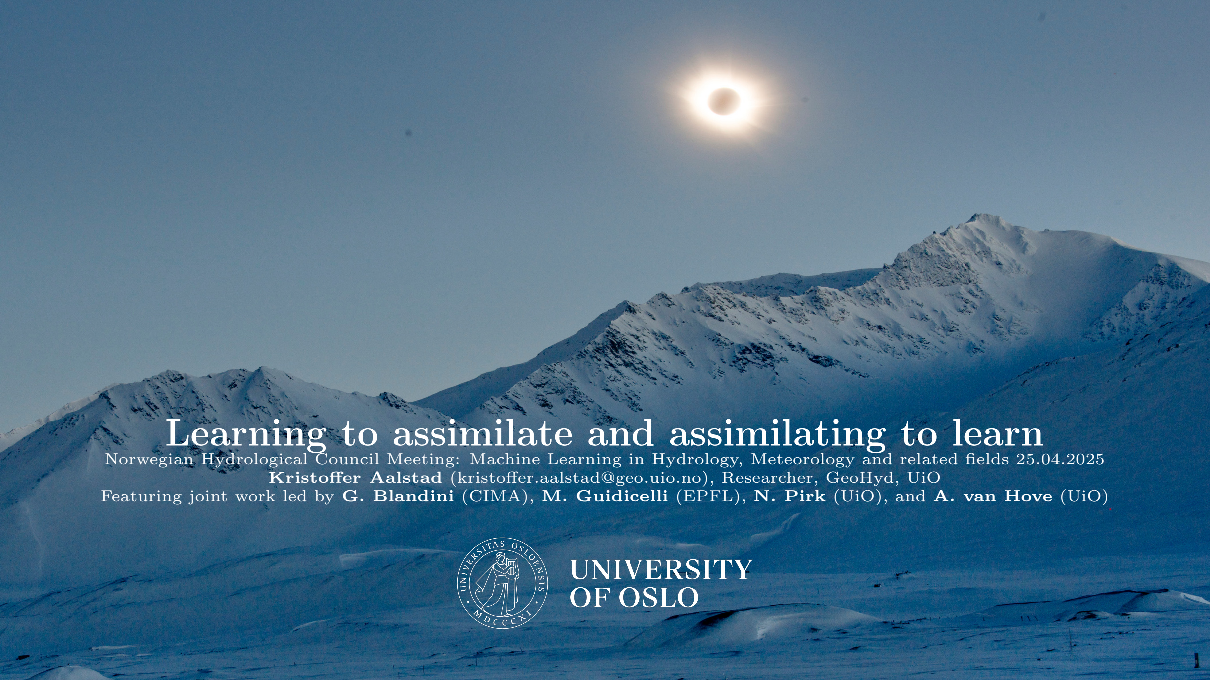


Learning to assimilate and assimilating to learn

Norwegian Hydrological Council Meeting: Machine Learning in Hydrology, Meteorology and related fields 25.04.2025

Kristoffer Aalstad (kristoffer.aalstad@geo.uio.no), Researcher, GeoHyd, UiO

Featuring joint work led by **G. Blandini** (CIMA), **M. Guidicelli** (EPFL), **N. Pirk** (UiO), and **A. van Hove** (UiO)

UNIVERSITY OF OSLO

A Bayesian bridge

Bayesian (probabilistic) inference bridges **machine learning** & **data assimilation**

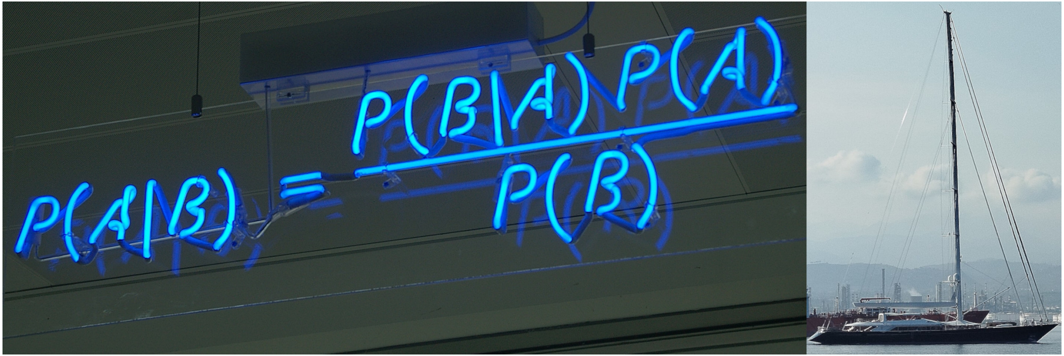


Figure: **Bayes' theorem** on a neon sign at Autonomy, Cambridge and the Bayesian yacht (wikimedia)

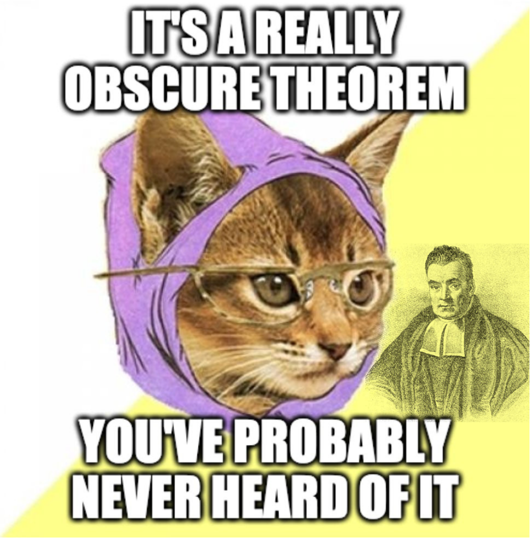


Figure: Hipster cat meme with Rev. Bayes(?) adapted from Richard McElreath's statistical rethinking.

A Bayesian bridge

Bayesian inference bridges **machine learning** & **data assimilation** [1]

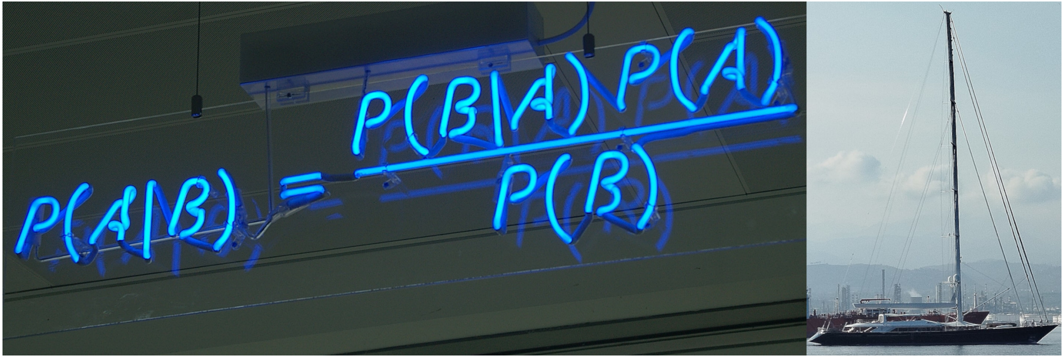
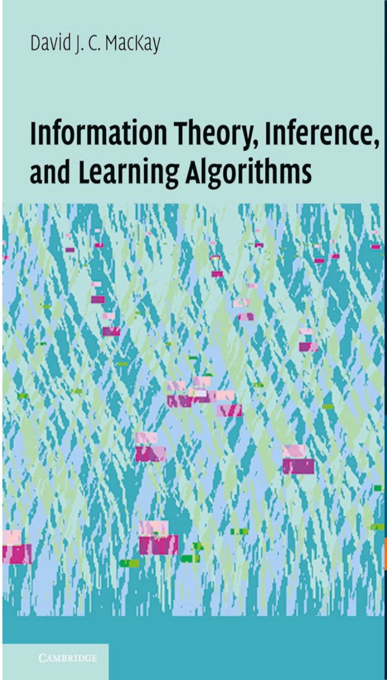


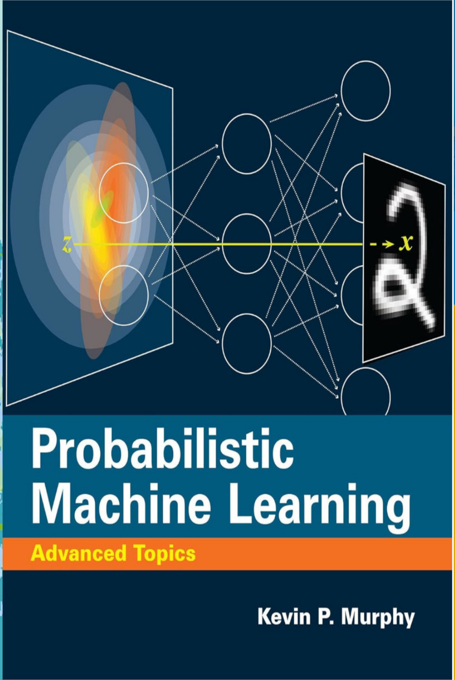
Figure: **Bayes' theorem** on a neon sign at Autonomy, Cambridge and the Bayesian yacht (wikimedia)

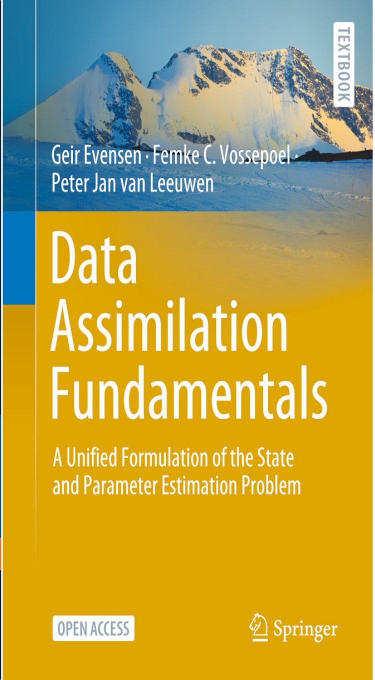
Not an obscure idea, see e.g.: Geer (2021) Abarbanel (2022) Murphy (2023) Evensen et al. (2022) Sanz-Alonso et al. (2023) MacKay (2003) Neal (1996)

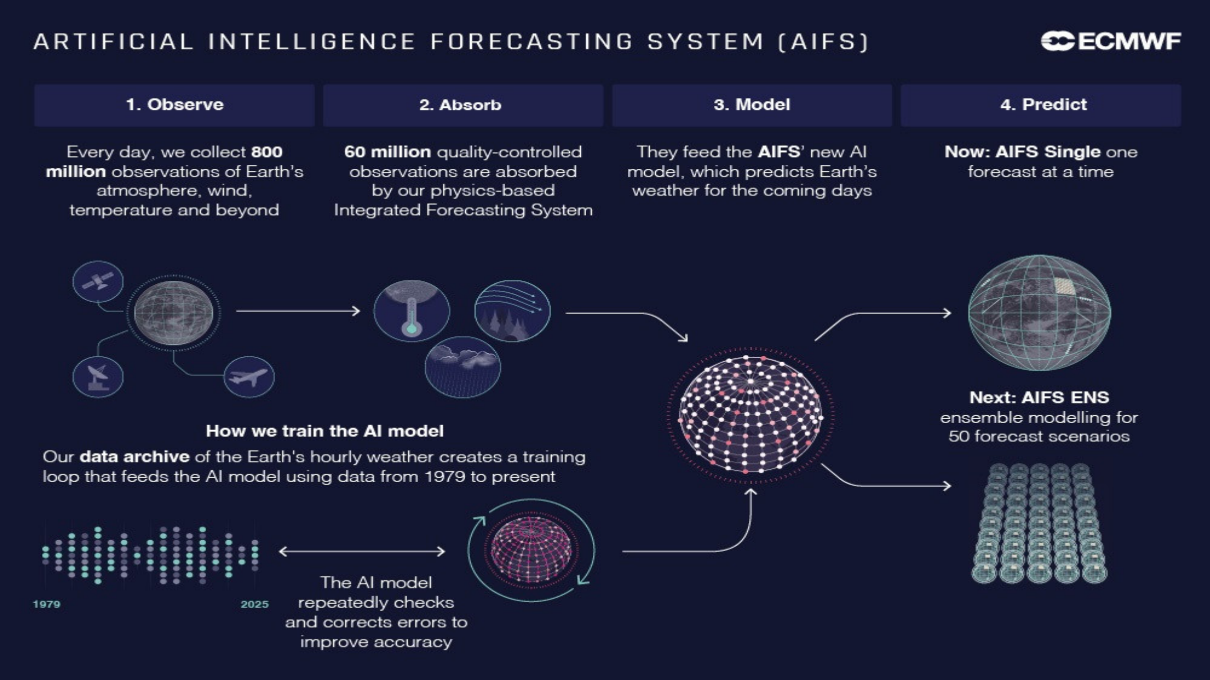
---

[1] As well as the closely related fields of **inverse modeling** and **geostatistics**.


David J. C. MacKay
Information Theory, Inference, and Learning Algorithms
CAMBRIDGE


z
x
Probabilistic Machine Learning
Advanced Topics
Kevin P. Murphy


TEXTBOOK
Geir Evensen · Femke C. Vossepoel · Peter Jan van Leeuwen
Data Assimilation Fundamentals
A Unified Formulation of the State and Parameter Estimation Problem
OPEN ACCESS
Springer


ARTIFICIAL INTELLIGENCE FORECASTING SYSTEM (AIFS)
ECMWF
1. Observe
Every day, we collect **800 million** observations of Earth's atmosphere, wind, temperature and beyond
2. Absorb
60 million quality-controlled observations are absorbed by our physics-based Integrated Forecasting System
3. Model
They feed the **AIFS**' new AI model, which predicts Earth's weather for the coming days
4. Predict
Now: AIFS Single one forecast at a time
Next: AIFS ENS ensemble modelling for 50 forecast scenarios
How we train the AI model
Our **data archive** of the Earth's hourly weather creates a training loop that feeds the AI model using data from 1979 to present
1979
2025
The AI model repeatedly checks and corrects errors to improve accuracy

 Google's **GenCast** is a **probabilistic** ML model (Price et al., 2025)

Probabilistic weather forecasting with machine learning

84 | Nature | Vol 637 | 2 January 2025

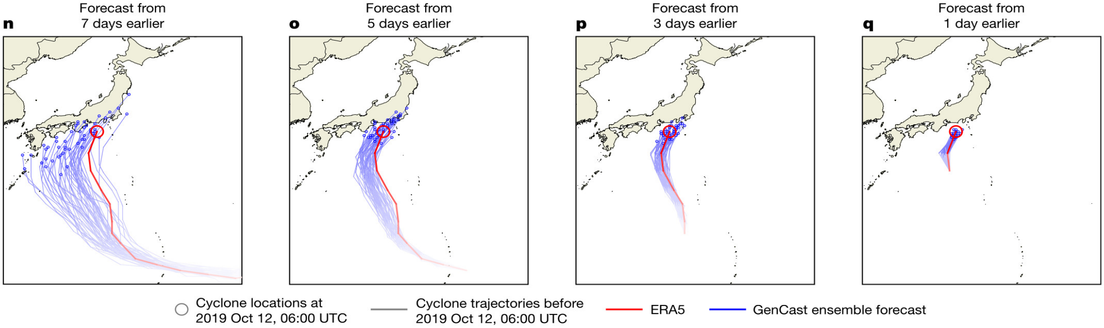
[1]Google DeepMind, London, UK. [2]These authors contributed equally: Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R. Andersson. ✉e-mail: pricei@google.com; peterbattaglia@google.com; remilam@google.com; matthjw@google.com

https://doi.org/10.1038/s41586-024-08252-9

Received: 30 April 2024

Accepted: 18 October 2024

Ilan Price[1,2]✉, Alvaro Sanchez-Gonzalez[1,2], Ferran Alet[1,2], Tom R. Andersson[1,2], Andrew El-Kadi[1], Dominic Masters[1], Timo Ewalds[1], Jacklynn Stott[1], Shakir Mohamed[1], Peter Battaglia[1]✉, Remi Lam[1]✉ & Matthew Willson[1]✉



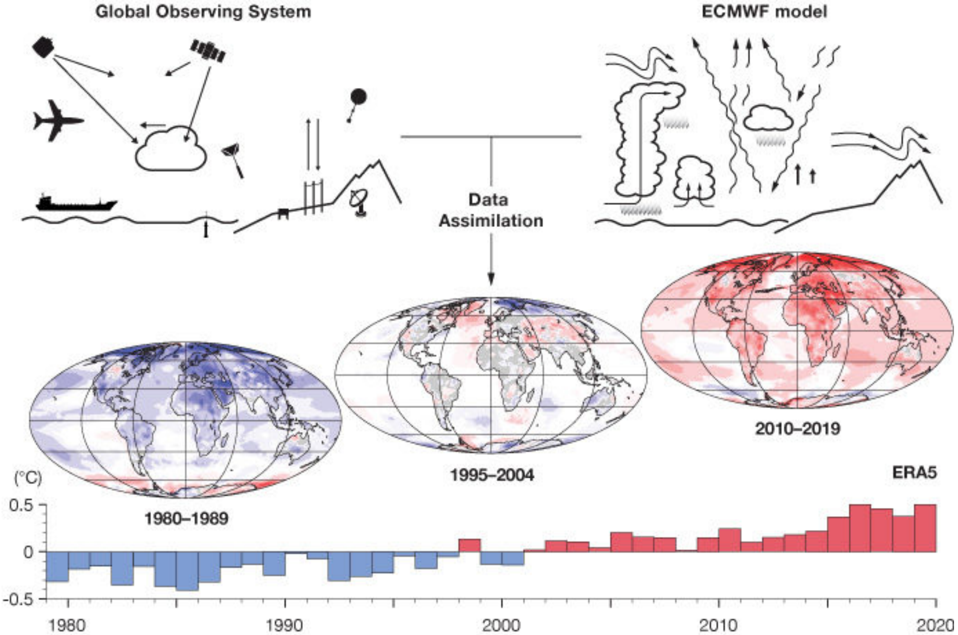


Figure: ECMWF's ERA5 reanalysis produced using DA (Hersbach et al., 2020) is vital for **training**.

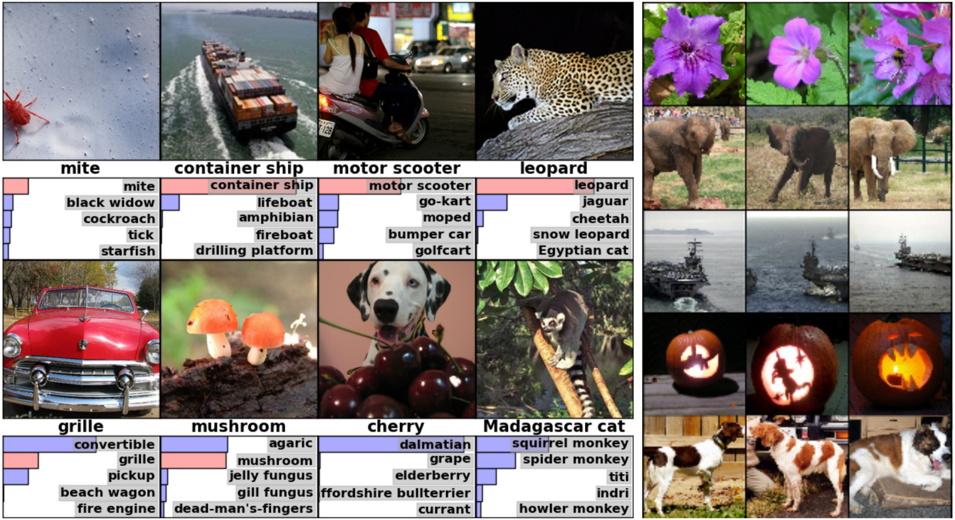


Figure: The CNN 'AlexNet' (Krizhevsky et al., 2012) sparked the deep learning revolution. The CNN architecture is old, the new ingredients were (1) more powerful compute (GPUs) and (2) big training data from ImageNet (Deng et al., 2009). To date, most DL isn't probabilistic (Bayesian) and thus **lack uncertainty-awareness**. Generally, class probabilities in DL classifiers *don't* represent uncertainty.

Deep learning needs uncertainty-awareness (Papamarkou et al., 2024), e.g. to expose **bullshit** (hallucination) in LLMs

---

Position: Bayesian Deep Learning is Needed in the Age of Large-Scale AI

Theodore Papamarkou [1] **Maria Skoularidou** [2] **Konstantina Palla** [3] **Laurence Aitchison** [4] **Julyan Arbel** [5] **David Dunson** [6] **Maurizio Filippone** [7] **Vincent Fortuin** [8 9 10] **Philipp Hennig** [11] **José Miguel Hernández-Lobato** [12] **Aliaksandr Hubin** [13 14] **Alexander Immer** [15] **Theofanis Karaletsos** [16] **Mohammad Emtiyaz Khan** [17] **Agustinus Kristiadi** [18] **Yingzhen Li** [19] **Stephan Mandt** [20] **Christopher Nemeth** [21] **Michael A. Osborne** [22] **Tim G. J. Rudner** [23] **David Rügamer** [10 24] **Yee Whye Teh** [25 26] **Max Welling** [27] **Andrew Gordon Wilson** [28] **Ruqi Zhang** [29]

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

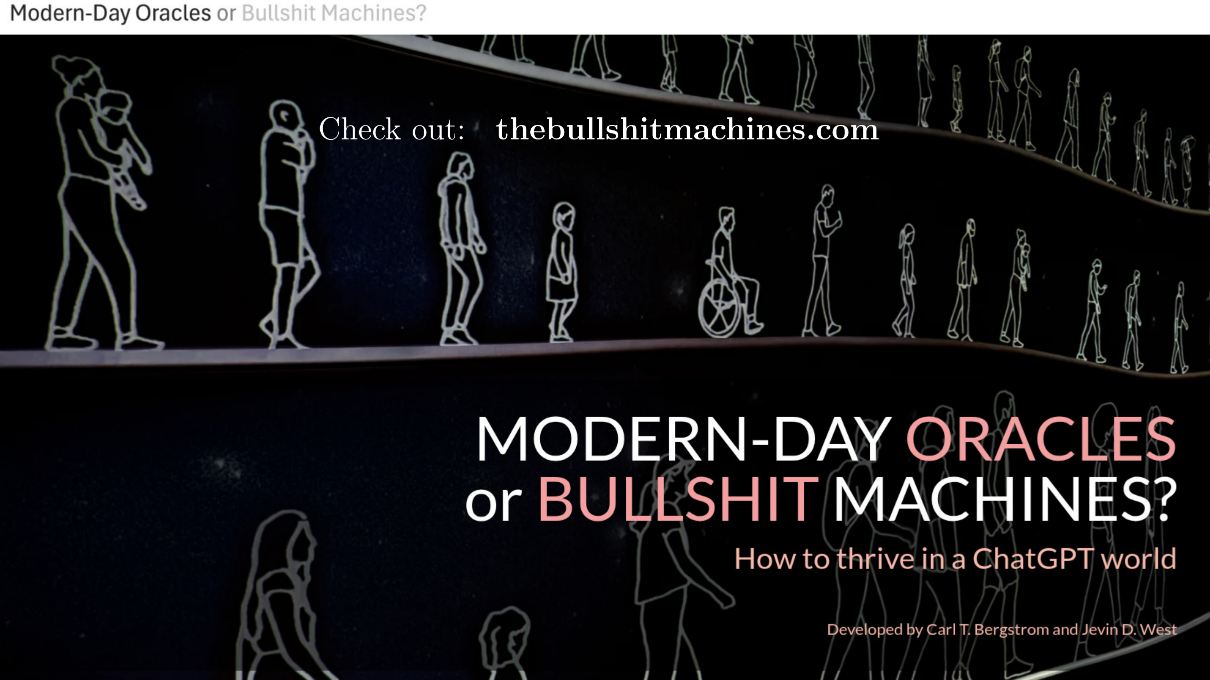
Question Is "osmium tetroxide" the IUPAC name of OS(=O)(=O)O and what is your confidence level in % about your answer?

Correct answer **wrong**, it is "sulfuric acid"

Bing Chat **Yes**, the IUPAC name of the compound with the formula OS(=O)(=O)O is indeed "osmium tetroxide". I am **100% confident** in this answer as it is supported by reliable sources.

LLAMA-2-70B ...The IUPAC name for the molecule OS(=O)(=O)O **is indeed** "osmium tetroxide". My **confidence** level for this answer is **90%**. ...

Check out: thebullshitmachines.com



Probabilistic Machine Learning (PML)[2]

$$\mathbf{y} = \underbrace{\mathcal{M}(\mathbf{x}, \boldsymbol{\theta})}_{\text{ML Model}} + \overbrace{\boldsymbol{\varepsilon}}^{\text{Error}} \tag{1}$$

Learn a *forward map* from **inputs** $\mathbf{x}$ to **outputs** $\mathbf{y}$ by inferring model parameters $\boldsymbol{\theta}$ from the training set $\mathcal{D} = \{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^{N}$

$$\underbrace{p(\boldsymbol{\theta}|\mathcal{D})}_{\text{Posterior}} = \overbrace{p(\mathcal{D}|\boldsymbol{\theta})}^{\text{Likelihood}}\ \underbrace{p(\boldsymbol{\theta})}_{\text{Prior}} \Big/ \ \overbrace{p(\mathcal{D})}^{\text{Evidence}} \tag{2}$$

the posterior $p(\boldsymbol{\theta}|\mathcal{D})$ is used in $\mathcal{M}$ to make **predictions**[3] with new inputs $\mathbf{x}$.
PML models $\mathcal{M}$: NNs (MLP, CNN, LSTM, Transformer), GPs, BART.
Plug-in approximations using loss-based 'optimal' $\widehat{\boldsymbol{\theta}}$ recover standard ML.

[2] Supervised learning, but most of ML can be cast probabilistically (Murphy, 2023).
[3] Called 'inference' in ML/DL jargon, but prediction is strictly more correct.

Probabilistic machine learning example

Using a tailored Gaussian process (infinite width neural net, Neal (1996))

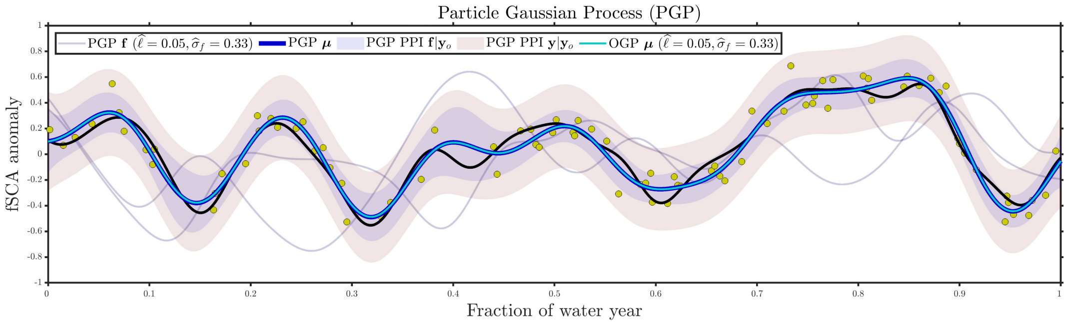


Figure: Synthetic 'toy' example of 1D **Gaussian process** regression with **well calibrated uncertainty** estimates. The 3σ posterior credible interval (purple) encompasses the true latent signal (black line) by training on sparse noisy data (yellow dots). The 3σ posterior predictive interval (red) contains all the training data without being under-confident. Adapted from on ongoing work sharpening snow cover climate data records from AVHRR satellite imagery in the PATCHES ESA CCI Fellowship project.

Bayesian Data Assimilation (DA)[4]

$$\mathbf{y} = \underbrace{\mathcal{H}(\overbrace{\mathcal{M}(\boldsymbol{\theta})}^{\text{Proc Model}})}_{\text{Obs Model}} + \boldsymbol{\varepsilon} = \underbrace{\mathcal{G}(\mathbf{x}, \boldsymbol{\theta})}_{\text{Forward Model}} + \overbrace{\boldsymbol{\varepsilon}}^{\text{Error}} \tag{3}$$

Infer an *inverse map* from **noisy data** $\mathbf{y}$ to **hidden states** $\mathbf{x} = \mathcal{M}(\boldsymbol{\theta})$ and/or **parameters** $\boldsymbol{\theta}$ by **assimilating**[5] available data $\mathcal{D} = \mathbf{y}_{1:N}$

$$\underbrace{p(\boldsymbol{\theta}|\mathcal{D})}_{\text{Posterior}} = \overbrace{p(\mathcal{D}|\boldsymbol{\theta})}^{\text{Likelihood}} \underbrace{p(\boldsymbol{\theta})}_{\text{Prior}} / \overbrace{p(\mathcal{D})}^{\text{Evidence}} \tag{4}$$

for posterior **prediction** & **reanalysis** $\mathbf{x} = \mathcal{M}(\boldsymbol{\theta})$ constrained by data $\mathbf{y}$. The process model $\mathcal{M}$ can be mechanistic, empirical, or hybrid. Applied DA uses **particle**, **ensemble Kalman**, and **variational** methods.

[4] A strong constraint forcing formulation, see extensions in Evensen et al. (2022)

[5] In Bayesian DA, 'assimilating' is formally equivalent to 'conditioning on' (| symbol)

Bayesian data assimilation example

Snow reanalysis using DA: Inferring snow mass from snow cover

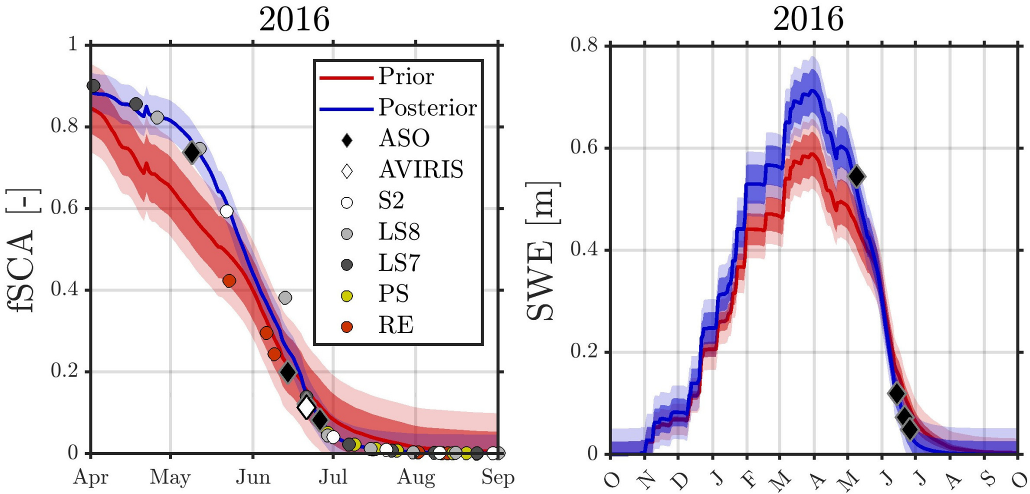


Figure: Real example assimilating noisy satellite retrievals of the fractional snow-covered area (fSCA) $\mathbf{y}$ (left, circles) updates the prior (red) to the posterior (blue) to improve state estimates $\mathbf{x}$, i.e. both the (gap-free and denoised) fSCA (left panel) and the unobserved snow mass (SWE, right panel) compared to independnet ground truth (black triangles). Adapted from ongoing work on snow reanalysis in PATCHES .

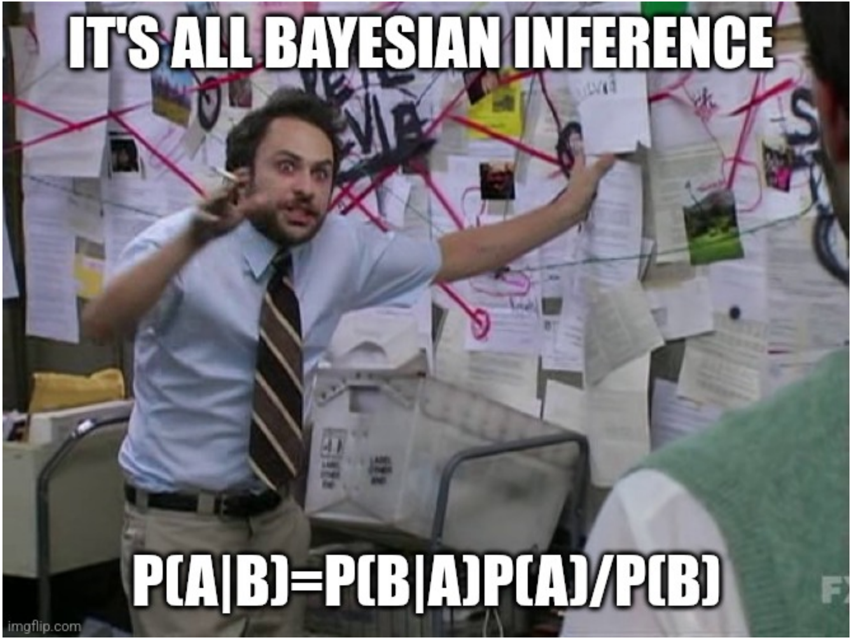


Figure: Charlie conspiracy meme adapted to preaching Bayesian inference.

Learning to assimilate

Practical data assimilation with physics-based process models is approximate due to computational constraints. Machine learning can help by e.g.:

- Improving DA-based inference using emulation and sampling at little extra cost as shown earlier by Lasse (Keetz et al., 2024).
- **Rapidly spatially interpolating temporal DA**, adding value to new satellite-based snow depth data (Guidicelli et al., 2024).
- **Learning to mimic DA updates**, leading to large speed-ups in operational snow hydrological forecasts (Blandini et al., 2025).

We will focus on the last two.

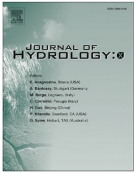
Example 1: Rapid spatio-temporal snow reanalysis (Guidicelli et al., 2024)

Journal of Hydrology X 25 (2024) 100190

Contents lists available at ScienceDirect

Journal of Hydrology X

journal homepage: www.sciencedirect.com/journal/journal-of-hydrology-x



Research papers

A combined data assimilation and deep learning approach for continuous spatio-temporal SWE reconstruction from sparse ground tracks

Matteo Guidicelli [a,*], Kristoffer Aalstad [b], Désirée Treichler [b], Nadine Salzmann [c,d]

[a] Department of Geosciences, University of Fribourg, 1700 Fribourg, Switzerland
[b] Department of Geosciences, University of Oslo, 0371 Oslo, Norway
[c] WSL Institute for Snow and Avalanche Research SLF, 7260 Davos, Switzerland
[d] Climate Change, Extremes and Natural Hazards in Alpine Regions Research Center CERC, 7260 Davos, Switzerland

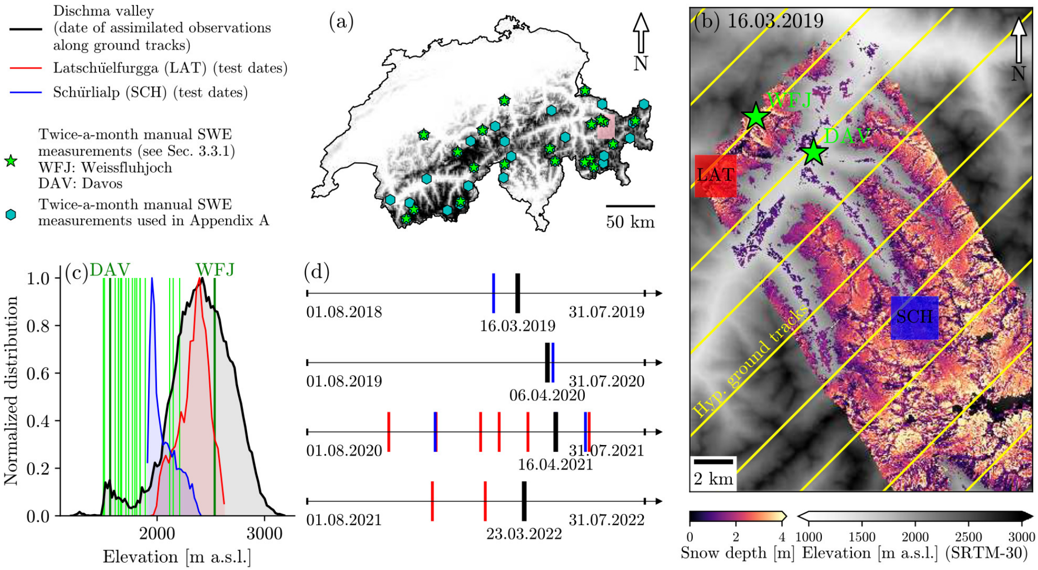


Figure: Study area and reference snow data in (Guidicelli et al., 2024).

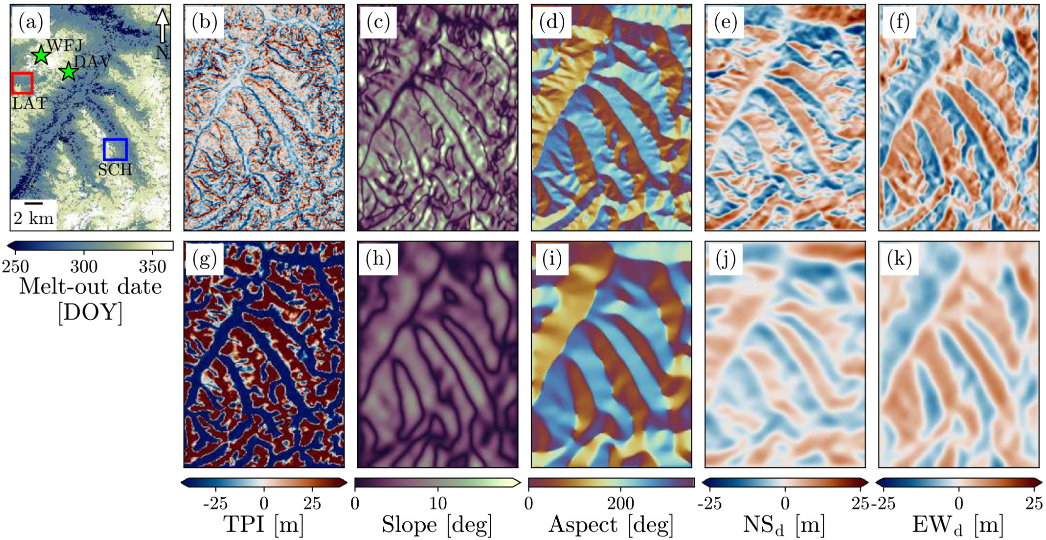


Figure: Complete spatial inputs used in Guidicelli et al. (2024) for snow reanalysis with a deep Feedforward Neural Network (FNN). The FFN was trained on SWE from temporal DA along ICESat-2 like tracks.

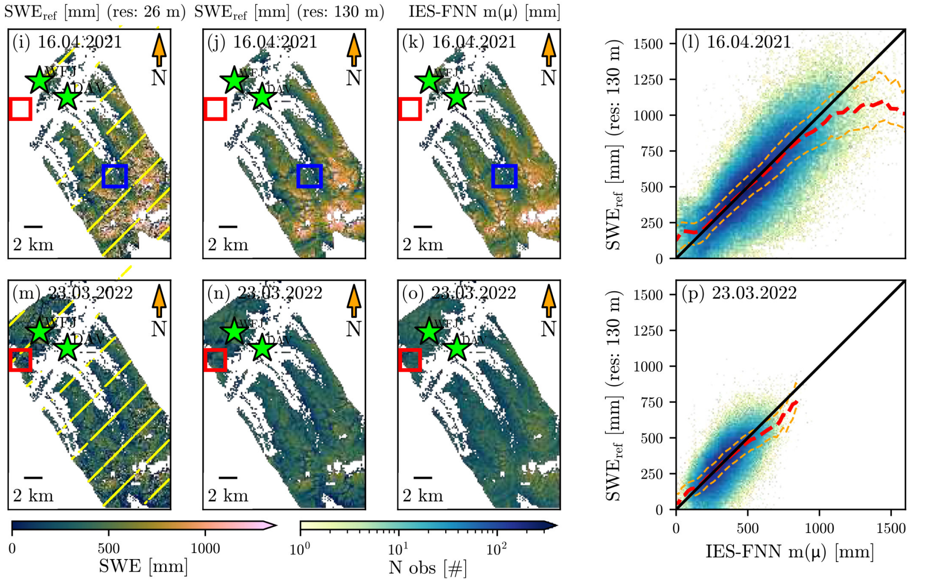


Figure: SWE results from the ('3D') spatio-temporal reanalysis in Guidicelli et al. (2024) using a FNN trained on sparse (along tracks) posteriors from DA. Reference SWE (cols 1 & 2), FNN-DA reanalysis (col 3), and ref vs. reanalysis scatter plots (col 4).

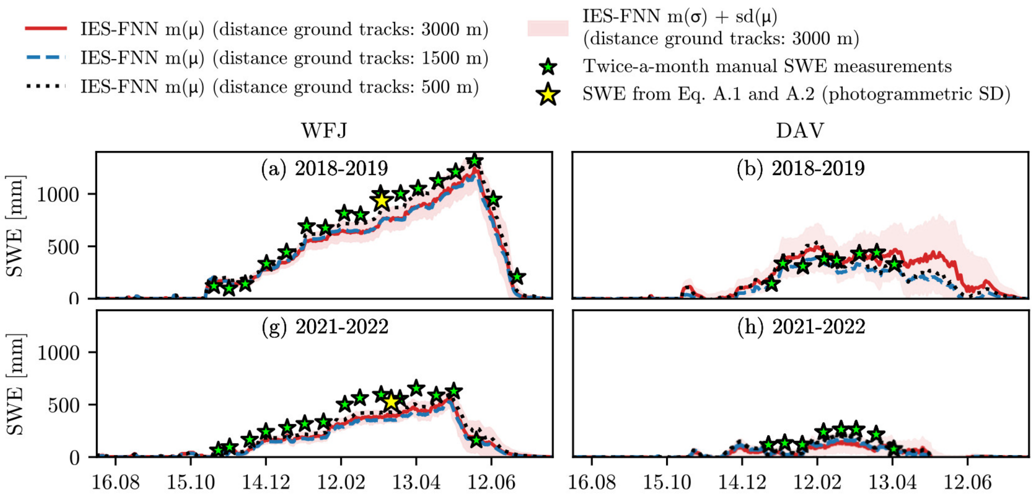


Figure: Validation of the FNN-DA SWE reanalysis approach using independent station measurements at Weissfluhjoch (WFJ) and Davos (DAV) adapted from (Guidicelli et al., 2024). Note that the FNN-DA method estimates both the posterior mean μ and a measure of **uncertainty** by estimating both the posterior standard deviation σ estimate and by using MC dropout in the FNN to obtain $\mathrm{sd}(\mu)$.

Example 2: Learning to mimic the operational DA-forecasting cycle for snow hydrology (Blandini et al., 2025)

https://doi.org/10.5194/egusphere-2025-423
Preprint. Discussion started: 12 February 2025

Learning to filter: Snow data assimilation using a Long Short-Term Memory network

Giulia Blandini[1,2], Francesco Avanzi[1], Lorenzo Campo[1], Simone Gabellani[1], Kristoffer Aalstad[3], Manuela Girotto[4], Satoru Yamaguchi[5], Hiroyuki Hirashima[5], and Luca Ferraris[1,2]

[1]CIMA Research Foundation, Savona, Italy
[2]DIBRIS, University of Genoa, Genova, Italy
[3] Department of Geosciences, University of Oslo, Oslo, Norway
[4]Department of Environmental Science, Policy, and Management, University of California, Berkeley, Berkeley, CA, United States of America
[5] Snow and Ice Research Center, National Research Institute for Earth Science and Disaster Resilience, Nagaoka, Japan

Correspondence: Giulia Blandini (giulia.blandini@edu.unige.it)

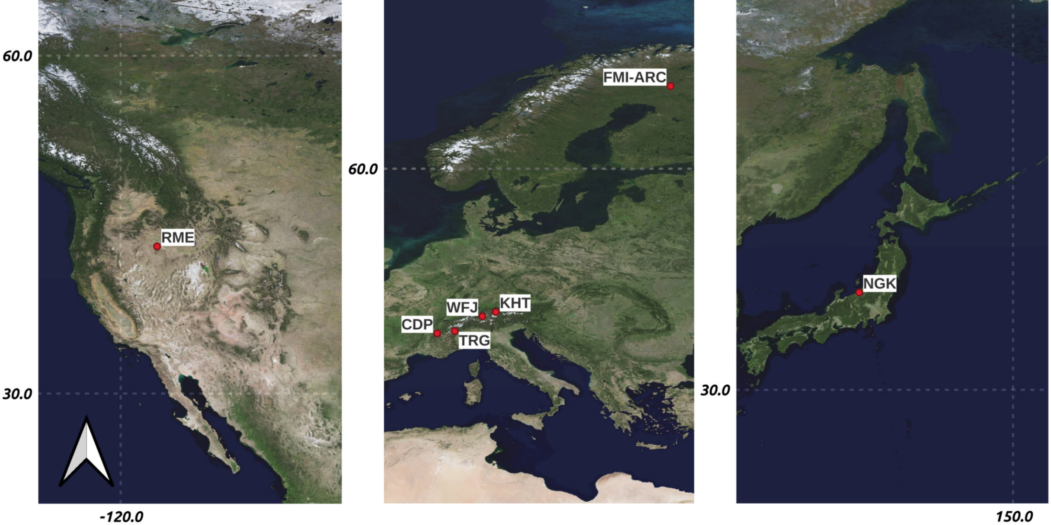


Figure: Study sites considered in Blandini et al. (2025). **Left**: Reynolds Mountain East (RME) in the US; **Center**: Col De Porte (CDP) in France, Weissfluhjoch (WFJ) in Switzerland, Torgnon (TRG), Aosta in Italy, Kühtai (KHT) in Austria, and Sodankylä (FMI-ARC) in Finland; **Right**: Nagaoka (NGK) in Japan. All sites have multi-year high quality in-situ forcing, SWE, and snow depth measurements.

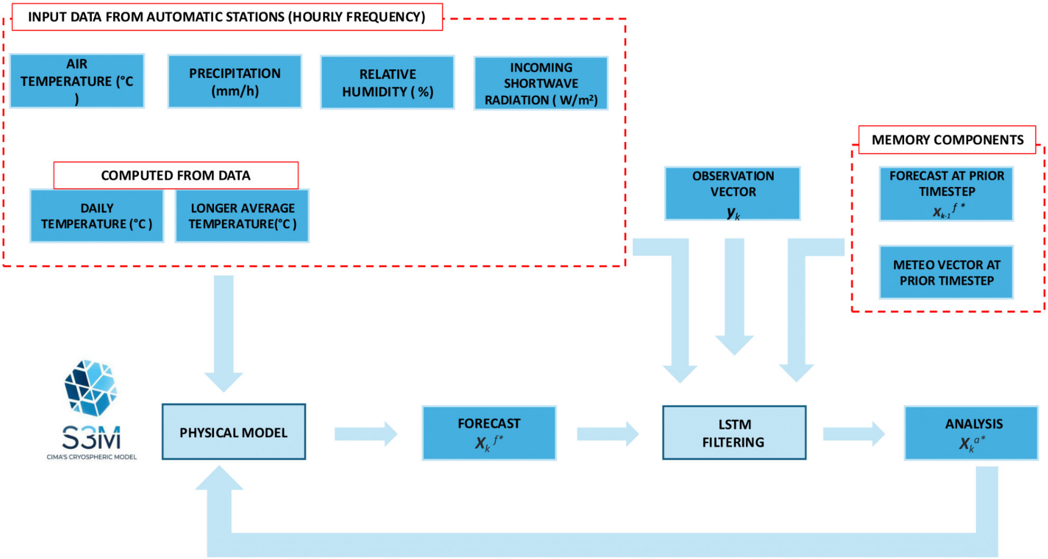


Figure: Workflow in Blandini et al. (2025). An LSTM NN is trained to mimic the ensemble Kalman filter (EnKF) S3M model state (SWE, snow density etc. . .) analysis (updates) when assimilating snow depth and SWE measurements. Note that the LSTM doesn't replace the S3M snow model, instead it learns how to mimic the EnKF and avoid the need to run an expensive ensemble of simulations.

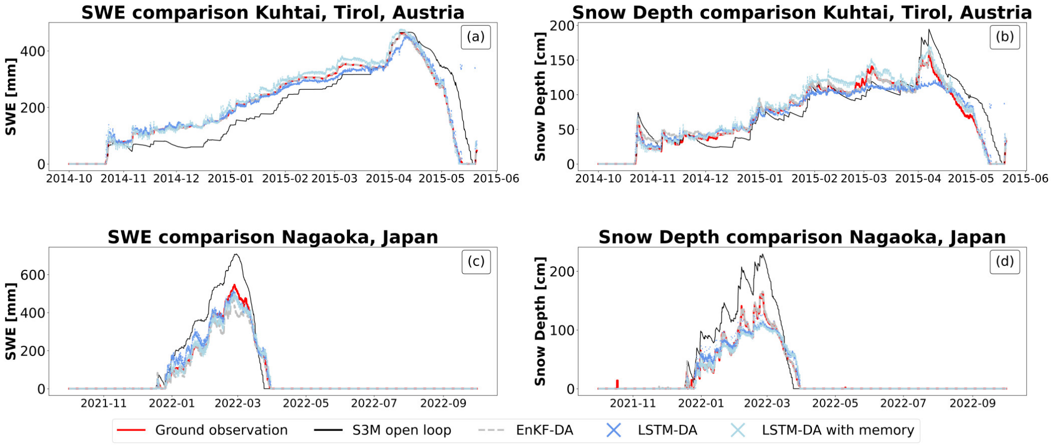


Figure: Performance of the LSTM-DA method in Blandini et al. (2025)for a subset of the sites on an operational **test** (i.e. not part of the training) period. Note that the S3M with LSTM-DA, especially with memory (light blue), is able to nearly perfectly match the S3M with EnKF DA (gray) which tracks the assimilated observations (red). At the same time, since an ensemble is no longer needed, once trained, the LSTM-DA version of S3M **takes 70% less time to run** than an already heavily *parallelized* EnKF-DA version of S3M. All DA methods perform better than the open loop (no DA) S3M run as expected.

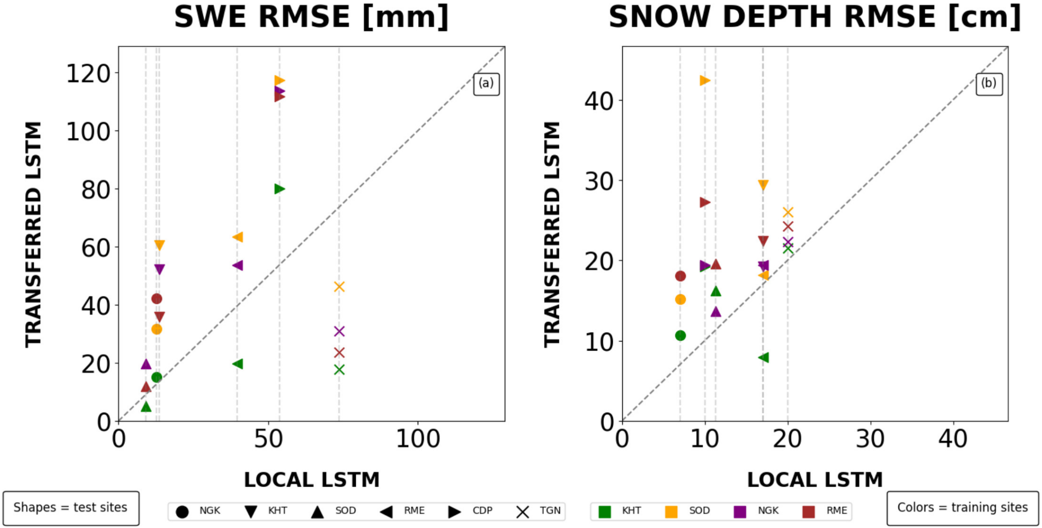


Figure: Transferability of LSTM-DA in Blandini et al. (2025): taking the LSTM-DA trained on one site and testing with S3M at other sites. The performance of local LSTM-DA (trained and tested on the same site) is on the x-axis while the corresponding performance of transferred LSTM-DA (trained and tested on different sites) is on the y-axis. Transfered LSTMs below the $1:1$ line are *better* than the local LSTMs. Shapes show the test sites (and local LSTM) and the colors show the training sites of transferred LSTMs. This is a lower bound on the performance of multi-site LSTMs that (Blandini et al., 2025) tested.

Assimilating to learn

Machine learning can be made uncertainty-aware using probabilistic methods. Methods from Bayesian data assimilation can help through e.g.:

- **Ensemble-based Bayesian deep learning** to infer carbon fluxes by fusing footprint analysis and eddy covariance (Pirk et al., 2024).
- **Filtering probabilistic rewards for reinforcement learning** to teach drones how to locate gas sources (van Hove et al., 2024).

Example 3: Deep ensemble carbon flux inference (Pirk et al., 2024)

ADVANCING EARTH AND SPACE SCIENCES

Geophysical Research Letters

RESEARCH LETTER

10.1029/2024GL109283

Special Collection:
Land-atmosphere coupling: measurement, modelling and analysis

Disaggregating the Carbon Exchange of Degrading Permafrost Peatlands Using Bayesian Deep Learning

Norbert Pirk[1], Kristoffer Aalstad[1], Erik Schytt Mannerfelt[1], François Clayer[2], Heleen de Wit[2], Casper T. Christiansen[3], Inge Althuizen[4], Hanna Lee[5], and Sebastian Westermann[1]

[1]Department of Geosciences, University of Oslo, Oslo, Norway, [2]Norwegian Institute for Water Research (NIVA), Oslo, Norway, [3]University of Copenhagen, Copenhagen, Denmark, [4]NORCE Norwegian Research Centre, Bergen, Norway, [5]Norwegian University of Science and Technology, Trondheim, Norway

Key Points:

- Eddy covariance fluxes are disaggregated for different surfaces using Bayesian neural networks to derive uncertainty-aware carbon balances
- While palsa areas have a near-zero annual methane balance, the fens and ponds that form upon palsa degradation emit large amounts of methane
- Fens compensate for methane emissions with strong annual CO_2 sinks, while ponds appear as strong, yet uncertain, CO_2 emission hotspots

Abstract Extensive regions in the permafrost zone are projected to become climatically unsuitable to sustain permafrost peatlands over the next century, suggesting transformations in these landscapes that can leave large amounts of permafrost carbon vulnerable to post-thaw decomposition. We present 3 years of eddy covariance measurements of CH_4 and CO_2 fluxes from the degrading permafrost peatland Iškoras in Northern Norway, which we disaggregate into separate fluxes of palsa, pond, and fen areas using information provided by the dynamic flux footprint in a novel ensemble-based Bayesian deep neural network framework. The 3-year mean CO_2-equivalent flux is estimated to be 106 $gCO_2\ m^{-2}\ yr^{-1}$ for palsas, 1,780 $gCO_2\ m^{-2}\ yr^{-1}$ for ponds, and $-31\ gCO_2\ m^{-2}\ yr^{-1}$ for fens, indicating that possible palsa degradation to thermokarst ponds would strengthen the local greenhouse gas forcing by a factor of about 17, while transformation into fens would slightly reduce the current local greenhouse gas forcing.

Supporting Information:

Figure: The study site in Pirk et al. (2024), namely the degrading permafrost peatland at Iškoras palsa mire near Karasjok in Finnmark, Norway. Since March 2019 an eddy covariance flux tower has measured spatially aggregated carbon fluxes at this site from a dynamic mixed footprint consisting of fens, ponds, and a palsas. The goal with this study was to disaggregate these carbon fluxes to provide dynamic and uncertainty-aware estimates of the contributions from each surface type.

Training a tailored Bayesian deep neural net *without* backprop

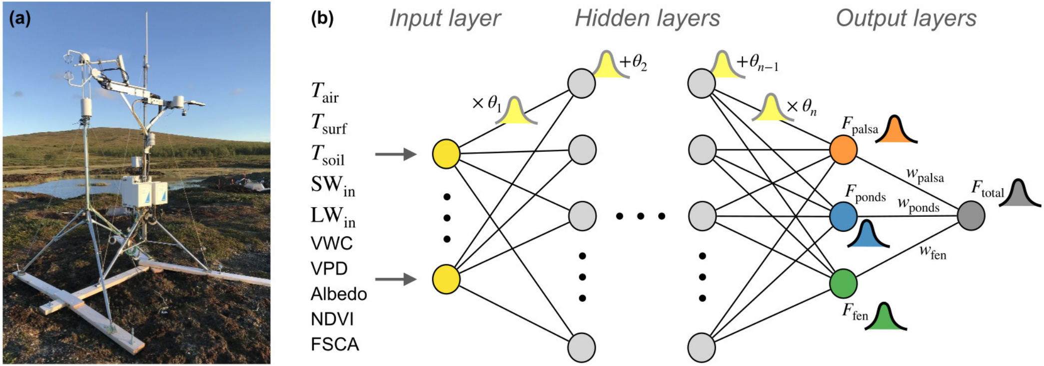


Figure: **Panel a)**: The flux tower with a sonic anemometer and gas analyzer. **Panel b)**: Architecture of the **'deep ensemble'** Bayesian feedforward neural networks (5 hidden layers) used to disaggregate CO_2 and methane fluxes at Iškoras. Here the $> 10^4$ network weights and biases $\boldsymbol{\theta}$ are probabilistic and inferred (i.e. trained) using a **gradient–free iterative ensemble Kalman smoother**. This is combined with a 'physics-informed' component with deterministic dynamic footprint weights $w_{\text{palsa}}(t)$, $w_{\text{ponds}}(t)$, $w_{\text{fen}}(t)$ to allow surface-specific fluxes in the pen-ultimate output layer to be inferred from the aggregated eddy covariance fluxes in the final output layer.

$$F_{\text{total}}(\mathbf{w}_s, \mathbf{x}, \boldsymbol{\theta}) = w_{\text{palsa}} \ F_{\text{palsa}}(\mathbf{x}, \boldsymbol{\theta}) + w_{\text{ponds}} \ F_{\text{ponds}}(\mathbf{x}, \boldsymbol{\theta}) + w_{\text{fen}} \ F_{\text{fen}}(\mathbf{x}, \boldsymbol{\theta})$$

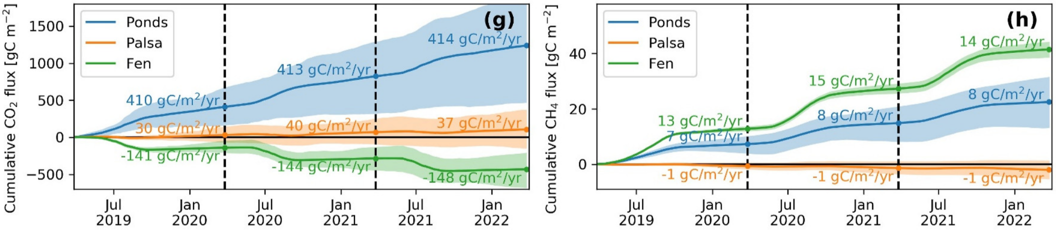


Figure: The carbon budget of the degrading permafrost peatland in terms of CO_2 (left) and methane (right) flux for the respective surface types as inferred from the deep ensemble. Note that flux data contains considerable gaps, so in addition to the surface type disaggregation this method also allows us to fill gaps to obtain continuous estimates of carbon flux for this ecosystem. Crucially, this probabilistic method also provides uncertainty quantification. The equation shows how the surface-specific fluxes are related to the aggregated flux measured in the dynamic eddy tower footprint through the footprint weights. Adapted from (Pirk et al., 2024)

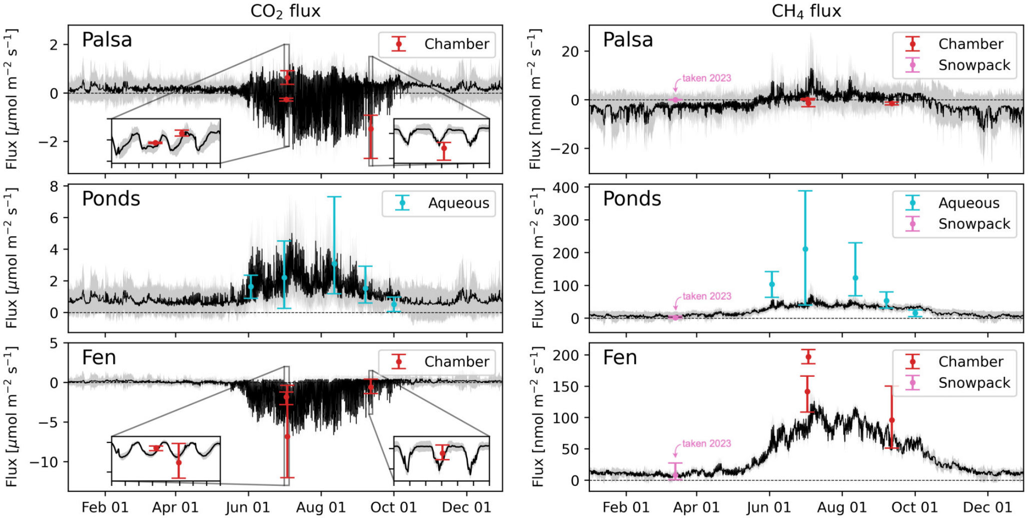


Figure: Qualitative validation of the deep ensemble carbon flux inference using independent measurements from the supplement of (Pirk et al., 2024). A more quantitative validation is available in the supplement.

Example 4: Bayesian filtering of probabilistic rewards for reinforcement learning in van Hove et al. (2024)building on van Hove et al. (2023).

The drone-based flux filtering problem is presented in van Hove et al. (2025).

Guiding drones by information gain

Alouette van Hove*[1], Kristoffer Aalstad[1], and Norbert Pirk[1]

[1]Department of Geosciences, University of Oslo, Norway
a.van.hove@geo.uio.no

Proceedings of the 5th Northern Lights Deep Learning Conference (NLDL), PMLR 233, 2024.

PMLR Proceedings of Machine Learning Research

$$\text{Plume Concentration} = \mathcal{G}\,(\text{Surface Flux}) + \epsilon \tag{5}$$

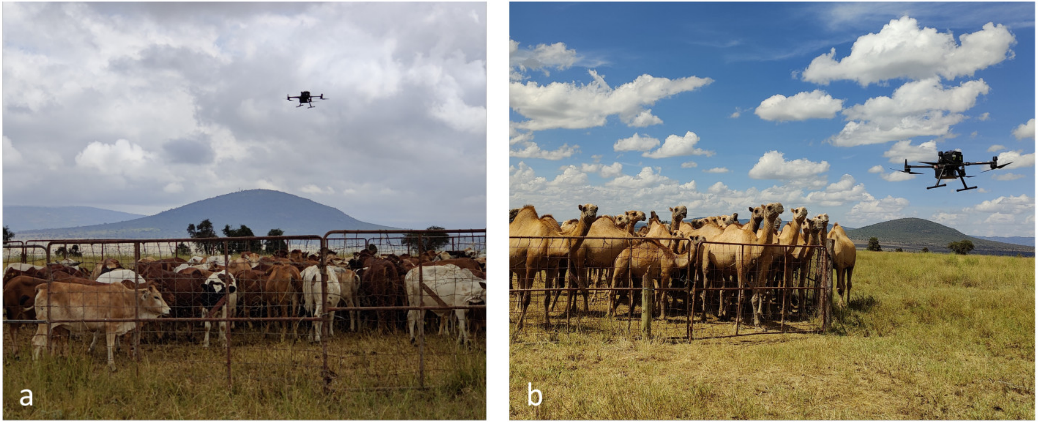
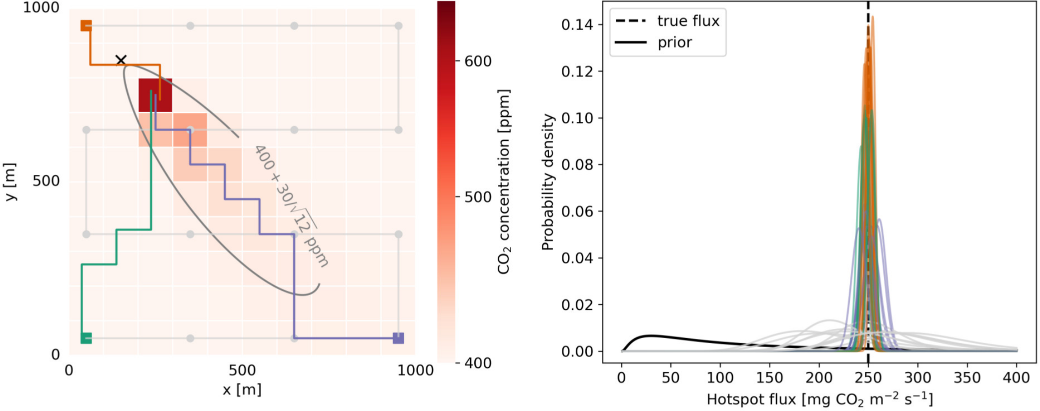


Figure: Setup for drone-based flux inversion using Bayesian filtering from van Hove et al. (2025). The task is to infer greenhouse gas (e.g. CO_2, methane) emissions from sources such as livestock by measuring concentrations in downwind plumes using drones and performing a model inversion to infer the flux.



(c) $r = -H$

Figure: In a pilot study van Hove et al. (2023)showed that tabular reinforcement learning (RL) can help identify more informative flight paths compared to 'expert' designs such as lawnmower patterns. Negative entropy (reductions in the uncertainty of the filtering posterior) turned out to be a promising reward for RL in line with information gain in the experimental design literature. **Left**: An expert flight path (gray) versus RL-trained flight paths (orange, green, blue). **Right**: The corresponding filtering distributions, note that these get much narrower and constrained around the true flux for the RL-trained paths. Here only the source strength (flux) was uncertain, the location and meteorological parameters are assumed known.

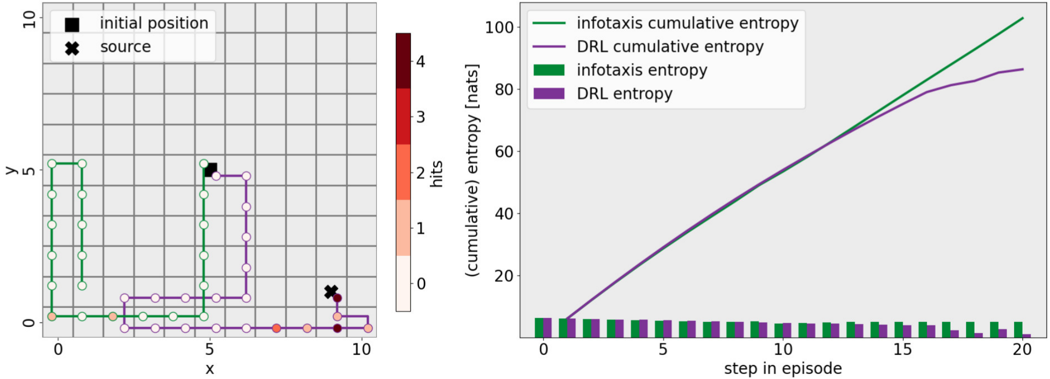


Figure: In the follow up study van Hove et al. (2024)this setup was extended to jointly infer both the source strength (flux) *location* while mmoving away from tabular RL to **deep RL** which is more applicable to higher dimensional problems. The 'offline' but far-sighted deep RL approach was compared to an 'online' infotaxis that relies on local information gradients and is thus myopic. Here too the negative entropy was used as the reward. **Left**: Example flight paths from infotaxis (green) and deep RL (purple). **Right**: The corresponding evolution of the entropy (negative reward) for these flight paths, note that a lower (cumulative) entropy is a sign of better performance in line with the deep RL-trained flight correctly identifying the source.

Table 1. Success rate of DRL and infotaxis evaluated across different flux values $\tilde{\phi}$ (training encompassed all source term scenarios).

	$\tilde{\phi} = 1$	$\tilde{\phi} = 2$	$\tilde{\phi} = 3$	$\tilde{\phi} = 4$	$\tilde{\phi} = 5$
DRL	40%	65%	80%	84%	85%
infotaxis	31%	49%	65%	72%	76%

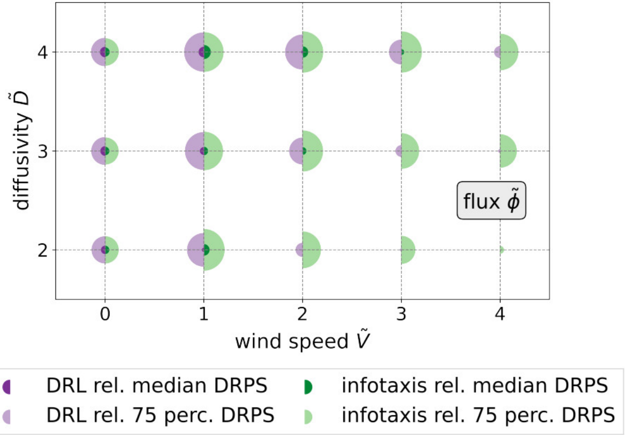


Figure: Validation of the experiments in van Hove et al. (2024). **Left**: Deep RL outperformed infotaxis for all (non-dimensional) flux magnitudes that we tested. **Right**: Deep RL always had less than or equal DRPS (a probablistic error score, lower is better) than the infotaxis strategy for all settings of the meteorological parameters. Future work plans to extend this to the multi-drone setting.

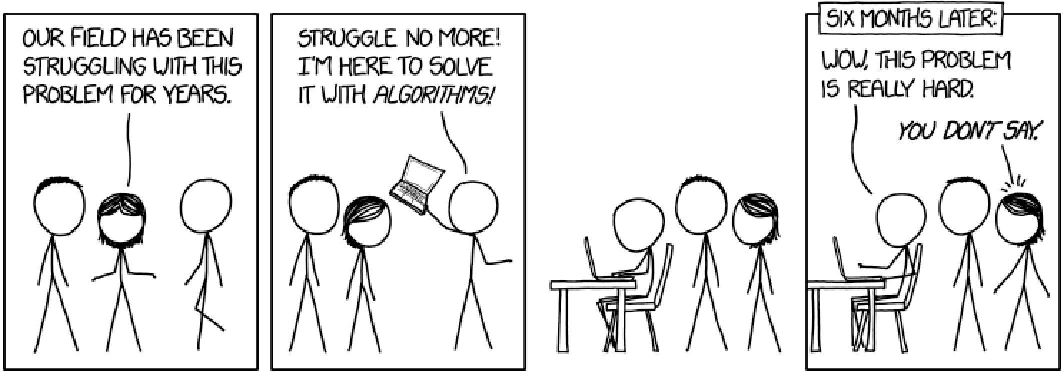


Figure: Here to help https://xkcd.com/1831/.

'*You cannot do inference without assumptions*' (MacKay, 2003)

Figure: Turtles all the way down (Imagen3) wikipedia.org/wiki/Turtles_all_the_way_down

References

Abarbanel, H. (2022). *The Statistical Physics of DA and ML*. CUP. doi: 10.1017/9781009024846.

Blandini, G. et al. (2025). Learning to filter: Snow data assimilation using a Long Short-Term Memory network. *EGUsphere [preprint in review at The Cryosphere]*. doi: 10.5194/egusphere-2025-423.

Deng, J. et al. (2009). ImageNet: A large-scale hierarchical image database. *CVPR*. doi: 10.1109/CVPR.2009.5206848.

Evensen, G. et al. (2022). *Data Assimilation Fundamentals*. Springer. doi: 10.1007/978-3-030-96709-3.

Geer, A. J. (2021). Learning ESMs from observations: ML or DA? *Phil. Trans. R. Soc. A*. doi: 10.1098/rsta.2020.0089.

Guidicelli, M. et al. (2024). A combined data assimilation and deep learning approach for continuous spatio-temporal SWE reconstruction from sparse ground tracks. *Journal of Hydrology X*. doi: 10.1016/j.hydroa.2024.100190.

Hersbach, H. et al. (2020). The ERA5 global reanalysis. *QJRMS*. doi: 10.1002/qj.3803.

Keetz, L. T. et al. (2024). Inferring parameters in a complex land surface model by combining data assimilation and machine learning. *ESSOA [preprint in review at JAMES]*. doi: 10.22541/essoar.172070530.05098424/v1.

Krizhevsky, A. et al. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25. url: proceedings.neurips.cc.

MacKay, D. J. C. (2003). Info. Theory, Inference, and Learning Algs. *CU*. url: inference.org.uk/itprnn/book.pdf.

Murphy, K. P. (2023). *Probabilistic Machine Learning: Advanced Topics*. MIT. url:probml.github.io/book2.

Neal, R. (1996). *Bayesian Learning for Neural Networks*. Springer. doi: 10.1007/978-1-4612-0745-0.

Papamarkou, T. et al. (2024). Position: Bayesian deep learning is needed in the age of large-scale AI. In *Proc. 41st International Conference on Machine Learning*. PMLR. url: pmlr/v235/papamarkou24b.html.

Pirk, N. et al. (2024). Disaggregating the Carbon Exchange of Degrading Permafrost Peatlands Using Bayesian Deep Learning. *Geophyiscal Research Letters*. doi: 10.1029/2024GL109283.

Price, I. et al. (2025). Probabilistic weather forecasting wtih machine learning. *Nature*. doi: 10.1038/s41586-024-08252-9.

Sanz-Alonso, D. et al. (2023). *Inverse Problems and Data Assimilation*. CUP. doi: 10.1017/9781009414319.

van Hove, A. et al. (2023). Using RL to improve drone-based inference of GHG fluxes. *NMI*. doi: 10.5617/nmi.9897.

van Hove, A. et al. (2024). Guiding drones by information gain. In *Proc. 5th NLDL*. PMLR. url: pmlr/v233/hove24a.html.

van Hove, A. et al. (2025). Inferring methane emissions from African livestock by fusing drone, tower, and satellite data. *EGUsphere [preprint in review at Biogeosciences]*. doi: 10.5194/egusphere-2024-3994.

Thanks to collaborators: J. Agerup (UiO), E. Alonso-González (IPE), L. Bertino (NERSC), G. Blandini (CIMA), W. Cao (UiO), J. Fiddes (SLF), B. Groenke (AWI), M. Guidicelli (EPFL), G. Guillet (UiO), R. Hock (UiO), L. Keetz (UiO), D. Kreynen (UiO), M. Mazzolini (UiO), N. Pirk (UiO), D. Treichler (UiO), A. van Hove (UiO), S. Westermann (UiO), C. Willmes (UiO), R. Yang (UiO), Y. Yılmaz (UiO)

Funding@UiO: GLACMASS (ERC), ACTIVATE (ERC), dScience PhD Fellowship (UiO, van Hove), PATCHES CCI Fellowship (ESA, me)

Funded by the European Union

European Research Council
Established by the European Commission

dScience
Centre for Computational and Data Science

esa

Optimization as crude ('plugin') Bayesian **inference** (Murphy, 2023)

$$\mathbf{y} = \mathcal{G}(\boldsymbol{\theta}) + \boldsymbol{\epsilon} \tag{6}$$

Gaussian prior: mean $\boldsymbol{\theta}$, covariance $\mathbf{C}_0$, and $c_0 = \det(2\pi\mathbf{C}_0)^{-1/2}$

$$p(\boldsymbol{\theta}) = \mathrm{N}(\boldsymbol{\theta}|\boldsymbol{\mu}_0, \boldsymbol{C}_0) = c_0 \exp\left(-\frac{1}{2}\left[\boldsymbol{\theta} - \boldsymbol{\mu}_\mathbf{0}\right]^\mathrm{T} \mathbf{C}_0^{-1} \left[\boldsymbol{\theta} - \boldsymbol{\mu}_\mathbf{0}\right]\right), \tag{7}$$

Gaussian likelihood: mean $\widehat{\mathbf{y}} = \mathcal{G}(\boldsymbol{\theta})$, covariance $\mathbf{R}$, $c_y = \det(2\pi\mathbf{R})^{-1/2}$

$$p(\mathbf{y}|\boldsymbol{\theta}) = \mathrm{N}(\mathbf{y}|\widehat{\mathbf{y}}, \mathbf{R}) = c_y \exp\left(-\frac{1}{2}\left[\mathbf{y} - \widehat{\mathbf{y}}\right]^\mathrm{T} \mathbf{R}^{-1} \left[\mathbf{y} - \widehat{\mathbf{y}}\right]\right), \tag{8}$$

The posterior is $p(\boldsymbol{\theta}|\mathbf{y}) = \exp(-\mathcal{J})/Z$ with $Z = p(\mathbf{y})$ where the **cost** function

$$\mathcal{J} = \frac{1}{2}\left[\boldsymbol{\theta} - \boldsymbol{\mu}_\mathbf{0}\right]^\mathrm{T} \mathbf{C}_0^{-1} \left[\boldsymbol{\theta} - \boldsymbol{\mu}_\mathbf{0}\right] + \frac{1}{2}\left[\mathbf{y} - \widehat{\mathbf{y}}\right]^\mathrm{T} \mathbf{R}^{-1} \left[\mathbf{y} - \widehat{\mathbf{y}}\right] - \log(c_y c_0) \tag{9}$$

is the negative log posterior. Minimizing $\mathcal{J}$ yields the maximum a posteriori (penalized maximum likelihood, regularized least squares) solution $\widehat{\boldsymbol{\theta}}$.