Supporting Information for "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], 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

Contents of this file

May 10, 2023, 11:58am

Text S1: Eddy covariance flux estimation

The EC instruments are installed at 2.8 m a.g.l. and sampled at a frequency of 20 Hz. The Li-7200 gas analyzer uses a 71 cm long heated (6 W) intake tube with a flow rate of 15 L min^{-1}. We processed the EC raw data to 30 minutes flux estimates following the conventional EC methodology (Gu et al., 2012) using EddyPro version 6.2.0. We extract turbulent fluctuations from block averages, use an anemometer tilt correction by double rotation, a constant time lag compensation, and a high- and low-pass filter correction following Moncrieff, Clement, Finnigan, and Meyers (2005) and Moncrieff et al. (1997), respectively. For quality control, we use statistical tests on the raw data proposed by Vickers and Mahrt (1997) and the flagging system proposed by Foken and Wichura (1996) to filter out flux estimates that are affected by instrument errors (e.g., rain or frost on the anemometer) or unfavorable micrometeorological conditions (e.g., non-stationarity or insufficient turbulent mixing). Following Vickers and Mahrt (1997), we estimate the number of spikes, drop-outs, as well as the absolute limits, amplitude resolution, skewness and kurtosis, and discontinuities for the pairs of raw data time series involved in the respective covariance-based flux estimates and discard data exceeding established thresholds. We also discard data with mean horizontal wind speeds below 1.5 m s^{-1} as well as all fluxes with quality flag 1 and 2 in the scheme of Foken and Wichura (1996).

The dynamic flux footprint function of each valid 30-minute flux is estimated based on friction velocity, wind direction, Obukhov length, and cross-wind variance, all obtained from EC measurements, as well as boundary layer height (linearly interpolated ERA5

atmospheric reanalysis data (Hersbach et al., 2020)), following the flux footprint model by Kljun, Calanca, Rotach, and Schmid (2015). For roughness length, we assume commonly used standard values of 10 cm for grasslands-like surfaces (i.e., for all snow-free areas) and 0.5 cm for snow-covered areas (Stull, 1988), and estimate a time series of roughness lengths as the area-weighted averages using the remotely sensed fractional snow-covered area (Aalstad, Westermann, and Bertino (2020), see Text S2 in this Supplement).

Text S2: Ancillary data for flux predictors

We use ancillary data from sensors on our measurements tower as well as from satellite remote sensing to quantify soil, surface, and atmospheric conditions during our flux measurements. Near-surface air temperature (T_{air}) is measured by a resistance temperature detector (PT-100) mounted in a radiation shield at 2 m a.g.l. Vapor pressure deficit (VPD) is derived from measurements of T_{air} and relative humidity (HMP155, Vaisala, Finland) mounted 2 m a.g.l. Soil temperature (T_{soil}) and soil volumetric water content (VWC) are measured at a depth of 8 cm (CS650, Campbell, USA). For incoming shortwave and longwave radiation ($\mathrm{SW_{in}}$ and $\mathrm{LW_{in}}$, respectively) we use a ventilated and heated radiometer (CNR4, Kipp&Zonen, Netherlands) mounted on a south-west-pointing boom at 2.8 m a.g.l. The same sensor is used to measure surface broadband albedo. Surface temperature (T_{surf}) is estimated using an infrared radiometer (SI-411, Apogee, USA) directed towards the surface approximately 2 m north of the tower. All these ancillary sensors are sampled every 10 s, filtered for corrupted measurements, and aggregated to 30 minute average values. Due to data logger problems, approximately 0.5% of the three year period lacks valid local measurements. For atmospheric and surface variables these short gaps are filled with estimates derived from a simple linear regression of the respective variable against its corresponding estimate from ERA5 atmospheric reanalysis data (Hersbach et al., 2020). Soil variables, which vary on longer timescales, are gap-filled with a simple linear interpolation of neighboring measurements.

The fractional snow covered area (FSCA) and Normalized Difference Vegetation Index (NDVI) are retrieved from multispectral satellite imagery covering the 500×500 m^2

area around the flux tower at a ground sampling distance of 10 m as described in Pirk et al. (2023). For this, we use surface reflectances from the Sentinel-2 satellite in 6 wavelength bands. The FSCA is retrieved using the spectral unmixing approach outlined in Aalstad et al. (2020). The NDVI, a widely used proxy for surface greenness, leaf area, and vegetation development, is calculated according to its usual definition (Jia et al., 2003). To avoid artifacts due to clouds in the reflectance data, we manually selected cloud-free scenes. This selection provided a total of 82 Sentinel-2 scenes for the entire study period, resulting in an average of just over 2 cloud-free scenes per month. The stack of cloud-free retrievals of FSCA and NDVI for each pixel were independently interpolated in time using Gaussian process regression (Rasmussen & Williams, 2005). We finally calculated and used spatial averages of FSCA and NDVI for the 500×500 m^2 area around the flux tower.

Text S3: Training the Bayesian neural networks

The BNN training is performed with an iterative ensemble Kalman method, namely the Ensemble Smoother with Multiple Data Assimilation (ES-MDA) (Emerick & Reynolds, 2013), a widely used scheme for computationally challenging Bayesian inference problems such as flux inversion with large eddy simulations (Pirk et al., 2022). In contrast to optimization, the Bayesian approach provides the necessarily probabilistic solutions to the ill-posed under-determined inverse problem of fitting the parameters of a nonlinear model (such as a neural network) to noisy data (Stuart, 2010). We use an ensemble size of $N_e = 100$ and $N_a = 128$ iterations with uniform observation error inflation ($\alpha = N_a$). To ensure that the matrix inversion in the ES-MDA is not poorly conditioned and to achieve a computational cost that is linear (rather than quadratic) in the size of the training data, we adopt ensemble subspace inversion (Evensen et al., 2022). To further improve computational efficiency and stability, each of the iterations only uses 10% of the training data (randomly sampled, without replacement), in a process referred to as mini-batching (Kovachki & Stuart, 2019; Murphy, 2022). Both the input and output data are standardized with a z-score transform following common practice in machine learning (Kovachki & Stuart, 2019; Murphy, 2022). This BNN training with flux data provides strong constrains for the model parameters $\boldsymbol{\theta}$ as exemplified by the prior and posterior distributions shown in Figure S3 in this Supplement.

To better represent the highly multi-modal posterior parameter distributions (Izmailov et al., 2021), we repeat the BNN training 100 times with different random seeds to capture local modes and combine these 100 local ensembles to form one global ensemble.

The global ensemble is subsequently marginalized to obtain the (approximate) posterior predictive distribution for the fluxes. This corresponds to a so-called deep ensembles approximation (Lakshminarayanan et al., 2017) of the posterior predictive distribution that is similar in spirit to the MultiSWAG method of Wilson and Izmailov (2020). A closely related iterative ensemble Kalman-based local updating method has been proposed by Zhang, Lin, Li, Wu, and Zeng (2018) to sample from multi-modal posterior distributions of hydrological model parameters. To the best of our knowledge, although extended (Singhal & Wu, 1988) and ensemble (Kovachki & Stuart, 2019; Lopez-Gomez et al., 2022) Kalman methods have been used for optimization-based training of neural networks, this is the first study to use ensemble Kalman methods to train an (approximately) Bayesian deep ensemble of neural networks.

Training our BNN model with an iterative ensemble Kalman method was found to be highly parallelizable and therefore computationally feasible, requiring less than one day on our server with 128 cores (two AMD EPYC 7742 64-core processors). The widely used gold-standard method for Bayesian inference, namely Markov Chain Monte Carlo, would require many more parameter samples while being inherently less parallelizable and would thus require a considerably longer wall clock time to train our BNN model (Kovachki & Stuart, 2019; Izmailov et al., 2021).

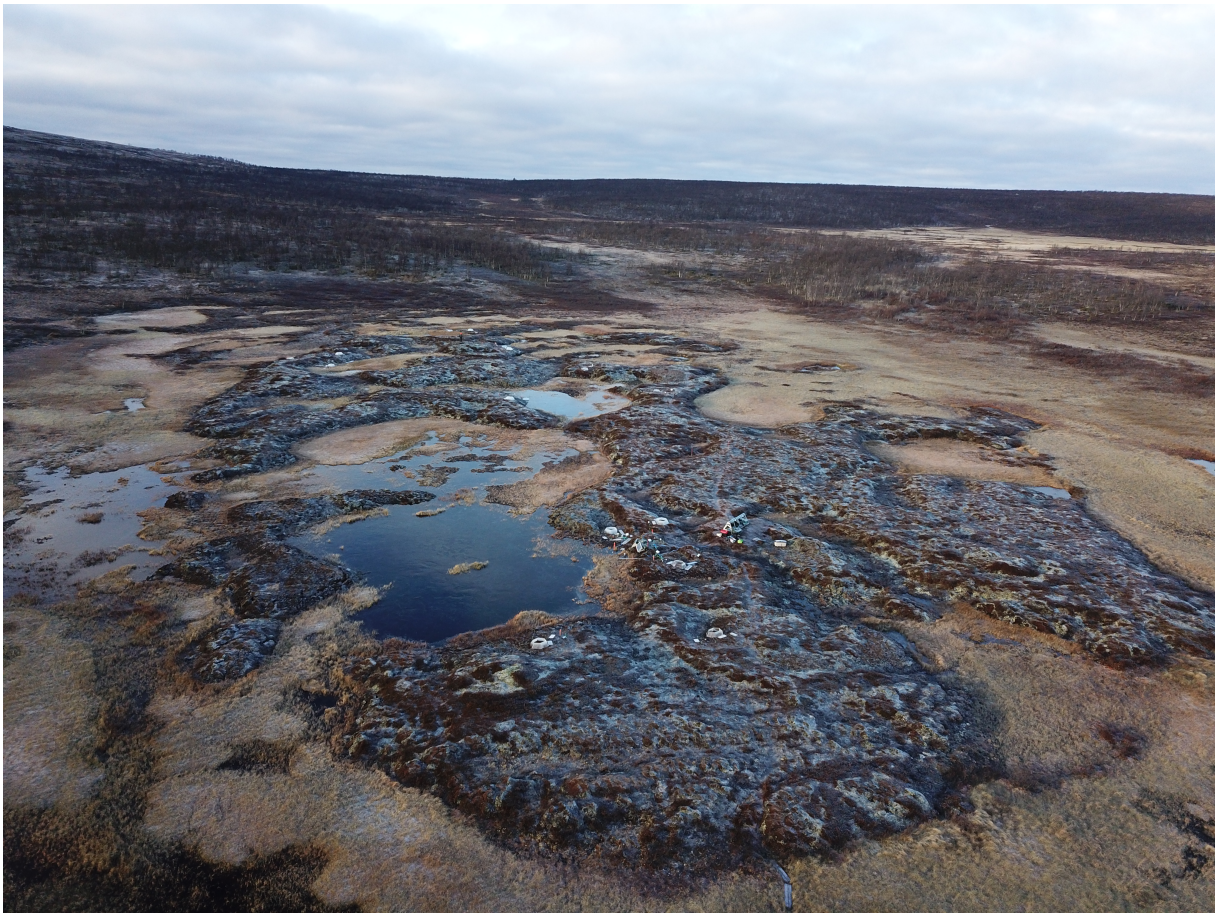

Figure S1. Overview aerial photograph of the Iškoras site taken on 20 October 2018, i.e., before the installation of the eddy covariance tower. The tower was set up with ample distance to the other installations seen in the picture (which are part of a different study).

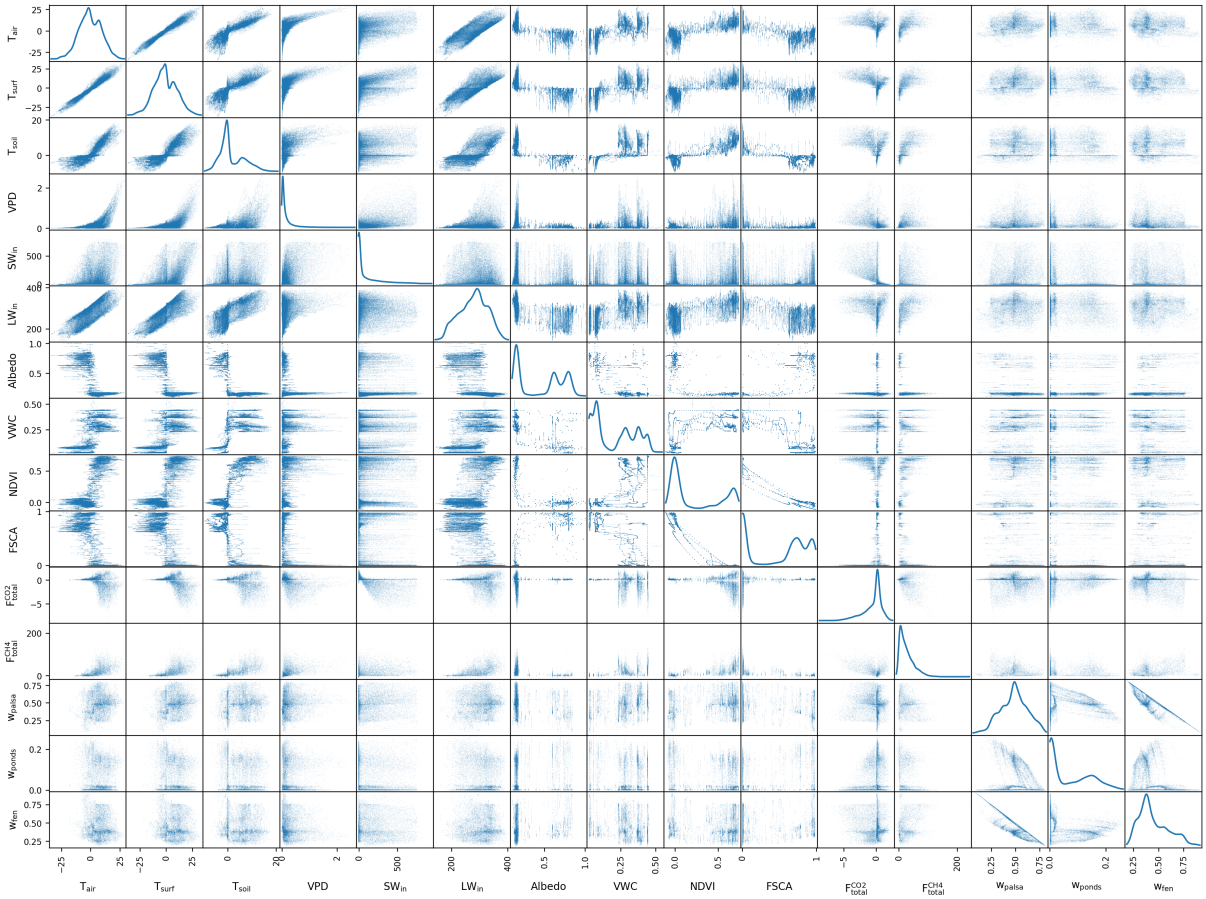

Figure S2. Scatter plot matrix of all predictors, fluxes, and footprint weights. The panels on the diagonal show kernel density estimates of the marginal distributions of the respective variables. The dataset is archived and available (Pirk, 2023).

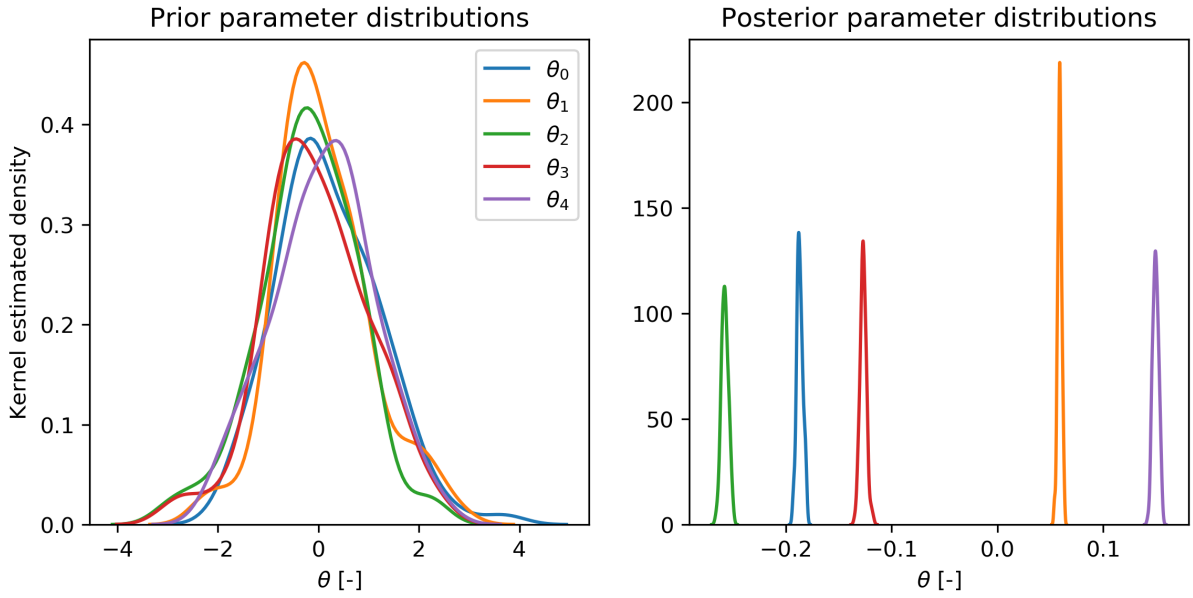

Figure S3. Kernel density estimates of the probability distributions of the first five (out of 11 919) parameters in a local BNN used to predict CH_4 fluxes. The priors are drawn from a standard normal distribution and the visual deviation of the KDE priors from the standard normal distribution are due to the finite ensemble size ($N_e = 100$). The posteriors are estimated using the ES-MDA scheme with $N_a = 128$ iterations.

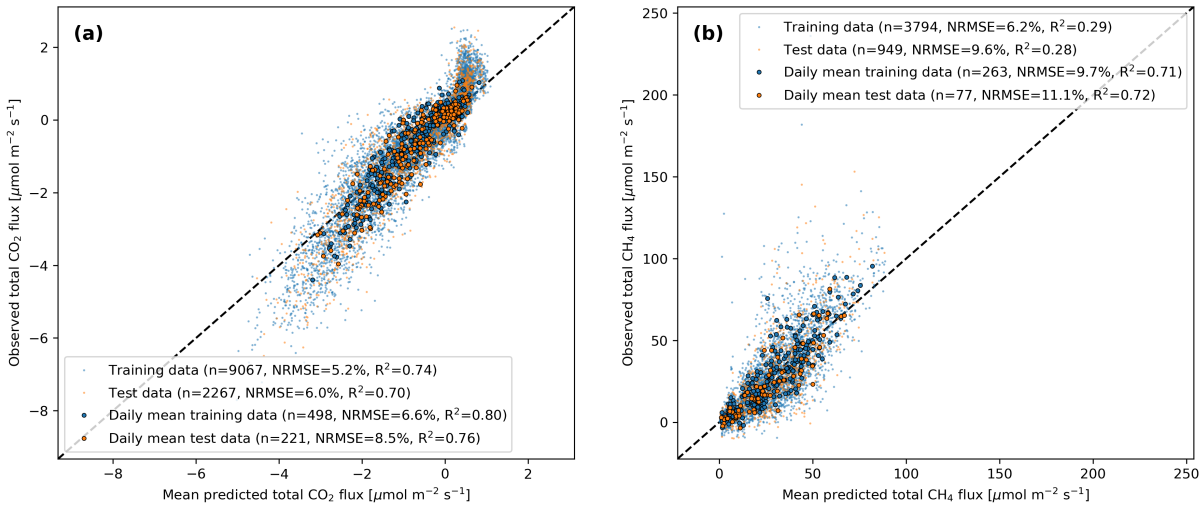

Figure S4. Train-test split evaluation of the BNN flux predictions of CO_2 (a) and CH_4 (b), both for 30-min (small dots) and daily average (larger circles) total fluxes. Statistics given in the figure legend include sample size (n), normalized root mean square error (NRMSE, normalized by the range), and the coefficient of determination (R^2).

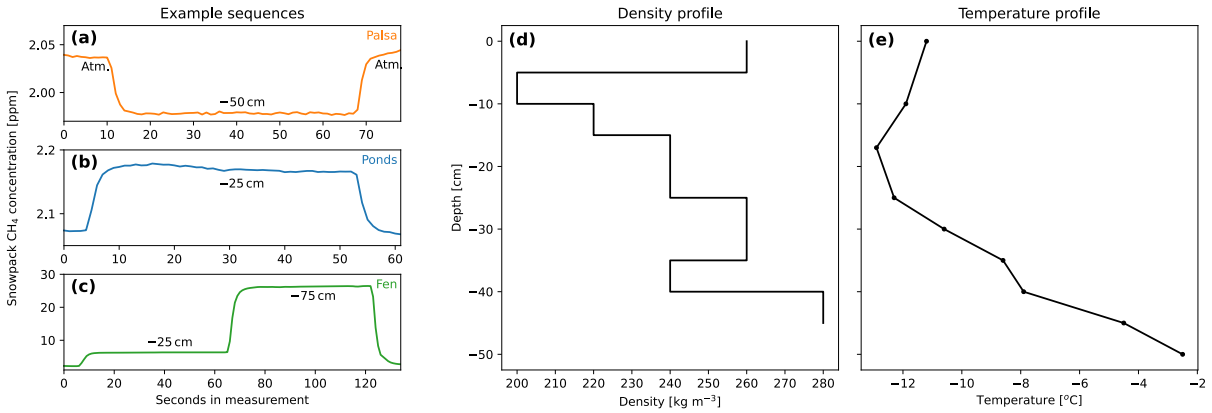

Figure S5. Data from a field campaign on March 14, 2023. Left: Example sequences of snowpack CH_4 concentration probing different depth levels for approximately one minute each, indicating CH_4 uptake on the ground surface at palsa areas (a), while ponds (b) and especially fen areas (c) feature CH_4 release (indicated by the increasing concentrations with depth). Center and right: Profiles of snow density (d) and temperature (e) obtained in a snow pit that were used for diffusive flux calculations.

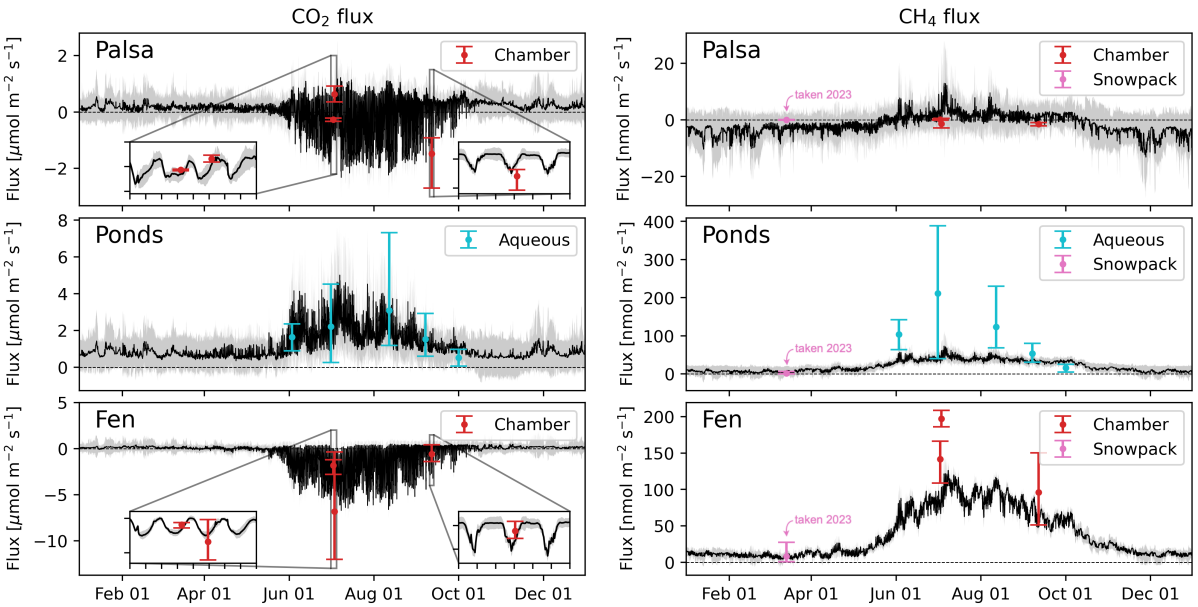

Figure S6. Flux time series obtained from the BNN disaggregation (black lines as ensemble means with shaded bands indicating the interquartile range of the ensemble predictions) in comparison with independent flux estimates from chamber, aqueous concentration, and snowpack gradient measurements. Error bars represent the range of spatial variability over the locations shown in Figure 1a in the main article. Inset sub-plots show a zoom-in view of the CO_2 fluxes for palsas and fens, which feature strong diurnal variability. Data is shown for 2021, but snowpack flux estimates were taken in 2023 (and assumed to be similar to 2021).

References

Aalstad, K., Westermann, S., & Bertino, L. (2020). Evaluating satellite retrieved fractional snow-covered area at a high-Arctic site using terrestrial photography. *Remote Sensing of Environment*, *239*, 111618. doi: 10.1016/j.rse.2019.111618

Emerick, A. A., & Reynolds, A. C. (2013). Ensemble smoother with multiple data assimilation. *Computers & Geosciences*, *55*, 3–15. doi: 10.1016/j.cageo.2012.03.011

Evensen, G., Vossepoel, F. C., & van Leeuwen, P. J. (2022). *Data Assimilation Fundamentals: A Unified Formulation of the State and Parameter Estimation Problem*. Springer. doi: 10.1007/978-3-030-96709-3

Foken, Th., & Wichura, B. (1996). Tools for quality assessment of surface-based flux measurements. *Agricultural and Forest Meteorology*, *78*(1-2), 83–105. doi: 10.1016/0168-1923(95)02248-1

Gu, L., Massman, W. J., Leuning, R., Pallardy, S. G., Meyers, T., Hanson, P. J., ... Yang, B. (2012). The fundamental equation of eddy covariance and its application in flux measurements. *Agricultural and Forest Meteorology*, *152*, 135–148. doi: 10.1016/j.agrformet.2011.09.014

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., ... Thépaut, J.-N. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999–2049. doi: 10.1002/qj.3803

Izmailov, P., Vikram, S., Hoffman, M. D., & Wilson, A. G. G. (2021). What are bayesian neural network posteriors really like? In M. Meila & T. Zhang (Eds.), *Proceedings of the 38th international conference on machine learning* (Vol. 139, pp. 4629–4640). PMLR.

Jia, G. J., Epstein, H. E., & Walker, D. A. (2003). Greening of arctic Alaska, 1981–2001. *Geophysical Research Letters*, *30*(20), 2003GL018268. doi: 10.1029/2003GL018268

Kljun, N., Calanca, P., Rotach, M. W., & Schmid, H. P. (2015). A simple two-dimensional parameterisation for Flux Footprint Prediction (FFP). *Geoscientific Model Development*, *8*(11), 3695–3713. doi: 10.5194/gmd-8-3695-2015

Kovachki, N. B., & Stuart, A. M. (2019). Ensemble Kalman inversion: A derivative-free technique for machine learning tasks. *Inverse Problems*, *35*(9), 095005. doi: 10.1088/1361-6420/ab1c3a

Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In I. Guyon et al. (Eds.), *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc.

Lopez-Gomez, I., Christopoulos, C., Langeland Ervik, H. L., Dunbar, O. R. A., Cohen, Y., & Schneider, T. (2022). Training Physics-Based Machine-Learning Parameterizations With Gradient-Free Ensemble Kalman Methods. *Journal of Advances in Modeling Earth Systems*, *14*(8). doi: 10.1029/2022MS003105

Moncrieff, J., Clement, R., Finnigan, J., & Meyers, T. (2005). Averaging, Detrending, and Filtering of Eddy Covariance Time Series. In X. Lee, W. Massman, & B. Law (Eds.), *Handbook of Micrometeorology* (Vol. 29, pp. 7–31). Dordrecht: Kluwer Academic Publishers. doi: 10.1007/1-4020-2265-4_2

Moncrieff, J., Massheder, J., de Bruin, H., Elbers, J., Friborg, T., Heusinkveld, B., ... Verhoef, A. (1997). A system to measure surface fluxes of momentum, sensible heat, water vapour and carbon dioxide. *Journal of Hydrology*, *188–189*, 589–611. doi: 10.1016/S0022-1694(96)

03194-0

Murphy, K. P. (2022). *Probabilistic machine learning: An introduction*. MIT Press.

Pirk, N. (2023). *Resources for "Disaggregating the carbon exchange of degrading permafrost peatlands using Bayesian deep learning" [Dataset]*. Zenodo. doi: 10.5281/zenodo.7913027

Pirk, N., Aalstad, K., Westermann, S., Vatne, A., van Hove, A., Tallaksen, L. M., ... Katul, G. (2022). Inferring surface energy fluxes using drone data assimilation in large eddy simulations. *Atmospheric Measurement Techniques*, *15*(24), 7293–7314. doi: 10.5194/amt-15-7293-2022

Pirk, N., Aalstad, K., Yilmaz, Y. A., Vatne, A., Popp, A. L., Horvath, P., ... Tallaksen, L. M. (2023). *Snow-vegetation-atmosphere interactions in alpine tundra* (Preprint). Biogeochemistry: Air - Land Exchange. doi: 10.5194/bg-2023-21

Rasmussen, C. E., & Williams, C. K. I. (2005). *Gaussian Processes for Machine Learning*. The MIT Press. doi: 10.7551/mitpress/3206.001.0001

Singhal, S., & Wu, L. (1988). Training multilayer perceptrons with the extended kalman algorithm. In D. Touretzky (Ed.), *Advances in neural information processing systems* (Vol. 1). Morgan-Kaufmann.

Stuart, A. M. (2010). Inverse problems: A Bayesian perspective. *Acta Numerica*, *19*, 451–559. doi: 10.1017/S0962492910000061

Stull, R. B. (1988). *An introduction to boundary layer meteorology*. Kluwer Academic Publishers.

Vickers, D., & Mahrt, L. (1997). Quality Control and Flux Sampling Problems for Tower and Aircraft Data. *Journal of Atmospheric and Oceanic Technology*, *14*(3), 512–526. doi:

10.1175/1520-0426(1997)014⟨0512:QCAFSP⟩2.0.CO;2

Wilson, A. G., & Izmailov, P. (2020). Bayesian deep learning and a probabilistic perspective of generalization. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (Vol. 33, pp. 4697–4708). Curran Associates, Inc.

Zhang, J., Lin, G., Li, W., Wu, L., & Zeng, L. (2018). An Iterative Local Updating Ensemble Smoother for Estimation and Uncertainty Assessment of Hydrologic Model Parameters With Multimodal Distributions. *Water Resources Research*, *54*(3), 1716–1733. doi: 10.1002/2017WR020906