}_t) \|^2 + \sum_{t=0}^{T} \| \boldsymbol{\omega} - \tilde{\boldsymbol{\omega}} \|^2_{{\boldsymbol{\sigma}_{\omega}}^{-1}}$$

Comparison of cost functions

$$J(\mathbf{x}) = \sum_{t=1}^{T} \|\mathbf{x}_t - M(\mathbf{x}_{t-1})\|^2_{\mathbf{Q}_t^{-1}} + \sum_{t=1}^{T} \|\mathbf{y}_t - H(\mathbf{x}_t)\|^2_{\mathbf{R}_t^{-1}} + \|\mathbf{x}_0 - \mathbf{x}_b\|^2_{\mathbf{B}^{-1}}$$

$$L(\boldsymbol{\omega}) = \sum_{t=0}^{T} \|\mathbf{x}_t - F_{\boldsymbol{\omega}}(\mathbf{y}_t)\|^2 + \sum_{t=0}^{T} \|\boldsymbol{\omega} - \tilde{\boldsymbol{\omega}}\|^2_{\boldsymbol{\sigma}_{\boldsymbol{\omega}}^{-1}}$$

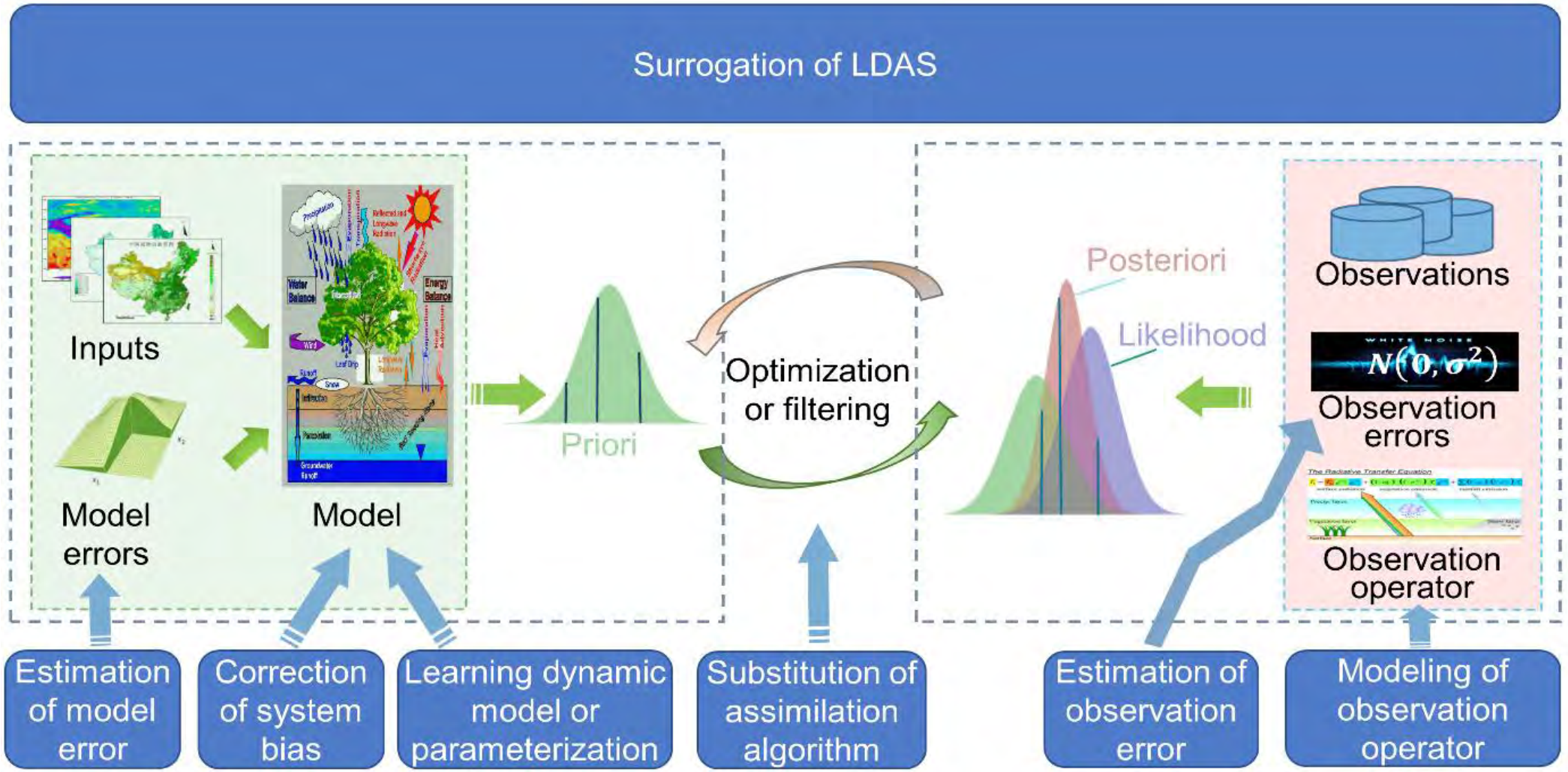
Comparison of DA and ML

	Data assimilation	Machine learning
Motivation and application scope	To improve state estimation and prediction accuracy by fusing observational data into physics-based models.	To improve prediction by estimating the model parameter from large datasets, with a lesser emphasis on dynamical systems.
Governing equations	Known system dynamics and observation model	No explicit system dynamics and observation model
Parameters	Usually with physical meanings and interpretable	Related to model structure and often abstracted from data, with varied interpretability
Cost functions	• Integrates dynamics and observations to align state estimates with physical laws and data. • Transforms states into the observation space.	• Optimizes prediction accuracy through data-driven approaches. • Maps observational data back to system states.
Optimization algorithms	Uses Bayesian filtering and variational methods to integrate data into models, often requiring explicit computations of adjoint model and Hessian matrix.	Employs algorithms like backward propagation and stochastic gradient descent in optimization.
Uncertainty quantification	Explicitly formulated and quantified in the DA process.	Embedded within the model structure, and implicitly handled.
Data utilization	Requires sparse observational data with high fidelity.	Relies on the availability of large-scale dense data.

Xin Li, Feng Liu. 2025, https://doi.org/10.1016/j.infgeo.2025.100001

4. Integration of machine learning and land data assimilation: Recent advances

Symbiotic integration of ML and LDA



Li et al., 2024, Reviews of Geophysics. https://doi.org/10.1029/2022RG000801

ML in LDA: Methods and applications


Hydrol. Earth Syst. Sci., 27, 1583–1606, 2023
https://doi.org/10.5194/hess-27-1583-2023
© Author(s) 2023. This work is distributed under the Creative Commons Attribution 4.0 License.

Hydrology and Earth System Sciences

Improving regional climate simulations based on a hybrid data assimilation and machine learning method

Xinlei He[1], Yanping Li[2], Shaomin Liu[1], Tongren Xu[1], Fei Chen[3], Zhenhua Li[2], Zhe Zhang[2], Rui Liu[4], Lisheng Song[5], Ziwei Xu[1], Zhixing Peng[1], and Chen Zheng[6]

Hybrid DA-ML for regional climate simulations


Contents lists available at ScienceDirect

Agricultural and Forest Meteorology

journal homepage:

ELSEVIER

Coupling data assimilation and machine learning to improve land surface conditions and near-surface temperature and humidity forecasts

Xinlei He[a,b], Shaomin Liu[b,*], Tongren Xu[b], Fei Chen[c], Zhitao Wu[b], Ziwei Xu[b], Xiang Li[d], Rui Liu[e]

Land states and near-surface forecasts


Contents lists available at ScienceDirect

Agricultural and Forest Meteorology

journal homepage:

ELSEVIER

Innovative approach for estimating evapotranspiration and gross primary productivity by integrating land data assimilation, machine learning, and multi-source observations☆

Xinlei He[a,b], Shaomin Liu[a,*], Sayed M. Bateni[c], Tongren Xu[a,*], Changhyun Jun[d], Dongkyun Kim[e], Xin Li[f], Lisheng Song[g], Long Zhao[b], Ziwei Xu[a], Jiaxing Wei[a]

Noah-MP soil moisture bias correction


https://doi.org/10.5194/egusphere-2026-5358
Preprint. Discussion started: 10 September 2026
© Author(s) 2026. CC BY 4.0 License.

EGUsphere

Score-based Filtering for Land Data Assimilation

Zhilong Fan[1,2], Lu Li[1], Zhongwang Wei[1], Shupeng Zhang[1], Sisi Chen[1], Xingjie Lu[1], Wei Shangguan[1], Nan Wei[1], Hua Yuan[1], Pan Zhang[1,2], Jiaxin Tian[3,4], Jianhong Zhou[5], Baoxiang Pan[6], Wei Han[4], and Yongjiu Dai[1]

Diffusion filtering with CoLM


Water Resources Research

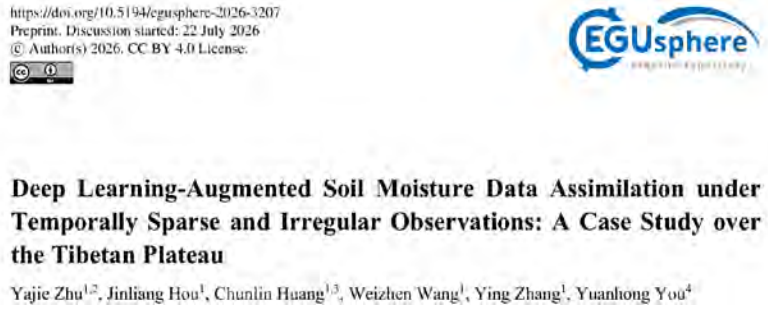
Research Article | Free Access

Improving Snow Estimates Through Assimilation of MODIS Fractional Snow Cover Data Using Machine Learning Algorithms and the Common Land Model

Jinliang Hou, Chunlin Huang, Weijing Chen, Ying Zhang

First published: 23 June 2021 | https://doi.org/10.1029/2020WR029010

Snow cover assimilation


https://doi.org/10.5194/egusphere-2026-3207
Preprint. Discussion started: 22 July 2026
© Author(s) 2026. CC BY 4.0 License.

EGUsphere

Deep Learning-Augmented Soil Moisture Data Assimilation under Temporally Sparse and Irregular Observations: A Case Study over the Tibetan Plateau

Yajie Zhu[1,2], Jinliang Hou[1], Chunlin Huang[1,3], Weizhen Wang[1], Ying Zhang[1], Yuanhong You[4]

Soil moisture DA with sparse data


Water Resources Research

Research Article | Open Access

Harnessing Generative Deep Learning for Enhanced Ensemble Data Assimilation

Ehsan Foroumandi, Hamid Moradkhani

First published: 07 July 2025 | https://doi.org/10.1029/2025WR040078 | VIEW METRICS

Hydrologic DA in latent space


Geophysical Research Letters

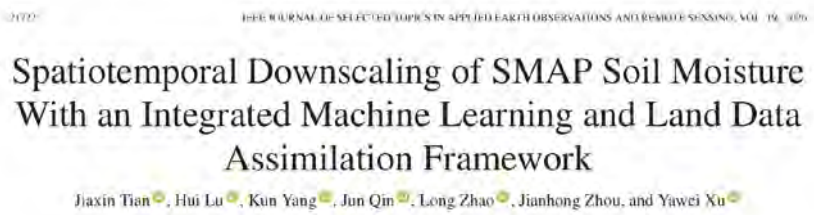
Research Letter | Open Access

The Role of Baseflow Data Assimilation in Hydrologic Modeling and Peak Flow Prediction

Parnian Ghaneei, Ehsan Foroumandi, Hamid Moradkhani

First published: 01 April 2026 | https://doi.org/10.1029/2025GL121456 | VIEW METRICS

Baseflow assimilation for peak flow prediction


IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, VOL. 19, 4006

Spatiotemporal Downscaling of SMAP Soil Moisture With an Integrated Machine Learning and Land Data Assimilation Framework

Jiaxin Tian, Hui Lu, Kun Yang, Jun Qin, Long Zhao, Jianhong Zhou, and Yawei Xu

SMAP soil moisture downscaling

01

ML as an assimilation algorithm

Diffusion-based assimilation

02

Learning dynamic models or parameterizations

Latent-space assimilation

03

Learning observation operators

Learned snow depletion curves

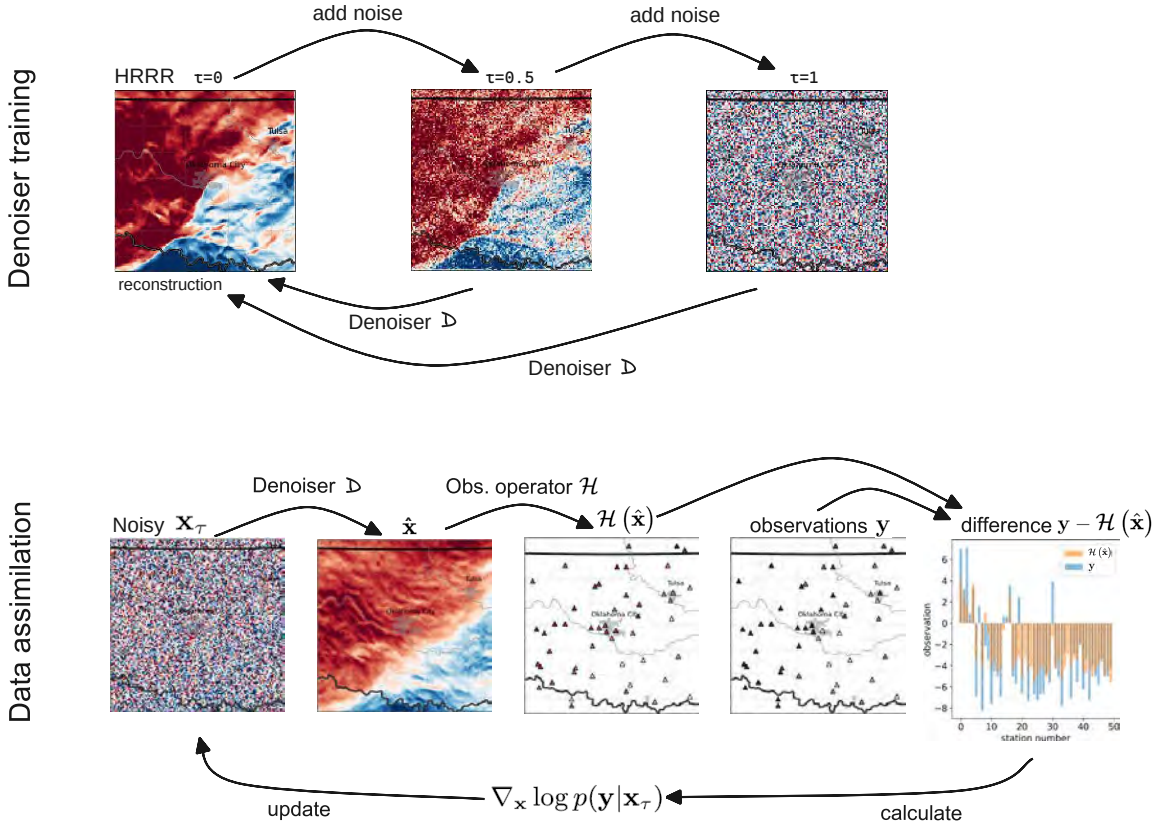
04

Bias correction

Model bias and observation errors

Diffusion models for data assimilation

Observation-guided score diffusion generates an analysis ensemble



1. Learn the forecast score

$$z_\tau = \sqrt{\bar{\alpha}_\tau}\, x^f + \sqrt{1-\bar{\alpha}_\tau}\,\epsilon$$

$$s_\theta(z_\tau, \tau) \approx \nabla_z \log p^f_\tau(z_\tau)$$

2. Condition the score on observations

$$\underbrace{s^a_\tau}_{\text{analysis}} = \underbrace{s^f_\tau}_{\text{forecast}} + \underbrace{\nabla_z \log p_\tau(y_k \mid z_\tau)}_{\text{observation guidance}}$$

3. Sample by reversing the diffusion

$$\mathrm{d}z_\tau = \left[-\tfrac{1}{2}\beta(\tau) z_\tau - \beta(\tau) s^a_\tau\right]\mathrm{d}\tau + \sqrt{\beta(\tau)}\,\mathrm{d}\bar{W}_\tau, \qquad \tau : T \to 0$$

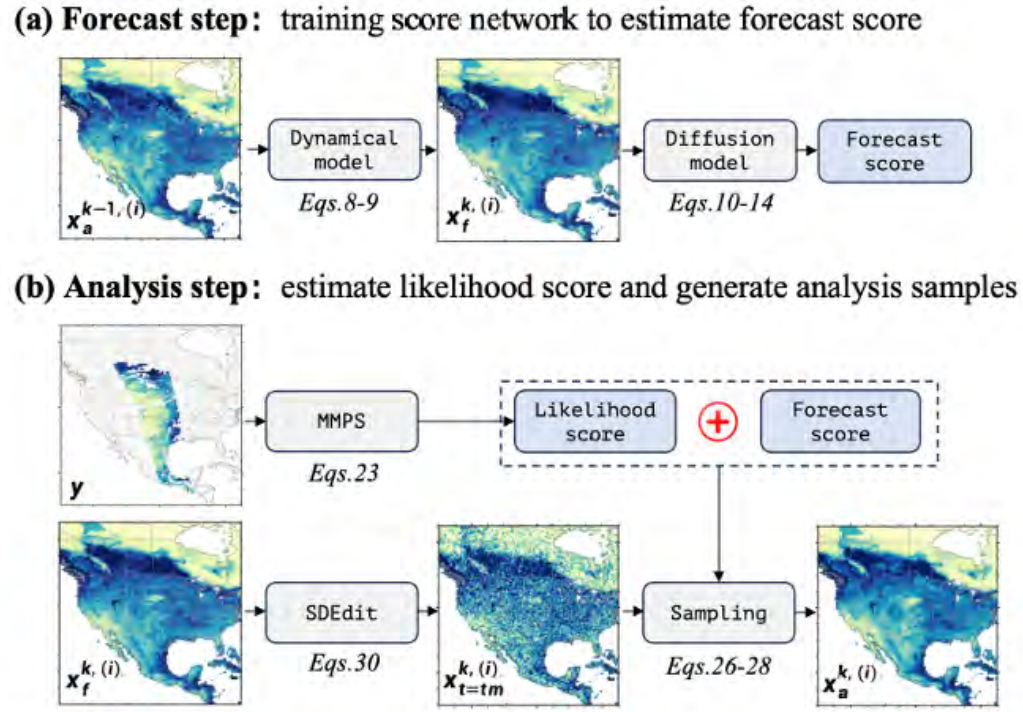
Manshausen et al., 2025, Journal of Advances in Modeling Earth Systems. https://doi.org/10.1029/2024MS004505

LandSF: diffusion filtering with CoLM

Refining diffusion filtering for **land**: SDEdit preserves forecast information and MMPS quantifies uncertainty

Score-based Filtering for Land Data Assimilation

Zhilong Fan[1,2], Lu Li[1], Zhongwang Wei[1], Shupeng Zhang[1], Sisi Chen[1], Xingjie Lu[1], Wei Shangguan[1], Nan Wei[1], Hua Yuan[1], Pan Zhang[1,2], Jiaxin Tian[3,4], Jianhong Zhou[5], Baoxiang Pan[6], Wei Han[4], and Yongjiu Dai[1]



CoLM, 100 members, real SMAP L2 soil moisture
Fine-tune a U-Net on the current CoLM forecast ensemble.

SDEdit: retain forecast information

B24 (Bao et al., 2024) is a score-based DA method that starts denoising from pure Gaussian noise.
LandSF starts from partially noised forecasts.

$$z_{\tau_m}^{(i)} = \sqrt{\bar{\alpha}_{\tau_m}}\, x_f^{(i)} + \sqrt{1-\bar{\alpha}_{\tau_m}}\, \epsilon^{(i)}$$

SDEdit reduces sensitivity to forecast-score errors.

Moment-matching posterior sampling (MMPS): account for uncertainty

Moment-match clean states given a noisy state.

$$p_\tau(y_k \mid z_\tau) \approx \mathcal{N}(y_k; H(\mu_\tau), S_\tau)$$

$$S_\tau \approx \underbrace{R}_{\text{observation error}} + \underbrace{J_H C_\tau J_H^\top}_{\text{conditional state uncertainty}}$$

Avoids B24's empirical guidance parameter.

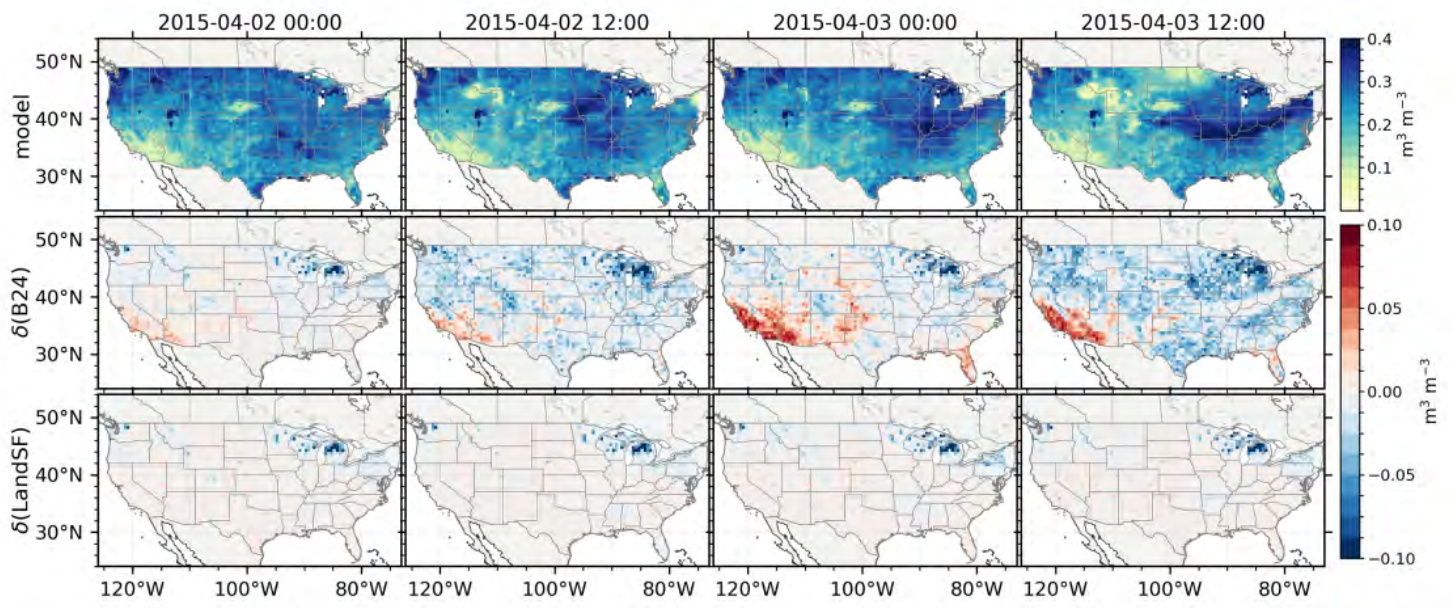
Fan et al., 2026, EGUsphere preprint. https://doi.org/10.5194/egusphere-2026-5358

LandSF: preserving forecast information and handling observation noise

Two experiments evaluate forecast preservation and robustness to noisy observations

SDEdit: preserve forecast information

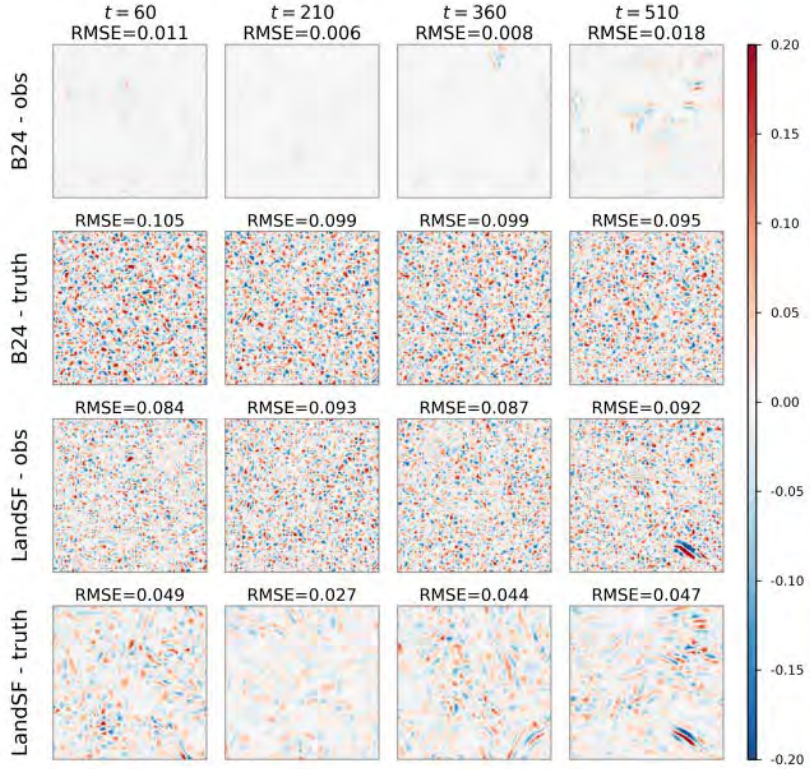
CoLM experiment without observation assimilation.



LandSF better preserves the CoLM forecast information

MMPS: account for observation uncertainty

Idealized Kolmogorov-flow experiment with known truth and noisy observations.



LandSF stays closer to the true state despite noisy observations

Generative Assimilation and Prediction (GAP)

npj | climate and atmospheric science

Explore content · About the journal · Publish with us

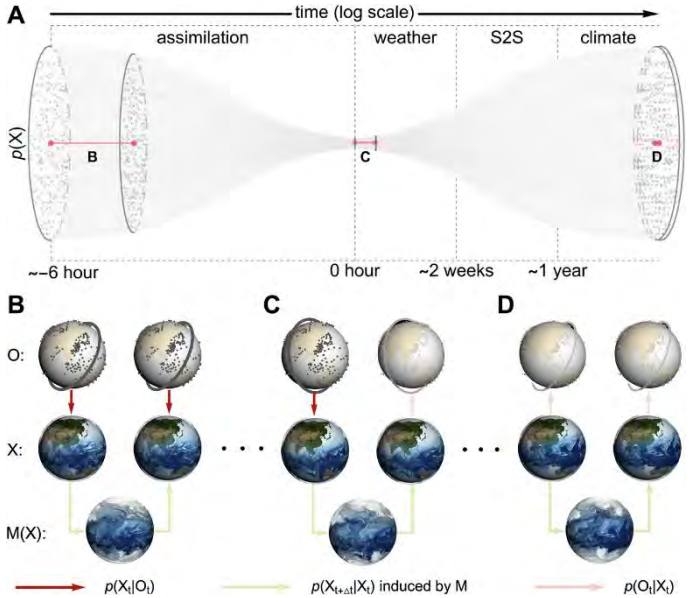
nature › npj climate and atmospheric science › articles › article

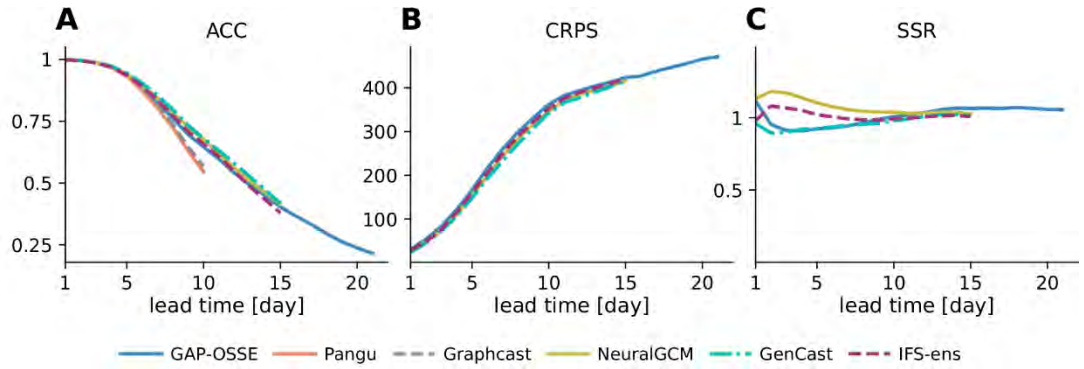
Article | Open access | Published: 02 November 2026

Generative assimilation and prediction for weather and climate

Shangshang Yang, Congyi Nai, Xinyan Liu, Weidong Li, Jie Chao, Jingnan Wang, Leyi Wang, Lu Li, Xichen Li, Xi Chen, Bo Lu, Xin Li, Ziniu Xiao, Niklas Boers ✉, Huiling Yuan ✉ & Baoxiang Pan ✉

$$d\mathbf{x} = [\mathbf{f}(\mathbf{x}, \tau) - g^2(\tau)\nabla_{\mathbf{x}} \log p(\mathbf{x})]d\tau + g(\tau)d\bar{\mathbf{w}}$$





- Learn a climatological prior for the atmosphere from ERA5.
- The same prior applies plug-and-play observational, predictive, and forcing constraints—no separate training across tasks.
- One seamless framework spans assimilation, weather, seasonal, and >1000-year climate runs with reliable uncertainty (ACC 3–14% higher).

Yang et al., 2026, npj Climate and Atmospheric Science. https://doi.org/10.1038/s41612-026-01585-4

01

ML as an assimilation algorithm

Diffusion-based assimilation

02

Learning dynamic models or parameterizations

Latent-space assimilation

03

Learning observation operators

Learned snow depletion curves

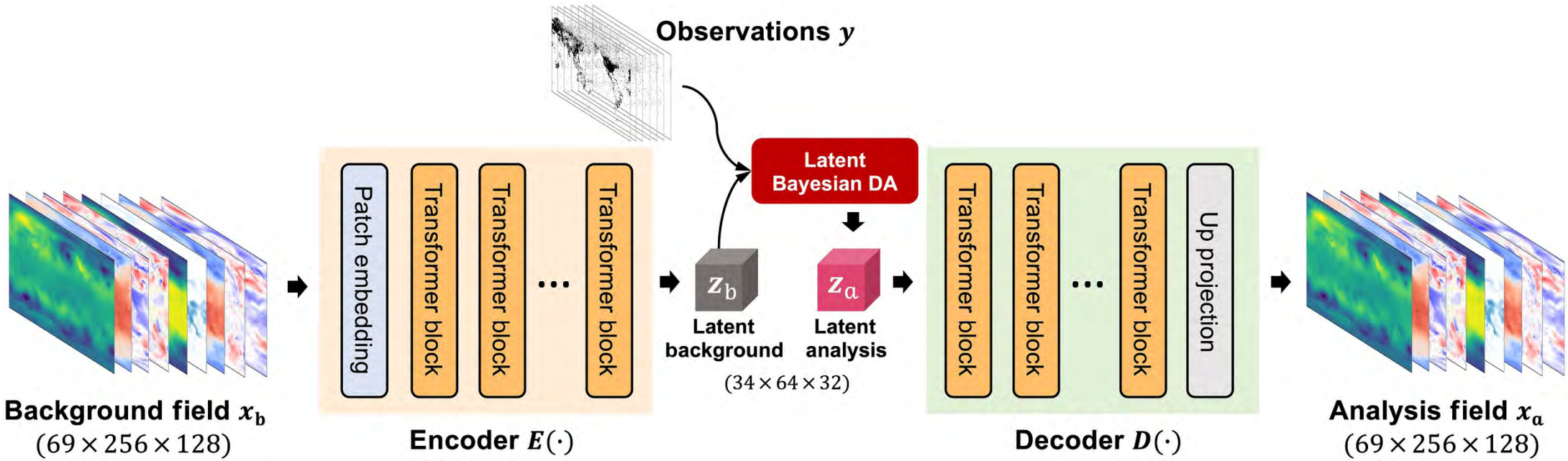
04

Bias correction

Model bias and observation errors

Latent data assimilation

Encode → assimilate in latent space → decode; ML learns the representation; DA performs the update



Why use a latent space?

- Reduces state dimensionality
- Captures relationships among coupled variables

Bayesian update in latent space

$$z^b = E(x^b), \qquad x^a = D(z^a)$$

$$p^a(z) \propto p(y \mid D(z))p^b(z)$$

Fan et al., 2026, Science Advances. https://doi.org/10.1126/sciadv.aea4248

HydroGEnDA: generative DA in latent space

Water Resources Research

Research Article | Open Access

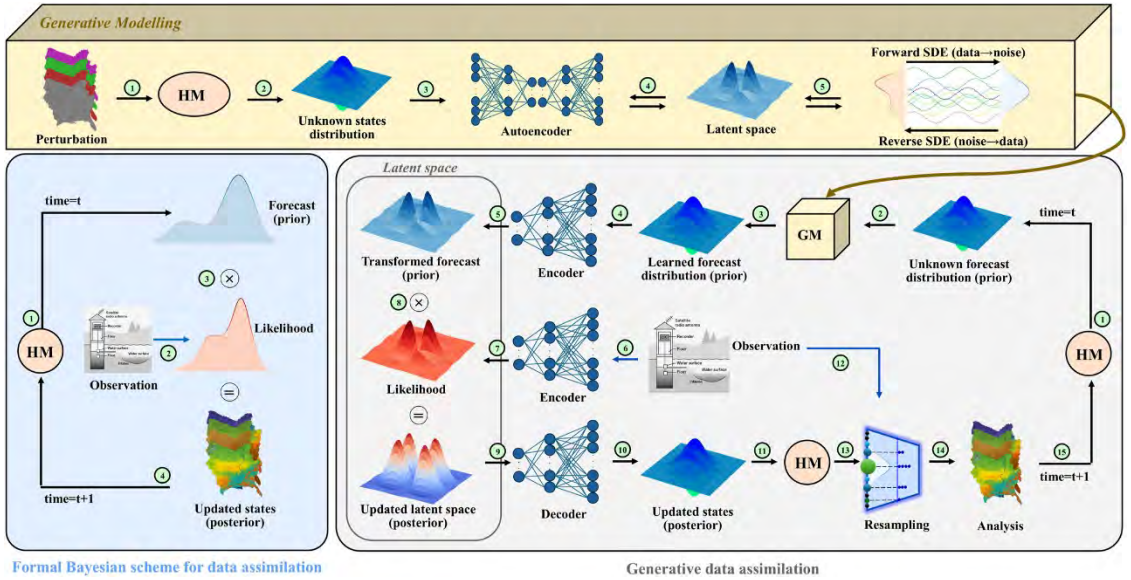
Harnessing Generative Deep Learning for Enhanced Ensemble Data Assimilation

Ehsan Foroumandi, Hamid Moradkhani

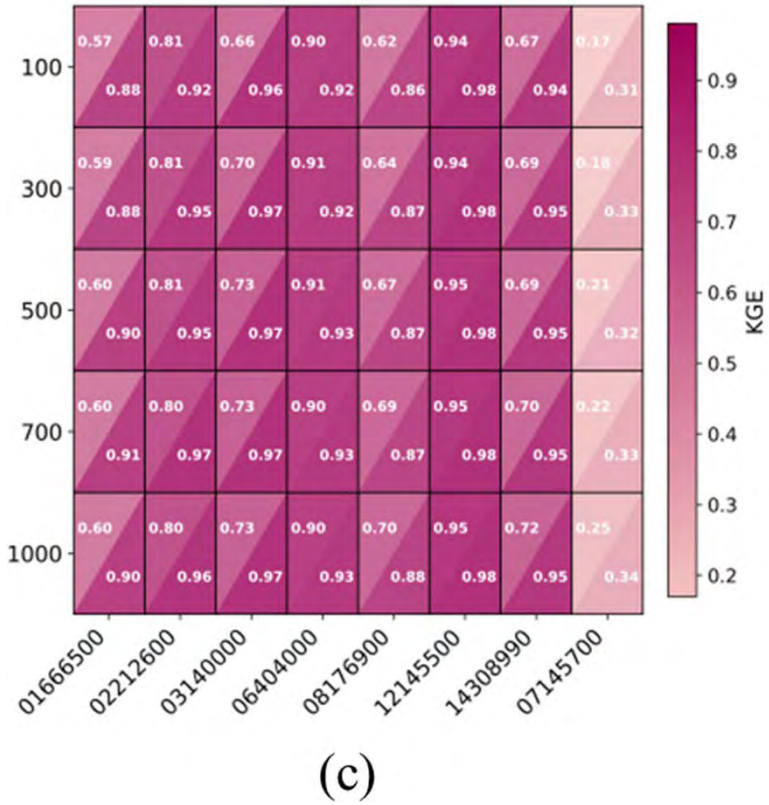
SNOW-17 + SAC-SMA, 8 U.S. basins

6 model states + 1 runoff flux compressed → 3 latent variables

USGS streamflow observations constrain latent-space sampling.



Decode and resample in physical space



Higher KGE than HEAVEN at all 8 gauges (100–1,000 members)

Foroumandi & Moradkhani, 2025, Water Resources Research.
https://doi.org/10.1029/2025WR040078

Baseflow assimilation with HydroGEnDA

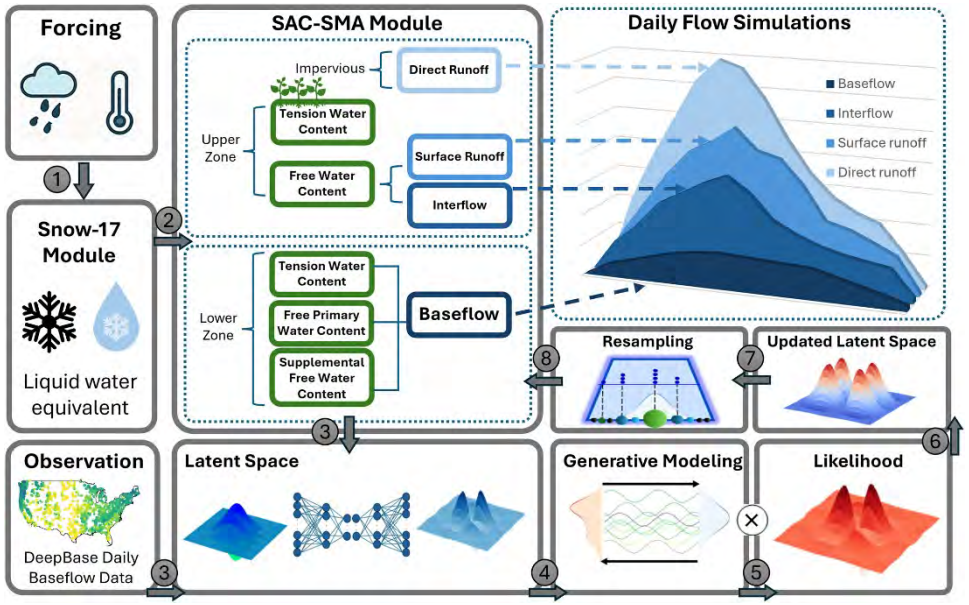
Geophysical Research Letters

Research Letter | Open Access

The Role of Baseflow Data Assimilation in Hydrologic Modeling and Peak Flow Prediction

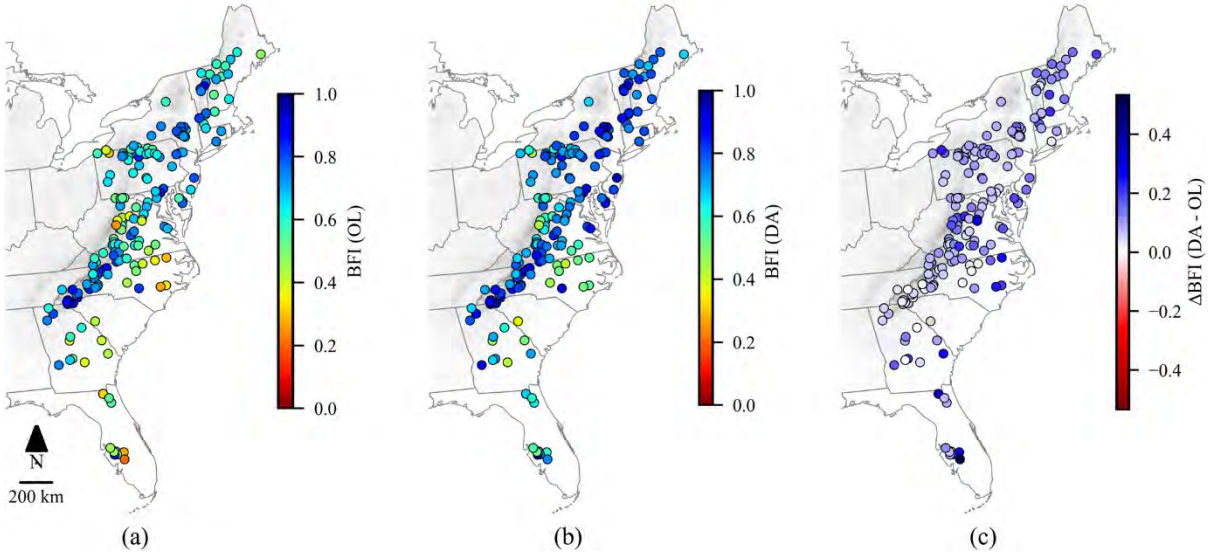
Parnian Ghaneei, Ehsan Foroumandi, Hamid Moradkhani

- SNOW-17 + SAC-SMA, 201 eastern U.S. basins
- Daily DeepBase baseflow estimates, 100 members



Baseflow constrains subsurface storage

DeepBase baseflow → latent-space update → decode and resample



Runoff partitioning

- Median BFI: 0.65 → 0.76
- BFI increases in 98% of basins

Streamflow prediction

- Median NSE improved by 30%
- Better flood-peak detection and magnitude estimates

Ghaneei, Foroumandi & Moradkhani, 2026, Geophysical Research Letters. https://doi.org/10.1029/2025GL121456

01

ML as an assimilation algorithm

Diffusion-based assimilation

02

Learning dynamic models or parameterizations

Latent-space assimilation

03

Learning observation operators

Learned snow depletion curves

04

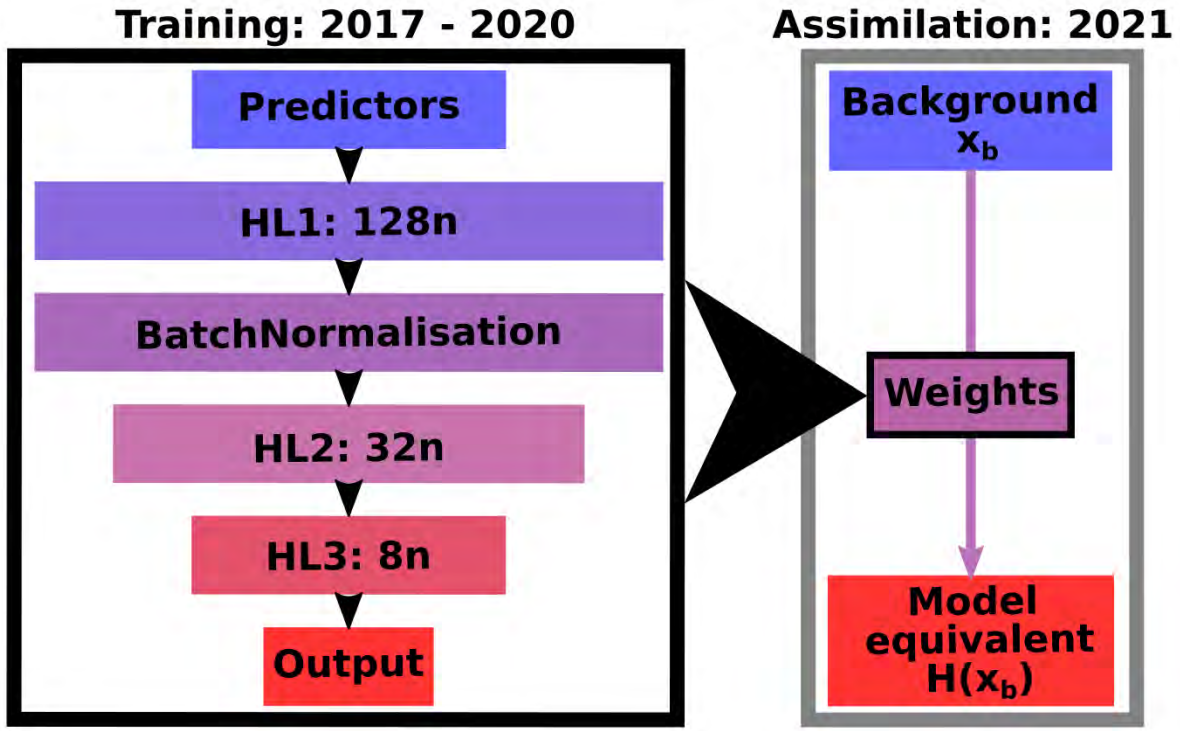
Bias correction

Model bias and observation errors

Machine learning for observation operators

A neural network learns the forward operator $\boldsymbol{H}$ that maps model states to simulated observations

Training and use of the operator



Observation operator $x \rightarrow \hat{y}$

Model state to simulated observation

Retrieval $y \rightarrow \hat{x}$

Observation to estimated land state

Simulate brightness temperature

$$\hat{y}^f = H_\theta(x^f, u)$$

ISBA model background and auxiliary inputs

Compute the innovation

$$d = y^{\text{obs}} - \hat{y}^f$$

Observed minus simulated brightness temperature

Update the land state

$$x^a = x^f + Kd$$

The SEKF updates LAI and multilayer soil moisture

Vural et al., 2026, Earth Observation. https://doi.org/10.5194/eo-1-155-2026

Learning snow depletion curves for CoLM

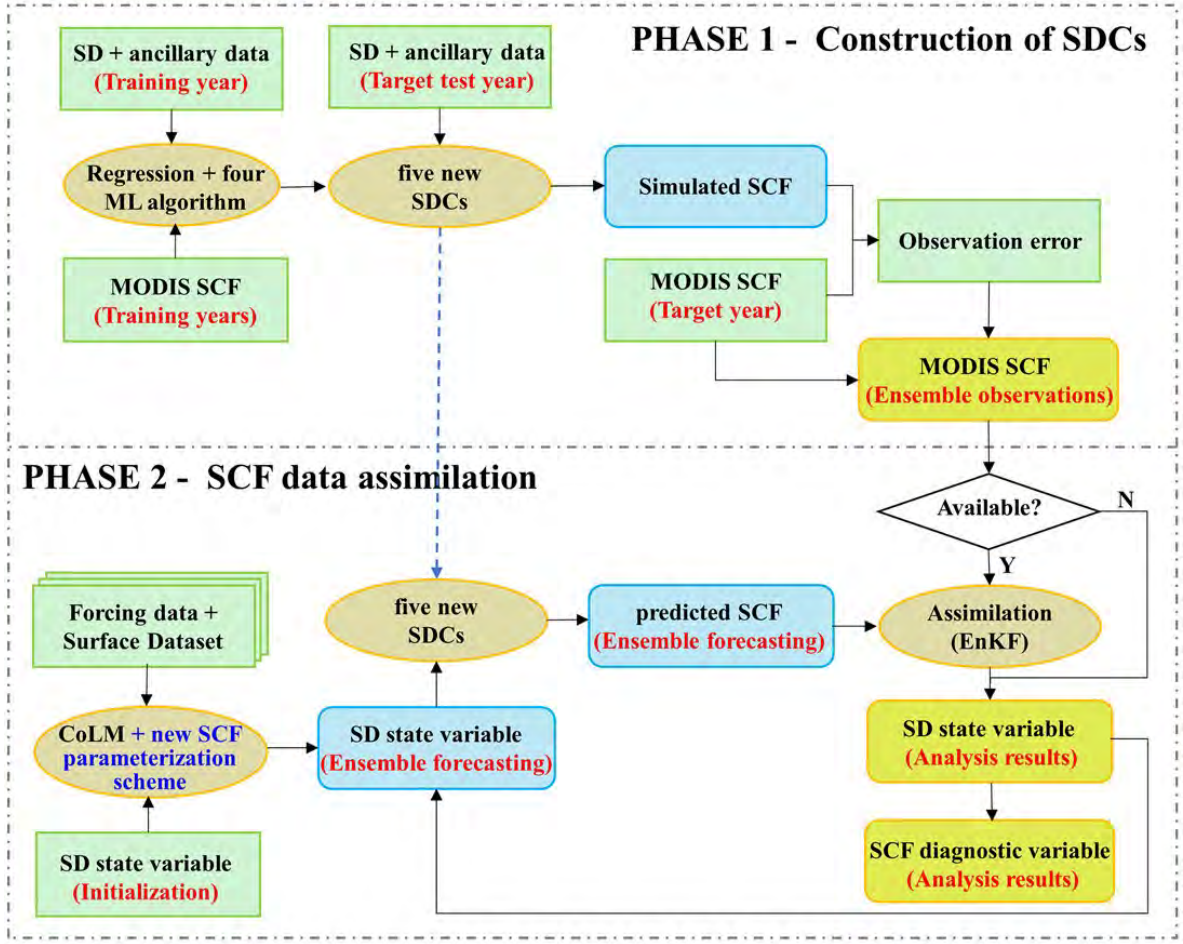
Water Resources Research

Research Article | Free Access

Improving Snow Estimates Through Assimilation of MODIS Fractional Snow Cover Data Using Machine Learning Algorithms and the Common Land Model

Jinliang Hou, Chunlin Huang ✉, Weijing Chen, Ying Zhang

- CoLM simulates snow depth, while MODIS observes snow cover fraction, so a mapping (**SDC**) is needed between model space and observation space.
- The learned **SDCs** were used to improve CoLM SCF parameterization and as the EnKF observation operator to assimilate original or gap-filled MODIS SCF and update snow depth.

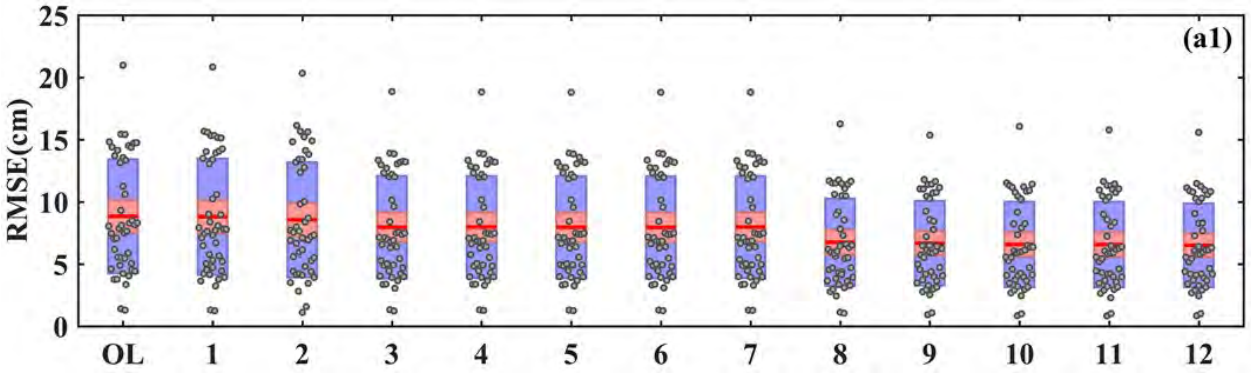


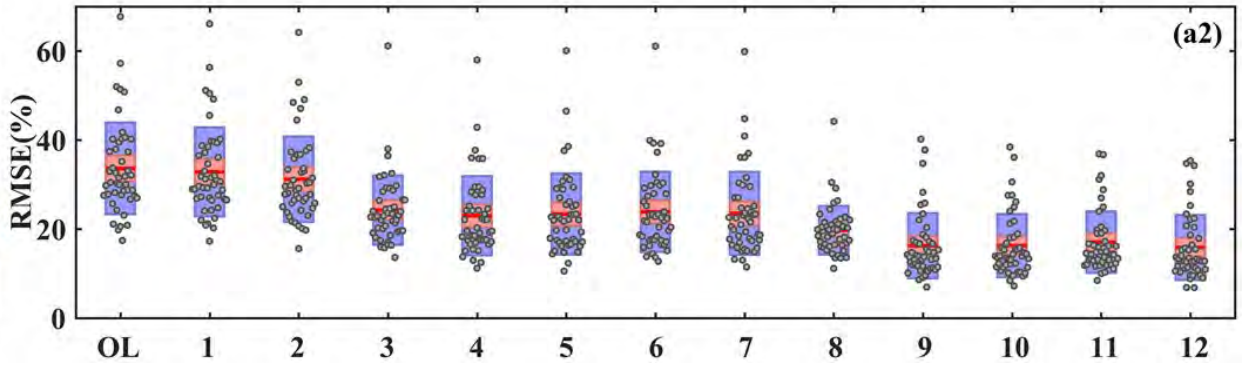
(b)

Hou et al., 2021, Water Resources Research. https://doi.org/10.1029/2020WR029010

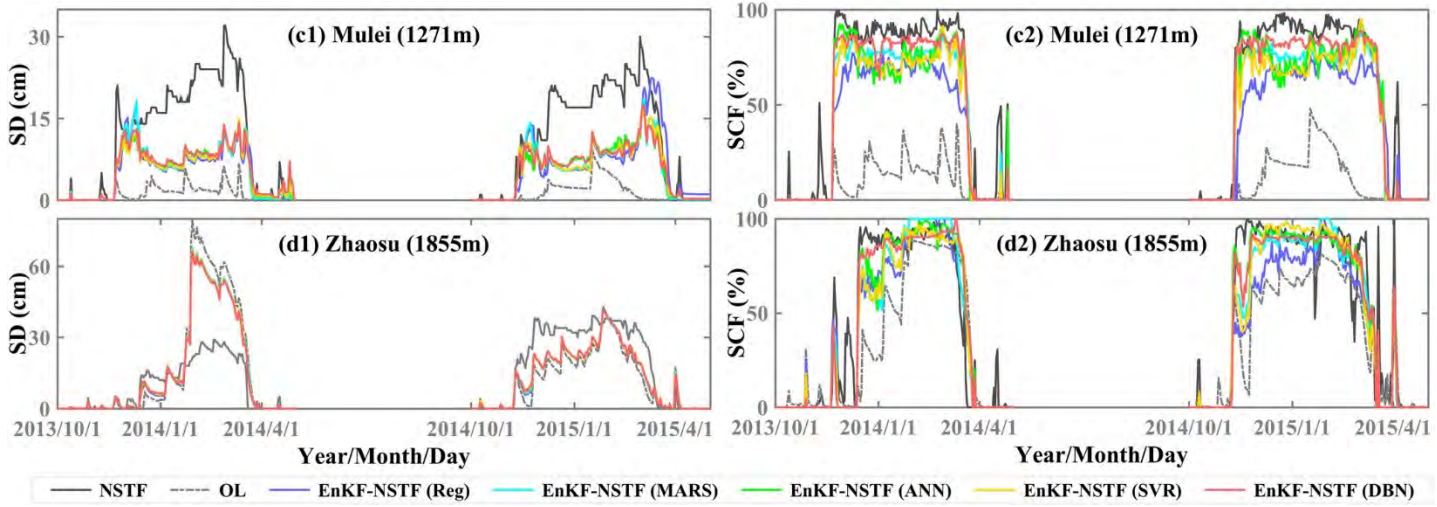
Snow data assimilation with learned operators

RMSE across 46 stations





1-2: default SDC + DI、3-7: learned SDC + EnKF-MOD、8-12: learned SDC + EnKF-NSTF



Best combined scheme

EnKF-NSTF (DBN)

SD RMSE ≈ 5.9 cm

SCF RMSE ≈ 16%

Mean NER vs. open loop

SD ≈ 32% SCF ≈ 48%

01

ML as an assimilation algorithm

Diffusion-based assimilation

02

Learning dynamic models or parameterizations

Latent-space assimilation

03

Learning observation operators

Learned snow depletion curves

04

Bias correction

Model bias and observation errors

Observation bias correction

Sources of bias in land DA

Systematic errors can skew the assimilation analysis.

Model and parameters

Missing processes or inaccurate parameters bias model predictions.

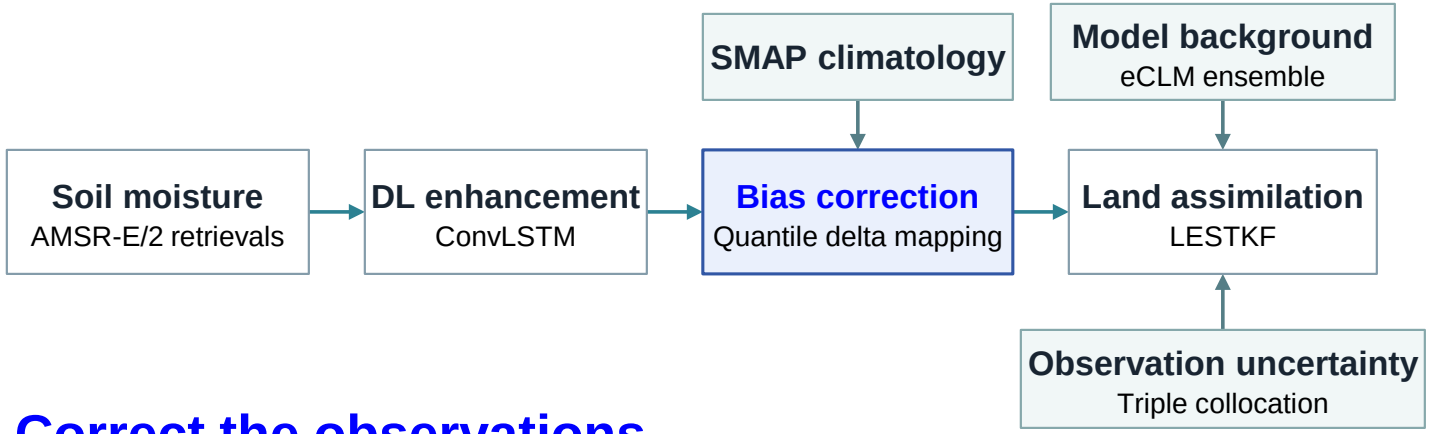
Meteorological forcing

Biased precipitation or other forcing data can distort the model state.

Observations

Calibration and retrieval errors or scale mismatches can introduce bias.

Observation bias correction in land DA



Correct the observations

$$y^{\mathrm{bc}} = \mathcal{C}_{\mathrm{ref}}(y^{\mathrm{ML}})$$

Apply a statistical correction to the ML-preprocessed observation

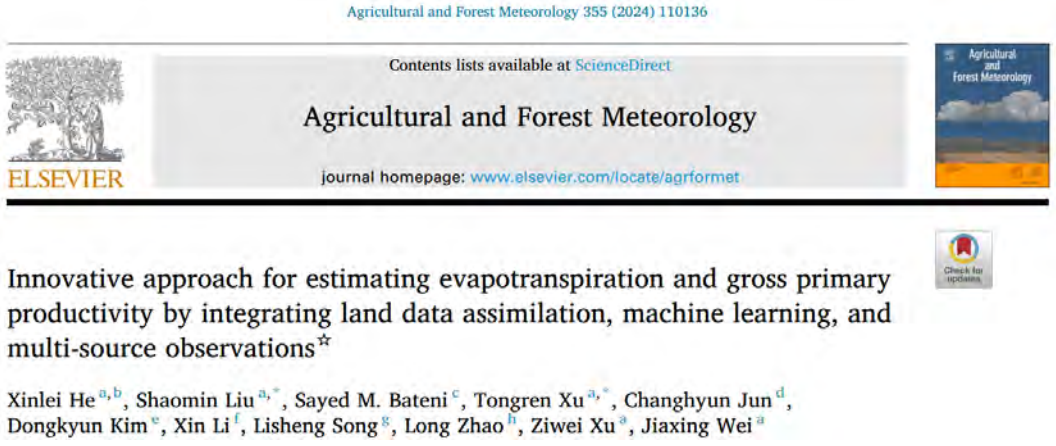
Update the model state

$$x^{a} = x^{f} + K[y^{\mathrm{bc}} - H(x^{f})]$$

Add the weighted innovation to the model background

Sivaprasad et al., 2026, Water Resources Research. https://doi.org/10.1029/2026WR043561

Soil moisture bias correction in Noah-MP


Agricultural and Forest Meteorology 355 (2024) 110136

Contents lists available at ScienceDirect

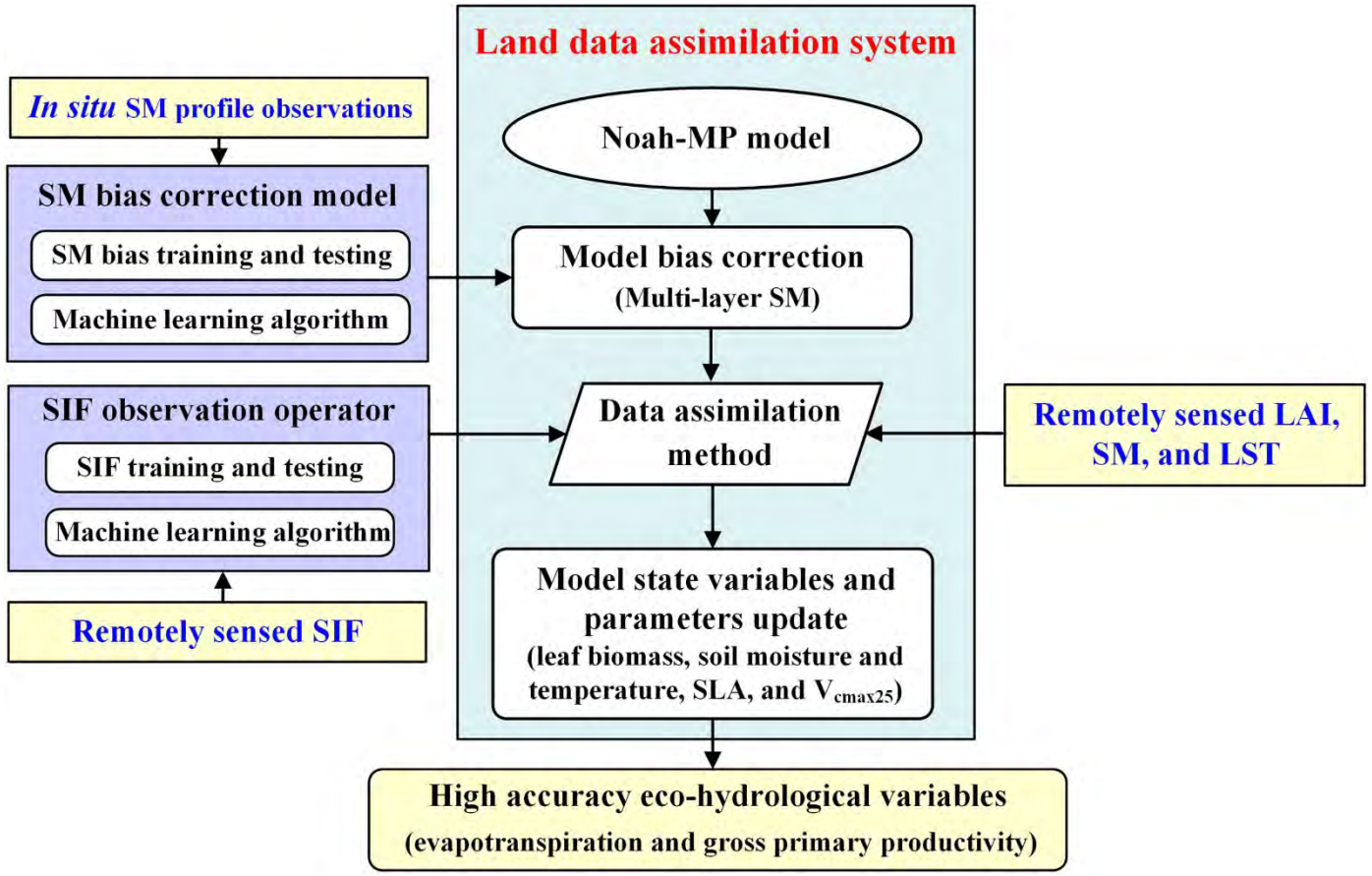
Agricultural and Forest Meteorology

journal homepage: www.elsevier.com/locate/agrformet

Innovative approach for estimating evapotranspiration and gross primary productivity by integrating land data assimilation, machine learning, and multi-source observations☆

Xinlei He[a,b], Shaomin Liu[a,*], Sayed M. Bateni[c], Tongren Xu[a,*], Changhyun Jun[d], Dongkyun Kim[e], Xin Li[f], Lisheng Song[g], Long Zhao[h], Ziwei Xu[a], Jiaxing Wei[a]

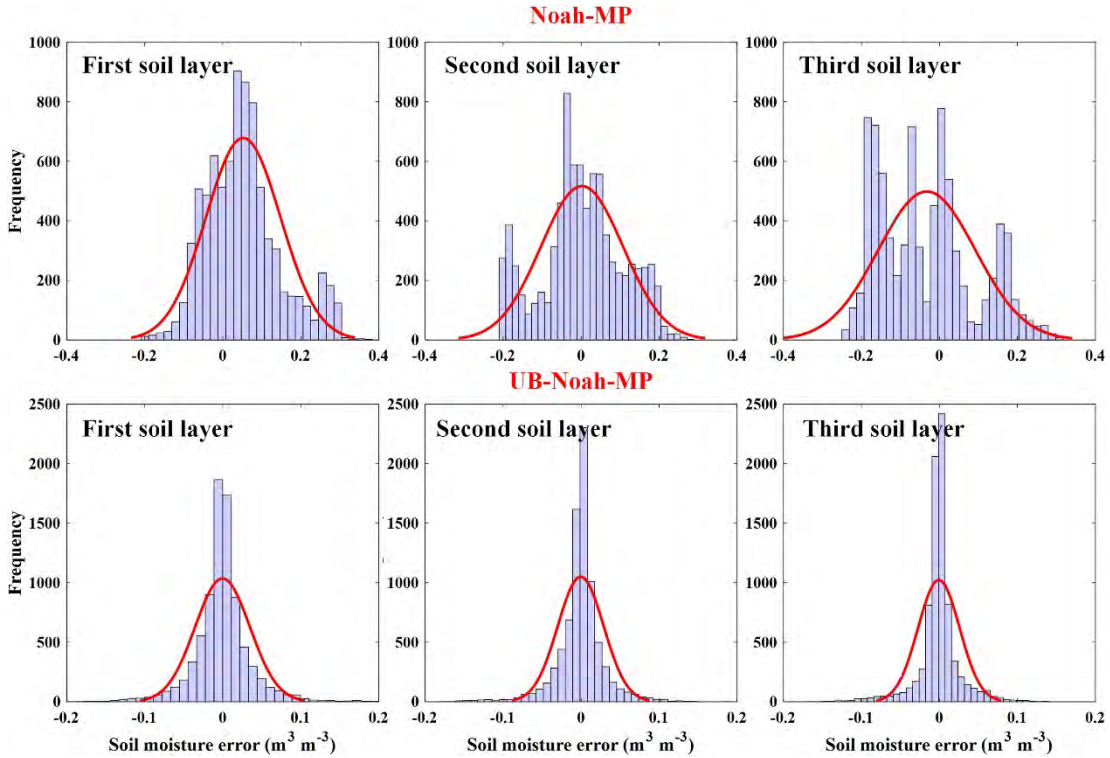
- XGBoost learns to correct soil moisture biases in three Noah-MP layers using in situ profiles.
- An ML-based observation operator maps Noah-MP outputs to solar-induced chlorophyll fluorescence (SIF).
- Satellite data assimilation updates Noah-MP to estimate ET and GPP.



He et al., 2024, Agricultural and Forest Meteorology. https://doi.org/10.1016/j.agrformet.2024.110136

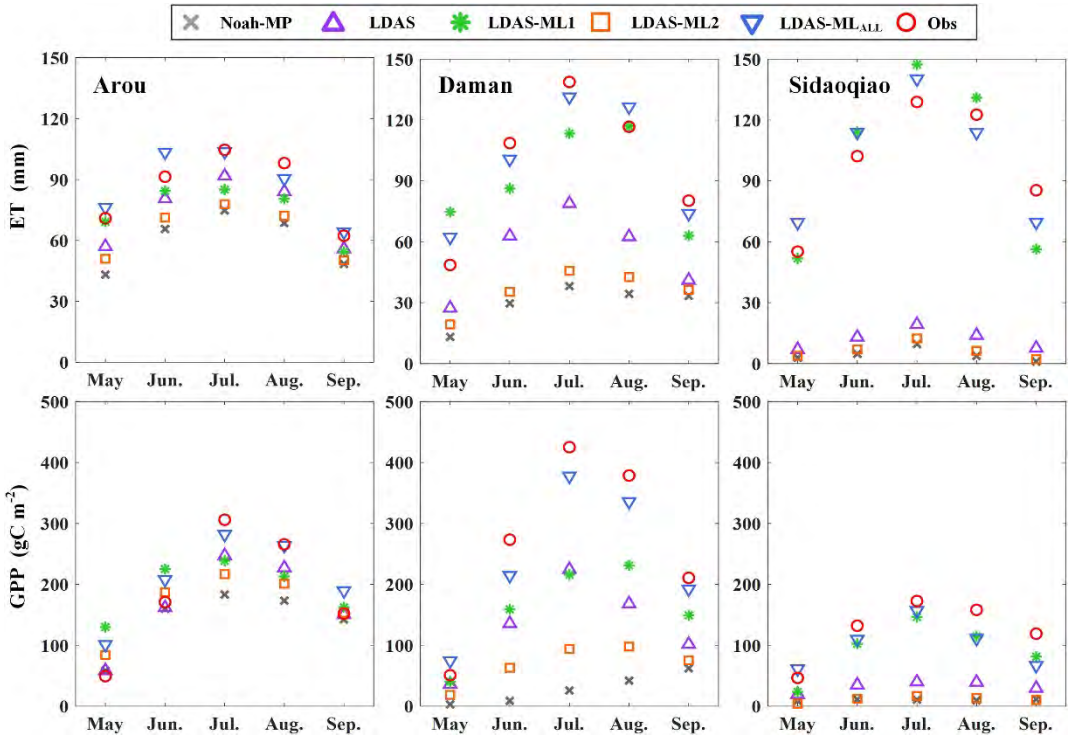
Bias correction improves soil moisture and flux estimates

Soil moisture errors cluster more tightly around zero



SM errors center closer to zero in all three layers.
UB-Noah-MP includes XGBoost correction.

Effect on ET and GPP estimations



LDAS-ML1 reduces the flux underestimation at Daman and Sidaoqiao relative to LDAS.

ML1: LDAS + SM bias correction.

ML2: learned SIF operator.

ALL: both ML components.

5. Key challenges and future research directions

Summary

1. LDA harmonizes theory and data under a Bayesian framework
2. DA and ML are mathematically comparable/isomorphic
3. ML serves four interfaces in LDA
4. Generative and diffusion methods open new DA pathways
5. Physics-informed ML integrates land DA and ML

Challenges and outlook

Challenges

1. Physical consistency and interpretability
2. Generalization across regions and scales
3. Uncertainty quantification
4. Stability and computational cost
5. Standardized evaluation

Outlook

1. End-to-end differentiable DA
2. Scalable multi-source assimilation
3. Coupled land–atmosphere DA
4. Physics-informed and generative ML
5. Social sensing and reanalysis

Machine learning for land DA in the AI era

nature reviews earth & environment
https://doi.org/10.1038/s43017-023-00409-w

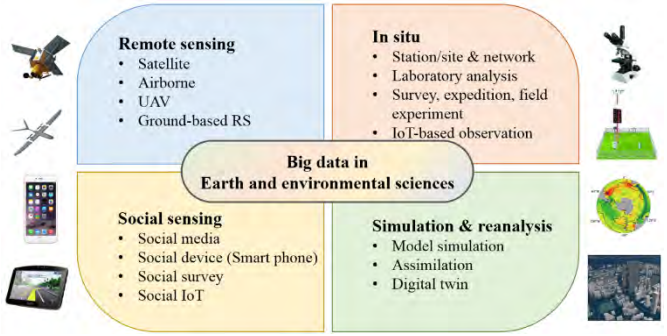
Review article

Check for updates

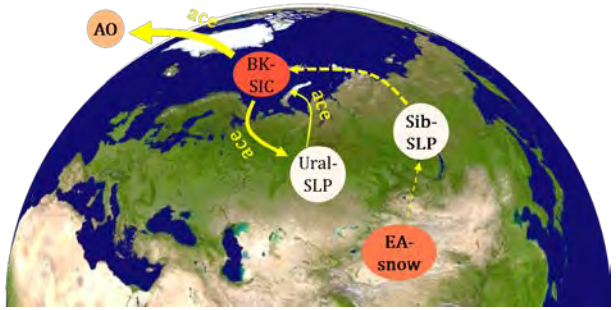
Big Data in Earth system science and progress towards a digital twin

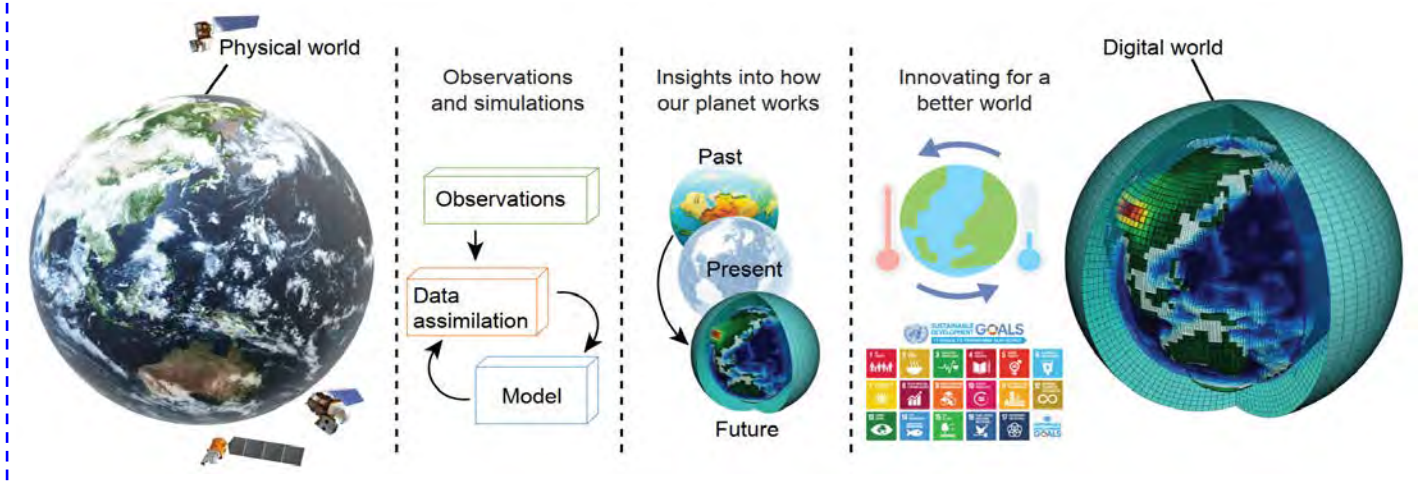
Xin Li[1,2,7], Min Feng[1,2,7], Youhua Ran[2,3], Yang Su[2,3], Feng Liu[3], Chunlin Huang[3], Huanfeng Shen[4], Qing Xiao[2,5], Jianbin Su[1], Shiwei Yuan[1] & Huadong Guo[2,5,6]

Big Earth data



Causal inference





Big data assimilation

Deep reinforcement learning

Li et al., 2023, Nature Reviews Earth & Environment. https://doi.org/10.1038/s43017-023-00409-w



Thank you for your attention!
I'd be happy to take any questions!

Acknowledgments:

Special thanks to Tinghua Mou and Dr. Feng Liu for their contributions to the literature review and slide preparation