Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Active-layer thickness prediction and regional differences across permafrost regions using a physics-residual machine-learning workflow

Seungwon Baek
Seoul National University
Seung Hee Kim
Korea Polar Research Institute
Yoon Taek Jung
Korea Polar Research Institute
Yongchae Cho
yc.cho@snu.ac.kr

Seoul National University

Article

Keywords:

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11671720/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Active-layer thickness prediction and regional differences across permafrost regions using a physics-residual machine-learning workflow

Seungwon Baek[1], Seung Hee Kim[2], Yoon Taek Jung[2], and Yongchae Cho[1,*]

[1]Seoul National University, Department of Energy Systems Engineering, Seoul, 08826, Republic of Korea
[2]Korea Polar Research Institute, Center of Remote Sensing and GIS, Incheon, 21990, Republic of Korea
[*]yc.cho@snu.ac.kr

ABSTRACT

Active-layer thickness (ALT) observations are sparse, and regional maps rely on the Stefan model or machine learning (ML). The Stefan model transfers to unobserved regions but uses one soil coefficient, whereas ML loses accuracy away from training sites. We propose a physics-residual workflow in which gradient boosting learns the residual of a locally recalibrated Stefan model from climate, terrain and soil covariates. Using 19,172 site-years (2010–2024) from Alaska, the Lena Delta and north-western Canada, each with the climate of its year, we compared it with the Stefan model, covariate-only ML (direct ML) and five products. Under spatial-block cross-validation it reduced the Stefan error by 3.6 cm (Lena Delta) and 3.7 cm (Canada). With the Lena Delta withheld, direct ML exceeded the Stefan error by 17.8 cm, whereas the physics-residual model stayed within 2.0 cm and gained 3.9 cm from 25 local labels. It matched two ML products in Alaska and Canada and surpassed them in the Lena Delta. It is stable without local training data and provides yearly 1 km maps with an area of applicability and prediction intervals.

Introduction

The active layer is the near-surface ground above permafrost that thaws each summer and refreezes each winter, and its thickness (ALT) is the maximum depth of seasonal thaw. ALT controls ground stability, near-surface hydrology and the amount of soil organic carbon exposed to decomposition [1–3]. It varies with the thermal conditions of each year, and this inter-annual variation is recorded by long-term programmes such as the Circumpolar Active Layer Monitoring (CALM) network [4, 5]. However, the observations are concentrated in a few field campaigns and long-term monitoring grids, many sites have been measured only once, and large parts of the permafrost region have no observations [6, 7]. ALT maps are therefore produced with models that extend the observations to unobserved grid cells and years, and the accuracy of the maps depends on the accuracy of the model away from the observed sites.

Regional ALT is mainly predicted with the Stefan model or with machine learning (ML). The Stefan model assumes that the thaw depth is proportional to the square root of the thawing degree-days (TDD) of the season and represents the thermal properties and moisture of the soil by one coefficient [4, 8, 9]. Because the model responds to the climate of each year and has a single parameter, its coefficient can be refitted with few observations and the model can be applied to regions without observations. However, one coefficient per region cannot represent the differences in soil and terrain within the region, and these differences remain as systematic residuals. ML models reproduce the observations at the training sites well from climate, terrain and soil variables, but they have to learn the response of ALT to climate from the training observations (labels). When they are validated with spatially separated data, their accuracy therefore decreases at locations unlike the training sites [10–13]. The published gridded ALT products follow one of these two approaches. The European Space Agency Climate Change Initiative (ESA CCI) product is computed with a ground heat-transfer model driven by satellite land-surface temperature [14, 15], and the products of Liu et al. [16] and Wei et al. [17] are ML products that relate site- and year-specific observations to yearly predictors. The products of Aalto et al. [18] and Ran et al. [19] provide the multi-year mean ALT. The ML products were validated by random or leave-one-site-out splits within their own site sets. Because the validation data differ, their reported errors (21.6 cm [16] to 60.5 cm [17]) cannot be compared directly, and none of them has been validated against observations of a stated year in a region without training sites.

Physics-guided learning, in which a physical model represents the dominant response and a learner corrects what the physical model does not explain, has improved accuracy and extrapolation in studies of lakes, rivers and soils [20–22]. For ALT, the Stefan model retained its accuracy at sites withheld from training in Alaska better than a random forest [23]. However,

it has not been tested whether a learner that corrects the Stefan model keeps the accuracy of the physical model in regions not used for training (transferability) while reaching the accuracy of ML. In addition, the conditions under which such a correction improves on the Stefan model, the number of local observations that a new region needs, and the accuracy of the correction relative to ML trained on the covariates alone (direct ML) and to the published products have not been compared on the same sites and years.

This study proposes a physics-residual workflow in which gradient boosting learns, from 23 climate, terrain and soil covariates, the residual of a Stefan model whose coefficient is recalibrated with local labels. Labels, climate inputs, predictions and scores are all defined for combinations of observation site and observation year (site-years), and the climate input is computed from ERA5-Land for that year and accumulated to the observation day [24]. Using 19,172 site-years of 2010–2024 from Alaska, the Lena Delta and north-western Canada, we compared the physics-residual model with the Stefan model and direct ML. We used five validation designs, from random splits to the hold-out of an entire region (region hold-out), and evaluated five published products on the same site-years. On this basis we analysed why the ranking of the methods changes with the validation design, under which conditions and with how many labels the residual learner improves on the Stefan model, and why the accuracy of the published products differs between regions. We also produced yearly 1 km ALT maps of the three regions for 2011–2024 with an area of applicability (the covariate range supported by the training data) and prediction intervals.

Data and methods

Study regions and active-layer observations

This study covers Alaska, the Lena Delta and north-western Canada (Canada (NW) in the figures and tables), where thaw-depth observations are dense and span several years (Fig. 1, Table 1). Observations for Alaska and Canada were taken from the Arctic-Boreal Vulnerability Experiment (ABoVE) soil moisture and active-layer thickness data set, version 2 [6], and from the CALM end-of-season grid values [5]. Observations for the Lena Delta were taken from the ALLena thaw-depth data set [7]. ABoVE thaw depths consist of probe measurements and ground-penetrating radar (GPR) measurements calibrated with probe measurements along the same transects (GPR values make up 70% of the Alaska site-years; Supplementary Methods). Measurements made before August fall well short of the seasonal maximum, so only ABoVE values of August–October were used, and they were restricted to values greater than 0 and less than 300 cm. Only ALLena values of August–September without the probe-limit flag (which marks measurements that reached the probe length) were used. The seasonal maximum of each location and year was taken as the label, and the mean day of its measurements as the observation day. CALM thaw depths are measured at the end of the thaw season, in late August or early September [4, 25], so CALM values without a date were assigned day of year 244 (1 September).

Observations were grouped by location, with coordinates rounded to 0.0001° (about 10 m; hereafter site), and by year, and they were averaged within each group together with their observation days. The measurement points of an ABoVE transect are therefore separate sites, and one 1 km cell can contain several sites. The three programmes differ in measurement protocol (probe and GPR transects in ABoVE, probe transects in ALLena and 1 km grid means of about 100 probe points in CALM) and in site selection, and this study used the values as reported. Only observations of 2010–2024, the years with year-specific climate inputs, were used, and a label of year y was paired only with the climate of year y. The labels comprise 15,540 site-years in Alaska (2010–2024), 2436 in the Lena Delta (2010–2022) and 1196 in Canada (2010–2019 and 2021–2024), distributed over 73, 16 and 26 blocks of 0.5° (Table 1). The sites lie in 321, 143 and 71 cells of 1 km (910, 244 and 144 cell-years). Up to 701 sites of one transect fall in one cell and year, so site-years are not independent samples at the 1 km scale; this was handled by the spatial-block validation. Sources, versions and licences are listed in Supplementary Table S1, the input variables in Supplementary Table S2, the years used in each analysis in Supplementary Table S3 and the processing steps in Supplementary Methods.

Year-specific climate input

Monthly ERA5-Land fields for January 2010 to December 2024 [24, 26] (2 m air temperature, top-layer soil temperature and snow water equivalent) were extracted at the nearest land grid cell of each site. The thawing degree-day sum [4, 9] of site c and year y is

$$\mathrm{TDD}_{c,y} = \sum_{m=1}^{12} \max(T_{c,y,m}, 0)\, d_{y,m}, \tag{1}$$

where $T_{c,y,m}$ is the monthly mean air temperature and $d_{y,m}$ the number of days in month m. Most observations were made before the end of the thaw season, so the model input is the TDD accumulated to the day of observation t,

$$\mathrm{TDD}^{\mathrm{obs}}_{c,y} = C_{c,y,k-1} + \frac{t - t_{k-1}}{d_{y,k}} \left(C_{c,y,k} - C_{c,y,k-1}\right), \tag{2}$$

Table 1. Labels, inputs and products. Top, ALT labels by region, 2010–2024 (blocks, 0.5° cells for spatial cross-validation; largest year, year with the most site-years). Bottom, inputs and products and the years used.

Region	Site-years	Sites	1 km cells	Blocks	Years	Sites ≥4 yr	Largest year
Alaska	15,540	13,253	321	73	2010–2024	203	2014 (51%)
Lena Delta	2436	2206	143	16	2010–2022	25	2016 (28%)
Canada (NW)	1196	730	71	26	2010–2024	48	2022 (37%)

Data	Domain	Resolution	Available	Used	Role
ERA5-Land monthly [24, 26]	global	0.1°	1950–	2010–2024	Climate input
Copernicus DEM [27]	global	30 m	static	–	Terrain
SoilGrids 2.0 [28]	global	250 m	static	–	Soil
MODIS [29, 30] MOD13Q1, MOD10A2	global	250, 500 m	2000–	2010–2024	Sensitivity
ESA CCI permafrost probability [31]	N. Hemisphere	1 km	1997–2021	mean 1997–2021	Map domain (≥ 50%)
Wei et al. [17, 32]	N. Hemisphere	1 km	2000–2024	2010–2024	Comparison
Liu et al. [16, 33]	N. Hemisphere	1 km	2003–2020	2010–2020	Comparison
ESA CCI ALT v4 [14]	N. Hemisphere	1 km	1997–2021	2010–2021	Comparison
ESA CCI ALT v5 [15]	N. Hemisphere	1 km	1997–2023	2010–2023	Comparison
Aalto et al. [18, 34]	circum-Arctic	30″	2000–2014	single map	Comparison
Yi and Kimball [35]	Alaska	1 km	2001–2015	2010–2015	Reference, not scored

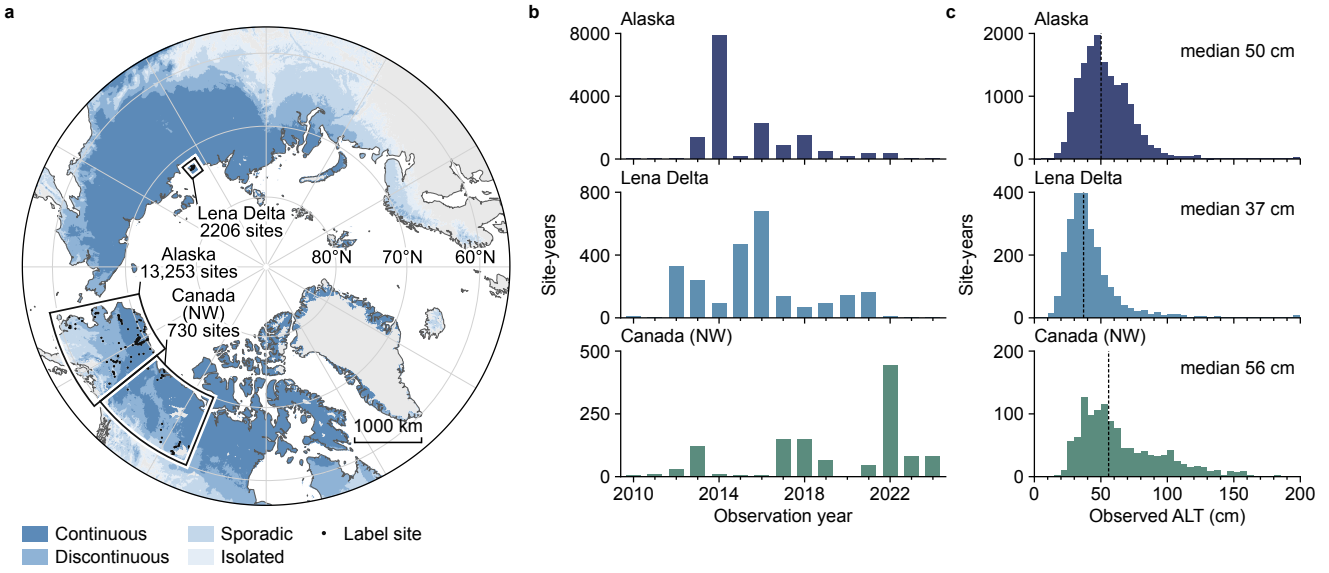

Figure 1. Study regions and observations. a, Permafrost zones (ESA CCI permafrost probability, mean of 1997–2021), label sites and the three study regions. **b**, Site-years per observation year. **c**, Distribution of the observed ALT and its median.

where $C_{c,y,k}$ is the cumulative TDD to the end of month k, k is the month containing t and t_{k-1} is the last day of the preceding month. Seven further climate variables were computed for each site and year. They are the mean annual air temperature, $\sqrt{\text{TDD}}$, the mean temperatures of the warmest and the coldest month, the freezing degree-days of the preceding winter (October of the previous year to April of the observation year), the mean top-layer soil temperature and the mean snow water equivalent of the hydrological year (October of the previous year to September of the observation year) (Supplementary Table S2 and Supplementary Methods). The series starts in January 2010, so the winter variables of 2010 were computed from January–April (freezing degree-days) and January–September (snow water equivalent) of 2010.

Two alternative inputs were also evaluated on the same site-years. One is the full-season TDD of the observation year. The other is the 2011–2024 mean of the yearly values (hereafter mean climate input), which corresponds to the multi-year inputs of existing products. Most observations were made in August, so the TDD accumulated to the observation day at the label sites is, in the median, 0.73, 0.81 and 0.89 times the full-season sum (Alaska, the Lena Delta and Canada). The yearly maps use the

TDD accumulated to day of year 244 (1 September in non-leap years), the end-of-season reference date of the CALM protocol. This date is 10 days later than the upper quartile of the pooled observation days (interquartile range 224–234) but within their full range (213–288). Map values therefore refer to the thaw depth reached by 1 September of the stated year.

Static covariates

Six terrain variables (elevation, slope, the sine and cosine of aspect, topographic position index and roughness) were computed from the 30 m Copernicus digital elevation model (DEM) [27] in a 33×33 pixel window (about 1 km) around each site. Nine soil variables (clay, sand, silt, bulk density, coarse fragments (>2 mm) and pH at 5–15 cm, and soil organic carbon at 0–5, 5–15 and 15–30 cm) were extracted from SoilGrids 2.0 at 250 m [28]. The main feature vector $\mathbf{x}$ consists of 23 variables (eight climate, six terrain and nine soil variables). Five alternative covariate sets were used in sensitivity tests: climate only; climate and terrain; the main set plus the MODIS summer normalised difference and enhanced vegetation indices (NDVI, EVI) and snow-off timing for 2010–2024 (MOD13Q1 and MOD10A2, version 6.1) [29,30]; the main set plus high-resolution soil organic carbon (NCSCD v2) [36,37], vegetation units (raster CAVM) [38] and land-cover fractions (ESA WorldCover 2021) [39]; and all variables including the climate of the preceding year. The ESA CCI permafrost probability (mean of 1997–2021) [31] was not used as a predictor. It was used only to define the map domain (probability of at least 50%) and the zones of Fig. 1a (continuous ≥90%, discontinuous 50–90%, sporadic 10–50%, isolated <10%).

Model formulation

The Stefan model describes the thaw depth as

$$\widehat{\mathrm{ALT}}_{\mathrm{S}} = E\,\sqrt{\mathrm{TDD}^{\mathrm{obs}}}, \tag{3}$$

where E (cm $(^{\circ}\mathrm{C\,d})^{-1/2}$) is a coefficient that summarises the thermal properties and moisture of the soil [8,9]. The relation holds well within a season at the Lena Delta monitoring sites, and it also holds among years at cells observed repeatedly, with more scatter (Supplementary Fig. S1). We used the Stefan model as the physical anchor. It has one parameter, which can be refitted from few observations and transferred to other regions, and its input is the TDD of each year. The Kudryavtsev model [40,41] with a scaling coefficient and a non-negative combination of the two models were evaluated as alternative anchors (Supplementary Table S4). Under spatial-block validation, the Kudryavtsev model was less accurate than the recalibrated Stefan model in Alaska (21.7 versus 18.6 cm) and more accurate in the Lena Delta (20.4 versus 21.0 cm) and Canada (25.4 versus 30.6 cm). Under region hold-out, however, it was less accurate in all three regions (25.1, 21.8 and 35.4 cm versus 19.4, 21.5 and 27.8 cm), so the Stefan model was kept as the anchor. Fitting this coefficient, often called the edaphic factor, to local observations is common practice in Stefan-based ALT mapping [4,42]. For n training labels y_i with $s_i = \sqrt{\mathrm{TDD}_i^{\mathrm{obs}}}$, the coefficient is recalibrated by least squares through the origin,

$$\hat{E} = \frac{\sum_{i=1}^{n} y_i\, s_i}{\sum_{i=1}^{n} s_i^2}. \tag{4}$$

The source-coefficient Stefan model uses $\hat{E}$ fitted on the other two regions, and the recalibrated Stefan model uses $\hat{E}$ fitted on the training labels of the target region. The physics-residual model takes the recalibrated Stefan prediction as the anchor and adds a learned correction. This is the residual form of physics-guided learning [20] (Fig. 2):

$$\widehat{\mathrm{ALT}} = \hat{E}\,s + \lambda\, g(\mathbf{x}), \qquad g = \arg\min_{f} \sum_{i=1}^{n} \left[\left(y_i - \hat{E}\,s_i\right) - f(\mathbf{x}_i) \right]^2, \tag{5}$$

where g is a CatBoost gradient-boosting regressor [43] (600 iterations, learning rate 0.03, depth 6, L2 regularisation 3) and $\lambda = 0.5$ shrinks the correction towards the physical anchor. CatBoost was used as the default learner because gradient-boosted trees need no feature scaling and are the usual reference model for tabular data of this size [44]. In a sensitivity test, the learner was replaced by a random forest, a multilayer perceptron, TabM and FT-Transformer [45–47], and no learner was the most accurate in all three regions (Supplementary Fig. S2). A larger λ increases the within-region gain (the reduction in root-mean-square error (RMSE) relative to the recalibrated Stefan model under spatial-block validation; −5.1 cm in the Lena Delta at $\lambda = 1$). It also increases the error under region hold-out (validation with one entire region withheld from training, defined below; +5.4 cm relative to the Stefan model in the Lena Delta at $\lambda = 1$). We therefore set $\lambda = 0.5$ as a compromise between the two. At λ = 0.25 and 1, the gains in the Lena Delta and Canada kept their sign, and the model remained more accurate than direct ML in the withheld Lena Delta and Alaska (Supplementary Table S4 and Supplementary Methods).

Three alternative structures, which differ in how they use physical information, were also compared. Direct ML uses no physical information, physics-input ML uses the physical predictions as input features, and the physics pseudo-label model

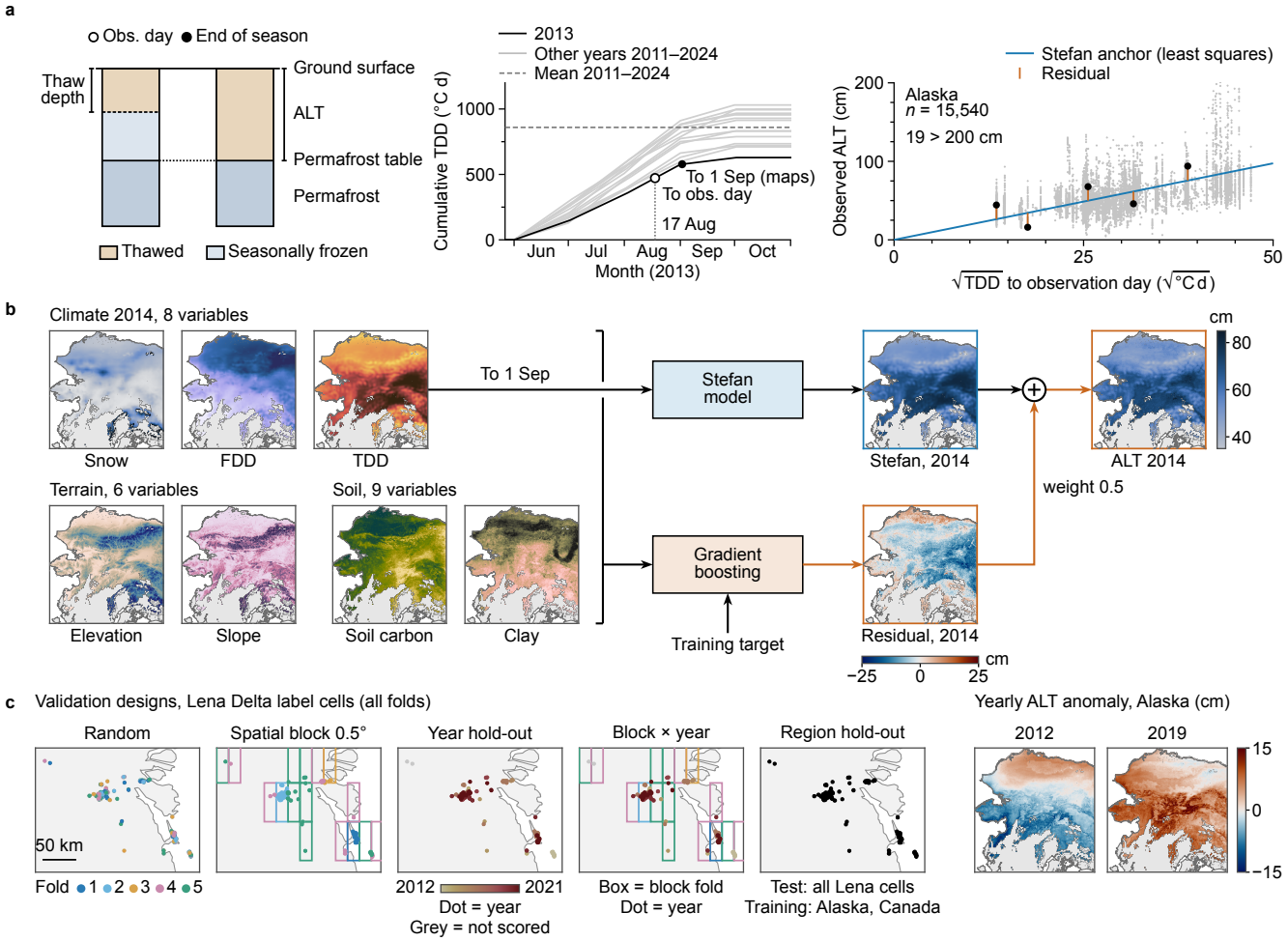

Figure 2. Site-year design, Stefan anchor, physics-residual model and validation designs. a, Thaw depth on the observation day and ALT at the end of the season; cumulative thawing degree-days of a Lena Delta cell in 2013, accumulated to the observation day (model input) and to 1 September (map input); observed ALT against $\sqrt{\text{TDD}_{\text{obs}}}$ for the Alaska site-years with the Stefan anchor. **b**, Physics-residual model (equation (5)), illustrated with the Alaska map of 2014. **c**, Validation designs on the Lena Delta label cells and ALT anomalies of the Alaska map in 2012 and 2019.

uses the physical prediction as an additional training target:

$$\widehat{\text{ALT}}_{\text{D}} = f_{\text{D}}(\mathbf{x}), \qquad \widehat{\text{ALT}}_{\text{I}} = f_{\text{I}}\left(\mathbf{x}, \hat{E}\,s, \widehat{\text{ALT}}_{\text{K}}\right),$$
$$\widehat{\text{ALT}}_{\text{A}} = f_{\text{A}}(\mathbf{x}) \text{ fitted to } \{(\mathbf{x}_i, y_i)\}_{i=1}^{n} \cup \{(\mathbf{x}_j, \hat{E}\,s_j)\}_{j=1}^{n}. \tag{6}$$

Direct ML (f_{D}) learns ALT from $\mathbf{x}$ alone. Physics-input ML (f_{I}) adds the recalibrated Stefan prediction and the Kudryavtsev [40] prediction $\widehat{\text{ALT}}_{\text{K}}$ as features. The physics pseudo-label model (f_{A}) is trained on the n training samples plus an equal number of samples j drawn from cells of the evaluation fold. The inputs of an added sample are the covariates of its cell, and its target is the Stefan prediction fitted on the training fold. All learned models were trained with three random seeds and their predictions averaged (one seed for the transfer models without target labels). For the physics-residual model and direct ML trained on all labels of each region, we computed SHAP values (SHapley Additive exPlanations) [48], which give the contribution of each input variable to a prediction.

Validation designs and uncertainty

All models were evaluated on out-of-fold predictions, that is, predictions of each site-year by a model trained without the test fold that contains it [10, 11]. Five validation designs were used. Random five-fold cross-validation assigns site-years to folds at random. Spatial-block five-fold cross-validation assigns 0.5° blocks to folds so that the folds contain similar numbers of

site-years, and it is the main design of this study. Year hold-out predicts each observation year with at least 20 site-years by models trained on the other years. Block-by-year hold-out excludes from training both the block fold and the year of the scored site-years (folds with at least 10 site-years). Region hold-out predicts an entire region by models trained on the other two regions, after training cells within 100 km of the test region are removed. Accuracy is expressed as

$$\mathrm{RMSE} = \left[\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2\right]^{1/2}, \qquad \Delta\mathrm{RMSE}_{A,B} = \mathrm{RMSE}_A - \mathrm{RMSE}_B, \tag{7}$$

where both terms of ΔRMSE are computed on the same site-years, and a negative value means that method A has the smaller error. Uncertainty was quantified with a paired block bootstrap [10,49]. The 0.5° blocks were resampled with replacement $B = 2000$ times (1000 times for the label-number experiments and for part of the covariate-set and learner sensitivity analyses), and ΔRMSE was recomputed from all site-years of the resampled blocks. The 2.5th and 97.5th percentiles of the replicates define the 95% interval,

$$\mathrm{CI}_{95} = \left[Q_{0.025}\left(\Delta\mathrm{RMSE}^{(b)}\right),\, Q_{0.975}\left(\Delta\mathrm{RMSE}^{(b)}\right)\right], \quad b = 1,\ldots,B. \tag{8}$$

A difference is described as distinguishable from zero when its 95% interval excludes zero. One 1 km cell can contain several sites (median number of sites per cell: 5 in Alaska, 3 in the Lena Delta and 5 in Canada), so the main comparisons were also scored on cell-year means of the observations and predictions (Supplementary Table S5). The conditions under which the gain appears were analysed in bins that divide each of eight covariates into thirds within each region (hereafter terciles). The eight covariates are the full-season TDD of the observation year, freezing degree-days, snow water equivalent, soil organic carbon, clay, slope, elevation and summer NDVI. Terciles with fewer than 30 site-years or fewer than five blocks were not evaluated. Year-wise RMSE was computed for years with at least 30 site-years. In box plots, the box shows the interquartile range and the median, and the whiskers show the 5th–95th percentile range.

Label-number and transfer experiments

In the within-region experiment, the training labels of each spatial-block fold were reduced to n = 25, 50, 100, 250, 500, 1000, 2500 or all labels (three random draws). The Stefan coefficient was recalibrated on the same n labels, and the complete test fold was scored. In the transfer experiment, models were trained on all labels of the other two regions plus n = 0, 25, 50, 100, 250 or 500 labels drawn from the training folds of the target region, and the spatial-block folds of the target region were scored. The Stefan coefficient was refitted on the n target labels. The residual learner was trained on the source-region samples, with residuals computed from the source coefficient, together with the target-region samples, with residuals computed from the refitted coefficient:

$$\begin{aligned} \widehat{\mathrm{ALT}}_\mathrm{T} &= \hat{E}_\mathrm{T}\, s + \lambda\, g_\mathrm{ST}(\mathbf{x}), \\ g_\mathrm{ST} &= \arg\min_f \left[\sum_{i\in\mathrm{S}}\left(y_i - \hat{E}_\mathrm{S}\, s_i - f(\mathbf{x}_i)\right)^2 + \sum_{i\in\mathrm{T}}\left(y_i - \hat{E}_\mathrm{T}\, s_i - f(\mathbf{x}_i)\right)^2\right], \end{aligned} \tag{9}$$

where S denotes the source-region samples, T the n target-region samples, $\hat{E}_\mathrm{S}$ the coefficient fitted on the source samples and $\hat{E}_\mathrm{T}$ the coefficient refitted on the target samples ($\hat{E}_\mathrm{T} = \hat{E}_\mathrm{S}$ for $n < 3$). In this experiment, source-region samples within 100 km of the target region were not removed.

Comparison with published products

Six publicly archived products were downloaded and sampled at the label sites (nearest 1 km pixel; Supplementary Methods and Supplementary Table S1). The Northern Hemisphere products are those of Wei et al. [17,32] (1 km, yearly 2000–2024; data set version 2), Liu et al. [16,33] (1 km, yearly 2003–2020), ESA CCI Permafrost ALT v4 (1997–2021) and v5 (1997–2023) [14,15] and Aalto et al. [18,34] (one value per cell for the 2000–2014 baseline). The Alaska-only product of Yi and Kimball [35] (2001–2015) covers one region and ends in 2015, so it was used only as a reference and was not scored. The multi-year mean of Ran et al. [19] (2000–2016) is distributed through a repository that requires registration and was not scored; Aalto et al. represents this type of product. Regional remote-sensing products do not cover all three regions and the years of the labels. The airborne radar product of the Arctic-Boreal Vulnerability Experiment [50] covers only flight lines in Alaska and north-western Canada in 2017, and the Lena Delta has no gridded ALT product.

For the yearly products, the value of the observation year was used, so that products and models were compared at the same site and in the same year. The comparison was made on common site-years, at which Wei et al., Liu et al. and ESA CCI v4 all have a value for the observation year (2010–2020). ESA CCI v5 and Aalto et al. were scored on the common site-years where they have a value, and the year-wise comparison used all site-years of a year with a value of Wei et al. Wei et al. published the

coordinates of their training sites, so the comparison was repeated after excluding label cells within 25 km of these sites. Liu et al. trained a random forest on CALM and on the ABoVE soil moisture and active-layer data set [16], an earlier release of the Alaska and Canada observations used in this study. Their product values at these sites are therefore not independent of our labels, and the training sites of Wei et al. overlap with our Alaska sites in the same way. To test whether the residual learner can also correct a product, the value p of the Wei et al. product for the observation year was recalibrated on the training fold,

$$\hat{p} = a + b\,p, \tag{10}$$

and the residual $y - \hat{p}$ was learned with the same covariates, learner, λ and folds as in equation (5) (Supplementary Methods).

Yearly maps, area of applicability and prediction intervals

In each of the three regions, the recalibrated Stefan model and the physics-residual model were trained on all site-years of the region. The map domain consists of the 1 km cells with an ESA CCI permafrost probability of at least 50%. This corresponds to the continuous and discontinuous permafrost zones in the classification of Obu et al. [51], where an active layer exists above permafrost and where the labels lie. Cells that are water (mean occurrence of at least 50% in the JRC Global Surface Water data set, v1.4 [52]) or lack an input were excluded (858,517, 38,086 and 1,151,169 cells remain). The two models were applied to these cells with the climate of each year from 2011 to 2024 (TDD accumulated to 1 September). The area of applicability followed the method of Meyer and Pebesma [53]. With $\mathbf{z}$ the standardised 23-dimensional feature vector (with the 2011–2024 mean climate), the dissimilarity of map cell k is $d_k = \min_j \|\mathbf{z}_k - \mathbf{z}_j\|_2$ over all training samples j, and the cell lies inside the area of applicability when

$$d_k \le \tau, \qquad \tau = Q_{0.95}\left\{ \min_{j:\, b(j)\neq b(i)} \|\mathbf{z}_i - \mathbf{z}_j\|_2 \;:\; i \in \text{training samples} \right\}, \tag{11}$$

where $b(\cdot)$ denotes the spatial block. The anomaly of cell k in year y is $a_{k,y} = \widehat{\mathrm{ALT}}_{k,y} - \overline{\widehat{\mathrm{ALT}}_k}$, where the overbar denotes the 2011–2024 mean. The regional mean anomaly is the mean of $a_{k,y}$ over all mapped cells, and the trend of each cell is the least-squares slope

$$\beta_k = \frac{\sum_y (y - \bar{y})\, a_{k,y}}{\sum_y (y - \bar{y})^2}. \tag{12}$$

Trends are given in cm per decade, as in previous yearly ALT products [16, 35]. Because the Stefan term changes with $\sqrt{\mathrm{TDD}}$ of each year, the regional mean anomaly is close to a linear function of the regional mean TDD anomaly, and departures from this relation come from the learned correction. The 90% prediction interval of a map value is

$$\left[\widehat{\mathrm{ALT}}_{k,y} + Q_{0.05}(e),\ \widehat{\mathrm{ALT}}_{k,y} + Q_{0.95}(e) \right], \tag{13}$$

where e denotes the spatial-block out-of-fold residuals of the region. This is a split-conformal interval [54] with a constant width within a region, and the uncertainty of individual cells is indicated by the dissimilarity d_k of equation (11). For display, the 1 km values were interpolated bilinearly onto a 250 m (Lena Delta) or 500 m (Alaska, Canada) hillshade grid.

For each region, we selected the year with the most observed 1 km cells that have a value in our map and in both products (2013 in Alaska and Canada, 2016 in the Lena Delta), and we compared the map of that year with the two ML products. Like our maps, these two products are yearly 1 km maps predicted from yearly covariates and trained on in situ ALT, and they were the most accurate products in Alaska and Canada in our comparison. The comparison used the product values of the same year in those cells, the cell means of the observations and the spatial-block out-of-fold predictions of our model. The product maps are displayed only on the cells of our map, and the three maps of a region share one logarithmic colour scale. The colour range is the 1st–99th percentile range of the pooled values of the three maps inside the area of applicability, widened by 10 cm at both ends, so that no single map sets the scale and the variation within each map remains visible.

Computation

Models were trained with fixed random seeds on a 96-core CPU node, and the neural learners on one NVIDIA A100 GPU (Supplementary Methods).

Code availability

The code used for data processing, model fitting, evaluation and figures is available from the corresponding author on reasonable request.

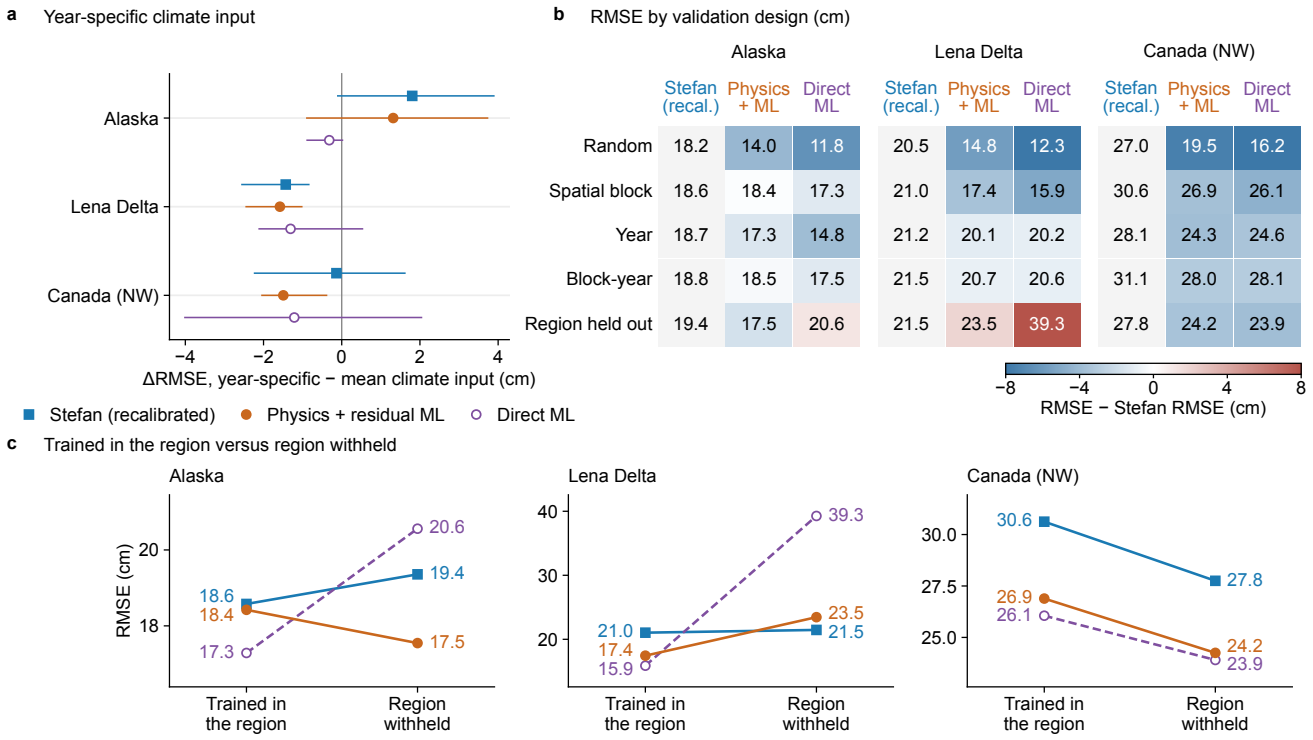

Figure 3. Climate input, validation design and the in-region versus out-of-region contrast. a, Change in spatial-block RMSE when the mean climate input of 2011–2024 is replaced by that of the observation year. **b**, RMSE of three methods under five validation designs; colour, difference from the recalibrated Stefan model. **c**, RMSE with the region's own labels and with the region withheld. Error bars, 95% block-bootstrap intervals.

Results

Climate input, validation design and the in-region versus out-of-region contrast

Using the climate of the observation year instead of the 2011–2024 mean climate reduced the spatial-block RMSE. The change in RMSE was −1.4 [−2.6, −0.8] cm for the recalibrated Stefan model in the Lena Delta, and −1.6 [−2.4, −1.0] cm (Lena Delta) and −1.5 [−2.0, −0.4] cm (Canada) for the physics-residual model. In Alaska the change could not be distinguished from zero (+1.3 [−0.9, +3.7] cm; Fig. 3a). Under spatial-block cross-validation with the climate of the observation year, the physics-residual model reduced the RMSE of the recalibrated Stefan model from 21.0 to 17.4 cm in the Lena Delta (ΔRMSE −3.6 [−7.8, −0.8] cm) and from 30.6 to 26.9 cm in Canada (−3.7 [−4.8, −0.8] cm). In Alaska the RMSEs of the two models were 18.6 and 18.4 cm, the smallest difference among the three regions (−0.2 [−0.6, +0.1] cm; Fig. 3b). The recalibrated Stefan model uses one coefficient per region, so its predictions spanned a narrow range, and the residual learner widened this range in the Lena Delta and Canada (Supplementary Fig. S3). With all labels of a region, direct ML had a slightly lower site-year RMSE than the physics-residual model (Alaska 17.3 versus 18.4, Lena Delta 15.9 versus 17.4 and Canada 26.1 versus 26.9 cm). When scored on 1 km cell-year means, the unit of the maps, this difference disappeared in Alaska (24.7 versus 24.0 cm; Supplementary Table S5).

The ranking of the methods depended on the validation design (Fig. 3b, Supplementary Table S4). Direct ML was the most accurate when test sites lay close to training sites (random cross-validation RMSE 11.8 cm in Alaska). Under spatial-block cross-validation, the error of direct ML increased by 5.4 cm in Alaska, 3.6 cm in the Lena Delta and 9.9 cm in Canada. When an entire region was withheld from training, the RMSE of direct ML increased from 15.9 to 39.3 cm in the Lena Delta (mean overestimation +33.6 cm) and from 17.3 to 20.6 cm in Alaska (Fig. 3c). Under the same condition, the physics-residual model stayed within 2.0 cm of the Stefan model with the coefficient of the other regions in the Lena Delta, and it had smaller errors than this model in Alaska (−1.8 cm) and Canada (−3.5 [−5.1, −0.1] cm). Direct ML transferred well only to Canada, whose inputs lie largely within the range of the training regions (site-years with all 23 inputs inside the 1st–99th percentile range of the other two regions: 42% in Canada, 0% in the Lena Delta and 0% in Alaska). The accuracy of the physics-residual model changed little with the region and the validation design. In all 15 combinations (three regions × five designs), its RMSE was

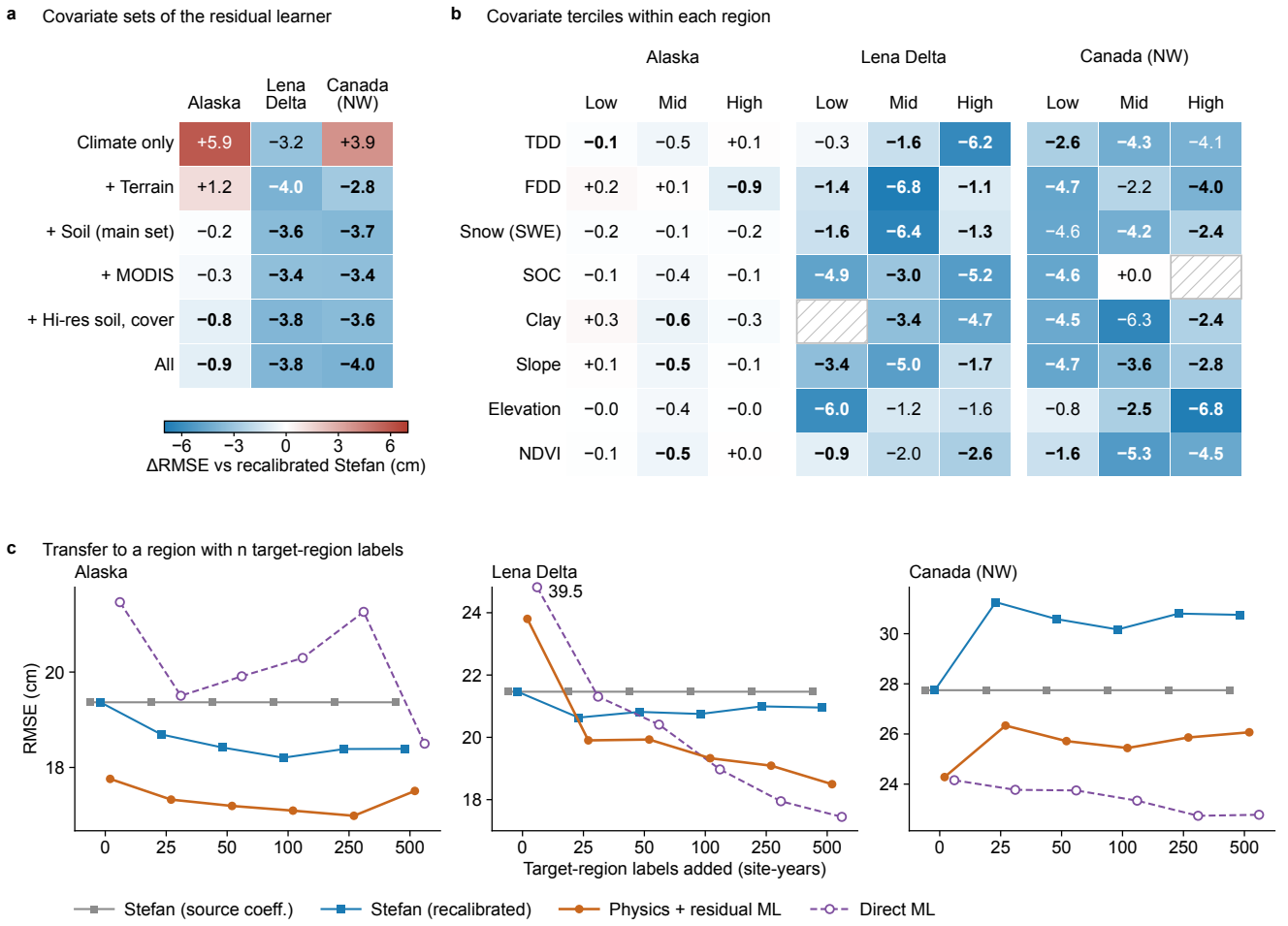

Figure 4. Conditions of the physics-residual gain and transfer to a region with few labels. a, Difference in spatial-block RMSE between the physics-residual model and the recalibrated Stefan model for six covariate sets. **b**, The same difference within regional terciles of eight covariates (hatched, fewer than 30 site-years or five blocks). **c**, RMSE of models trained on the other two regions and *n* labels of the target region. Bold, 95% interval excluding zero.

within 3.3 cm of the most accurate of the three methods. In contrast, the RMSEs of direct ML and the recalibrated Stefan model exceeded that of the most accurate one by up to 17.8 and 10.8 cm, respectively.

Conditions of the physics-residual gain

The gain of the residual learner depended on the covariates used (Fig. 4a). With climate covariates only, the residual model did not improve on the Stefan model in Alaska (+5.9 [−0.3, +14.2] cm) or Canada (+3.9 [−5.1, +10.3] cm). Adding terrain and soil covariates produced the gains of the previous section in the Lena Delta and Canada, but not in Alaska. Adding MODIS vegetation and snow timing or high-resolution soil and land-cover classes changed the RMSE difference in the Lena Delta and Canada by less than 0.5 cm. In the regional terciles of the eight covariates, the gain appeared in most terciles of the Lena Delta (terciles with 95% intervals excluding zero, 19 of 23) and Canada (17 of 23). The gain was larger in terciles where the Stefan model had larger errors (correlation of ΔRMSE with the Stefan RMSE across terciles −0.99 and −0.79; Fig. 4b), and it was −6.2 [−12.0, −0.7] cm in the warmest TDD tercile of the Lena Delta. In Alaska, the change did not exceed 0.9 cm in any tercile. Terrain variables made up 50% of the mean absolute SHAP value of the residual model in the Lena Delta, and soil variables 39% in Canada (Supplementary Fig. S4). The physics pseudo-label structure gave results close to the Stefan model, and the physics-input structure close to direct ML. Replacing CatBoost with the four other learners did not change the sign of the gain in the Lena Delta and Canada (Supplementary Fig. S2).

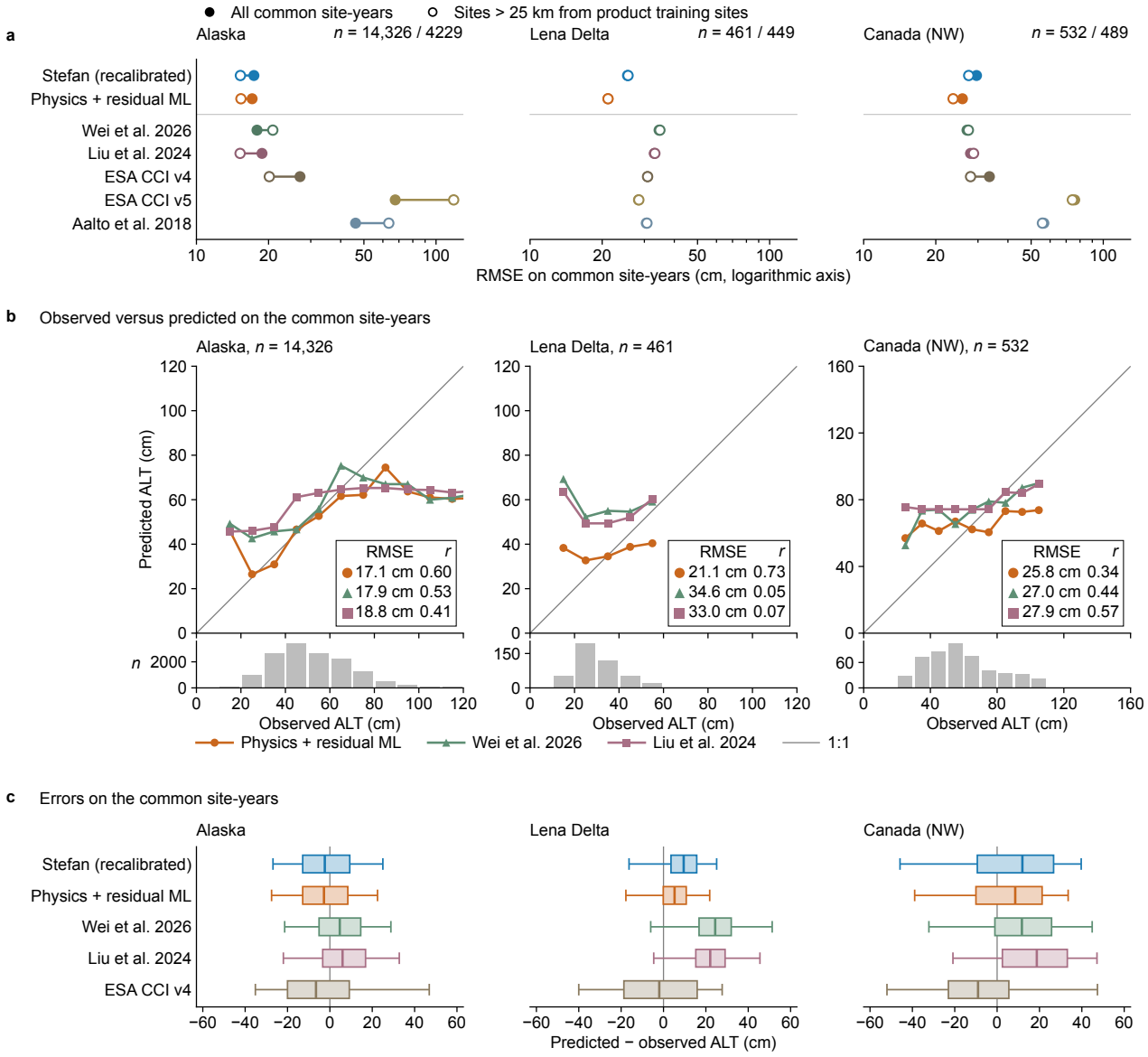

Figure 5. Comparison with published products on the same sites and years. a, RMSE on the common site-years of 2010–2020, for all sites and for sites more than 25 km from the training sites of Wei et al. **b**, Median prediction per 10 cm bin of observed ALT and number of site-years per bin. **c**, Prediction errors on the same site-years (boxes, quartiles and median; whiskers, 5th–95th percentiles).

Number of labels within a region and transfer to a new region

In the within-region experiment, the physics-residual model was at least as accurate as the Stefan model recalibrated on the same labels at every label number from 25 upwards (25 labels: Alaska −0.9 [−1.9, +0.1], Lena Delta −0.8 [−2.2, −0.1] and Canada −0.3 [−3.2, +1.7] cm). With 100 labels, the gain was distinguishable from zero in all three regions (−0.7 [−1.4, −0.01], −2.0 [−5.3, −0.3] and −1.9 [−3.7, −0.3] cm; in Alaska only at 100 and 250 labels; Supplementary Fig. S5). Direct ML trained on 25 labels had a larger site-year RMSE than the Stefan model in Alaska (+2.8 [−0.3, +4.5] cm) and Canada (+1.2 [−1.5, +3.4] cm). For the point estimate of direct ML to fall below that of the physics-residual model, at least 50 labels were needed in the Lena Delta, 250 in Canada and 500 in Alaska.

In the transfer experiment (Fig. 4c), the RMSE in the Lena Delta without target labels was 39.5 cm for direct ML, 21.5 cm for the source-coefficient Stefan model and 23.8 cm for the physics-residual model. With 25 target labels, the RMSE decreased to 19.9 cm for the physics-residual model and to 21.3 cm for direct ML. In Alaska, the physics-residual model had the lowest

error at every n (17.8 cm without target labels). In Canada, direct ML trained on the source regions was the most accurate (24.2 cm). Recalibrating the Stefan coefficient on 25 Canadian labels increased its error from 27.7 to 31.3 cm, and the error of the physics-residual model, whose anchor uses this coefficient, also increased from 24.3 to 26.3 cm. The thaw depths of the two Canadian data sources differ greatly (mean 63.4 cm at the transects and 89.1 cm at the CALM grids), and we consider that one coefficient fitted to 25 labels could not represent both groups.

Comparison with published products on the same sites and years

Fig. 5 compares our models with the five products on the same site-years. Our 19,172 site-years outnumber the training data of the two ML products (Supplementary Methods), but they are confined to three regions and consist mostly of transect points measured in a single year. The products, in contrast, were trained on long-record sites across the Northern Hemisphere. In this comparison, the products were therefore evaluated away from most of their training sites, and our models at spatial blocks withheld from training. On the 461 common site-years of the Lena Delta (2013–2020), the RMSE of the physics-residual model was 21.1 cm. The RMSEs of the products were 28.4 cm for ESA CCI v5, 30.6 cm for Aalto et al., 31.0 cm for ESA CCI v4, 33.0 cm for Liu et al. and 34.6 cm for Wei et al. (Fig. 5a, Supplementary Table S6). At all Lena Delta site-years more than 25 km from the training sites of Wei et al. that have a product value for the observation year, the paired differences were −12.5 [−15.6, −5.8] cm relative to Wei et al. and −5.0 [−9.5, −2.5] cm relative to ESA CCI v5 (802 and 2352 site-years, respectively). In Alaska (17.1 cm versus 17.9 cm for Wei et al. and 18.8 cm for Liu et al.) and Canada (25.8 versus 27.0 and 27.9 cm), the physics-residual model and the two ML products had similar accuracy.

In the Lena Delta, the two ML products predicted about 40–70 cm at most sites regardless of the observed ALT (Fig. 5b). The Pearson correlation coefficient (r) between observed and predicted values was 0.05 for Wei et al., 0.07 for Liu et al. and 0.73 for the physics-residual model. The mean bias was +21.9 and +19.8 cm for the two products and +2.8 cm for the physics-residual model. The overestimation was largest where the observed active layer was thinnest (Fig. 5b). In the Lena Delta, the error distributions of the two ML products were shifted by about 20 cm, and their spread was similar to that of the physics-residual model. In Alaska and Canada, the errors of all series were centred near zero (Fig. 5c). The Alaska label sites lie at a median distance of 7 km from the training sites of Wei et al., whereas the Lena Delta sites lie at a median distance of 107 km (Supplementary Fig. S6). A model trained without any Lena Delta label was also more accurate than Wei et al. on the same site-years (26.0 versus 30.3 cm; ΔRMSE −4.3 [−6.1, −0.8] cm). Year by year, the physics-residual model had a smaller error than Wei et al. in all 6 Lena Delta years with at least 30 site-years and in 7 of 13 Alaska years (Supplementary Fig. S7).

The maps of the year with the most observed cells showed the same differences (Fig. 6). Our map kept the level of the recalibrated Stefan model from which it starts (medians inside the area of applicability: 60 and 64 cm in Alaska, 38 and 40 cm in the Lena Delta, and 66 and 66 cm in Canada). The products, in contrast, were deeper in the Lena Delta (56 and 52 cm) and Canada (84 and 74 cm). At the 126 observed cells in Alaska, the RMSEs of the three maps were similar (14.9, 13.4 and 12.9 cm), and they were also similar at the 67 cells more than 25 km from a training site of Wei et al. (9.1, 10.9 and 8.5 cm). The correlation with the observations was lower for our model (0.34) than for the two products (0.53 and 0.57). In the Lena Delta (14 observed cells), both products were deeper than the observations (mean bias +23.6 and +27.3 cm; RMSE 26.2 and 30.4 cm, against 12.7 cm for our model). At the 33 observed cells in Canada (NW), the RMSE was 23.7 cm for our model, 29.9 cm for Wei et al. (mean bias +20.5 cm) and 20.2 cm for Liu et al. At the 27 cells more than 25 km from a training site of Wei et al., the RMSEs were 17.5, 31.4 and 21.0 cm, respectively.

The residual learner can also use a product as its anchor (equation (10)). In the Lena Delta, the residual model with the Stefan anchor was more accurate than the residual model anchored to the Wei et al. product (−2.8 [−6.2, −0.1] cm; Supplementary Fig. S8). In Alaska and Canada, the difference between the two anchors could not be distinguished from zero.

Yearly 1 km ALT maps, 2011–2024

Fig. 7 shows the yearly maps of the physics-residual model for 2011–2024. Relative to the 2011–2024 mean, the regional mean anomaly ranged from −2.8 to +6.0 cm in Alaska, from −8.9 to +9.0 cm in the Lena Delta and from −3.8 to +6.4 cm in Canada (Fig. 7d). The correlation (r) between the regional mean anomaly and the regional mean TDD to 1 September was 0.92 in Alaska, 0.96 in the Lena Delta and 1.00 in Canada. The regional mean ALT was deepest in the years with the largest TDD (2019, 2021 and 2023) and shallowest in the years with the smallest TDD (2014, 2017 and 2020). Supplementary Fig. S9 compares the yearly changes of the maps and the observations at long-record sites. The linear trend of the yearly maps averaged +2.1, +6.7 and +1.5 cm per decade over the cells of Alaska, the Lena Delta and Canada, and the trend was positive in 78%, 100% and 91% of the cells (Fig. 7e). The area of applicability covers 56% of the Alaska cells, 62% of the Lena Delta cells and 43% of the Canada cells, and the cells outside it lie mainly in mountains and at the southern margin of the permafrost zone (Supplementary Fig. S10). The half-width of the 90% prediction interval was 25.5 cm in Alaska, 24.5 cm in the Lena Delta and 41.3 cm in Canada (dissimilarity and residual-correction maps in Supplementary Figs S10 and S11). In Alaska, the anomaly of a warm year was largest in the interior lowlands and smallest on the North Slope (Supplementary Fig. S12).

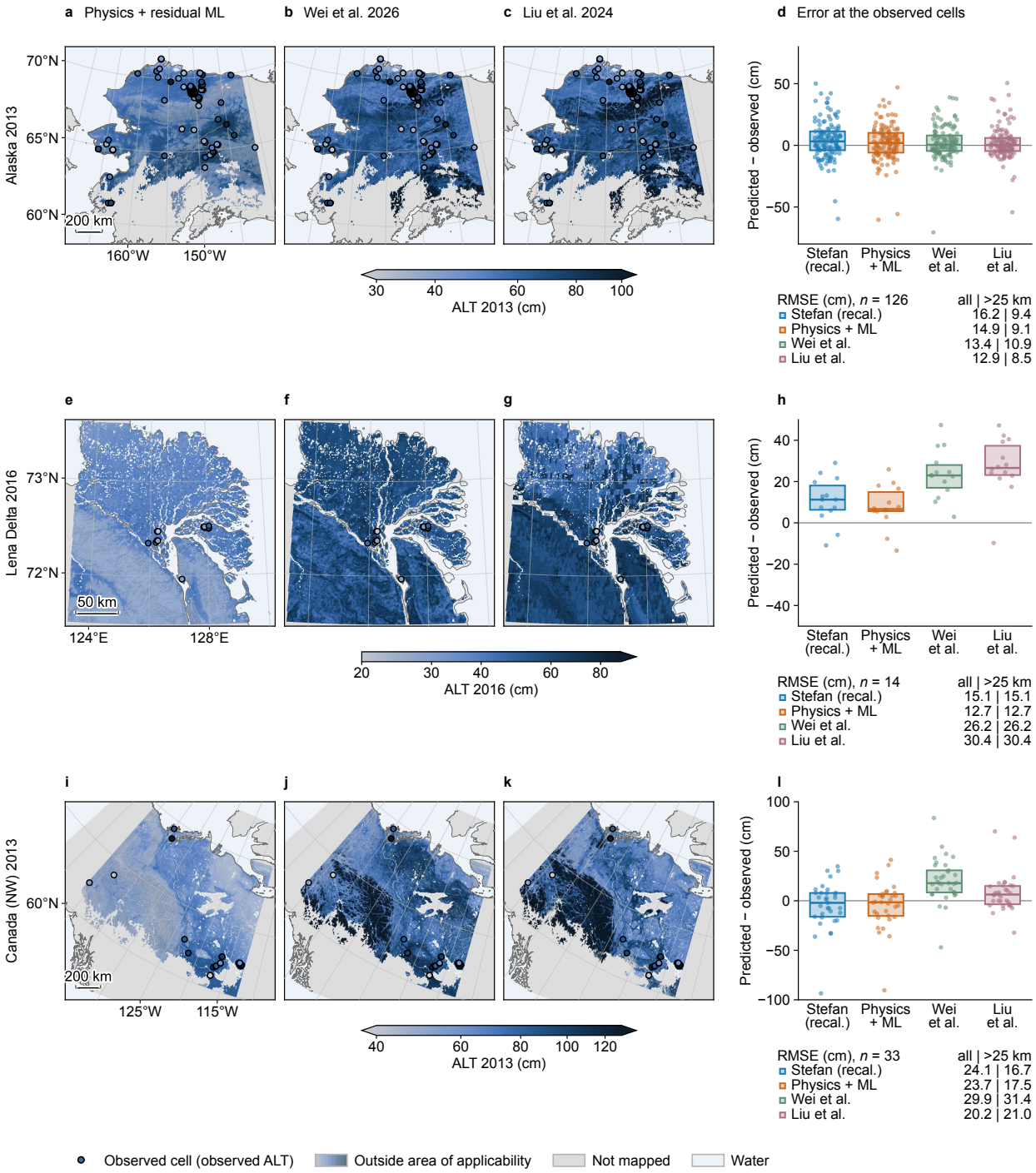

Figure 6. ALT maps of the physics-residual model and two published products in the year with the most observed cells. **a–c**, Alaska, 2013; circles, observed cells. **d**, Prediction errors at the observed cells (boxes as in Fig. 5c), with the RMSE for all cells and for cells more than 25 km from a training site of Wei et al. **e–h**, Lena Delta, 2016. **i–l**, Canada (NW), 2013. Lighter colours, outside the area of applicability.

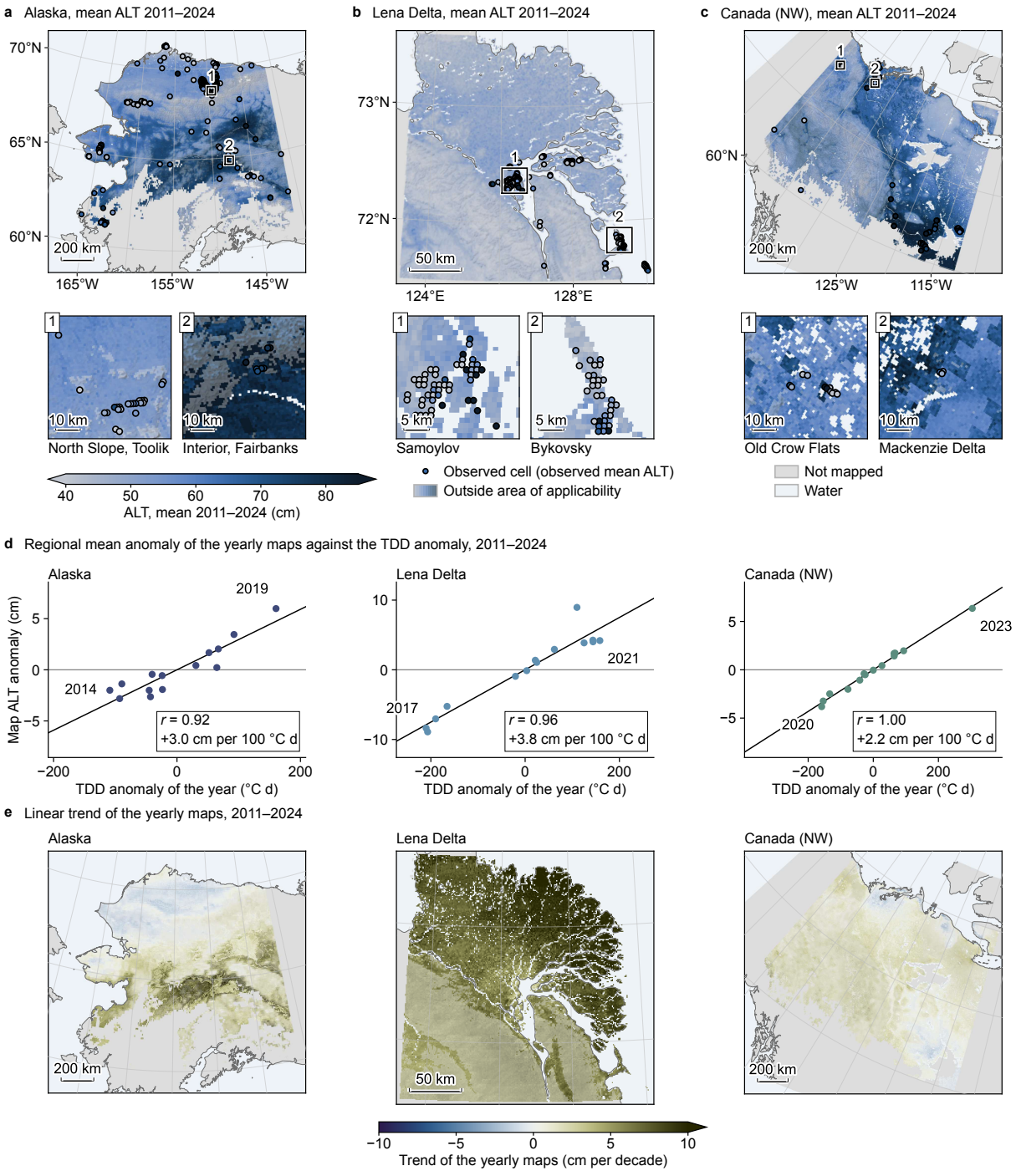

Figure 7. Yearly 1 km ALT maps, 2011–2024. a–c, Mean ALT of 2011–2024 in Alaska, the Lena Delta and Canada (NW), with two windows per region enlarged below; circles, observed cells; lighter colours, outside the area of applicability. **d**, Regional mean ALT anomaly against regional mean TDD anomaly, one point per year. **e**, Linear trend of the yearly maps.

Discussion

In the physics-residual model, the Stefan term represents the response to the thaw-season climate of each year, and the learner corrects the effects of soil and terrain that one regional coefficient cannot represent (equation (5)). In this study, the gain over the Stefan model appeared when terrain and soil covariates were included, and it was larger where the Stefan model had larger errors. These results are consistent with the interpretation that terrain and soil determine the effective coefficient of each site. Direct ML, in contrast, has to learn the response to climate from the labels as well. It was slightly more accurate when all labels of a region were used, but its accuracy decreased greatly when it predicted a region withheld from training. The physics-residual model was within 3.3 cm of the most accurate of the three methods in every region and validation design. Most of the permafrost region lies far from any label, so mapping requires a model that keeps its accuracy outside the training data.

Alaska was a region where residual learning had little room for improvement. In Alaska, the spatial-block RMSE of the recalibrated Stefan model was 18.6 cm, and it differed from that of the physics-residual model by less than 1 cm. When the region was withheld from training, the physics-residual model was the most accurate. This result can be explained by two properties of the Alaska labels. First, the labels in Alaska are dense, and the recalibrated Stefan model already fits the regional level and the year-to-year variation well, so little systematic residual is left for the covariates to explain. Second, most of the remaining error is the spread between measurement points within one 1 km cell (pooled standard deviation 13.0 cm), which covariates at 1 km resolution cannot distinguish. The difference was small in every tercile (Fig. 4b), on cell-year means (Supplementary Table S5) and without the GPR-derived labels (+0.3 [−0.4, +0.6] cm; Supplementary Methods). The spatial-block comparison with the recalibrated Stefan model can therefore show whether a region needs the residual learner. A negative ΔRMSE that is distinguishable from zero, as in the Lena Delta and Canada, indicates systematic residuals that the covariates can explain. Where it is not, as in Alaska, the learned correction stays small, and the workflow keeps the accuracy of the recalibrated Stefan model.

When compared on the same sites and years, the accuracy of the ML products depended on the distance to their training sites. This difference cannot be detected by validation within the products' own site sets. Where the training sites are close to our label sites (Alaska, and Canada for Liu et al., who used an earlier release of the same observations for training), the products were as accurate as our model, but these comparisons are not independent. In the Lena Delta, which has no nearby training site (median distance 107 km), the products predicted 40–70 cm regardless of the observed ALT and overestimated ALT by about 20 cm. The physics-residual model trained without Lena Delta labels was also more accurate than the Wei et al. product there. This result follows the same pattern as that of direct ML with the Lena Delta withheld from training. The predictors of the products are yearly, but their response to climate is learned at training sites elsewhere. In the physics-residual model, the response to the TDD of the year is represented by the Stefan term, which does not depend on the training sites. ESA CCI, a process-based product that is not trained on ALT observations, was less accurate than our model in all three regions, but unlike the ML products it was not systematically biased in the Lena Delta (Fig. 5c).

Gautam et al. [23] reported, from one year of data at 68 Alaska CALM sites, that a random forest lost much of its accuracy at test sites whereas the Stefan model kept its accuracy. The same tendency appeared in this study for the site-years of 2010–2024, in three regions and under region hold-out. The physics-residual model kept the stability of the Stefan model and added the within-region accuracy of the learner. Studies of physics-guided learning for lakes and rivers reported that process knowledge helps mainly when data are scarce or extrapolation is needed [20–22]. For ALT, this benefit appeared under region hold-out. Direct ML transferred to Canada, whose inputs lie largely within the range of the training regions, but not to the Lena Delta and Alaska, where extrapolation was needed. Random cross-validation underestimated the spatial-block error of direct ML by 3.6–9.9 cm, which agrees with previous studies on spatial validation [10–12].

Each map uses the ERA5-Land climate of one year, including the TDD accumulated to 1 September, so it can be compared with observations or products of the same year. In a new region without local observations, a first map can be made with the Stefan model using the coefficient of other regions or with the physics-residual model trained on other regions, and its error does not depend on the distance to training sites. The map can then be improved once labels that represent the region are obtained. In the Lena Delta, 25 labels reduced the error of the physics-residual model from 23.8 to 19.9 cm. In Canada, however, the error increased from 24.3 to 26.3 cm because of the difference between the two observation protocols. When a new region is calibrated, labels obtained with one consistent protocol should therefore be used.

Limitations and ways to address them are as follows. The Lena Delta labels come from one data set distributed over 16 blocks, so the intervals for the Lena Delta are wide. Half of the Alaska labels come from the 2014 survey, so the effect of the year-specific input in Alaska could not be distinguished from zero, and the result was the same when 2014 was excluded (0.0 [−0.7, +0.9] cm). Observations from more sites and years are expected to reduce both problems. The labels combine probe and GPR transects and CALM grid means from three observation programmes, and the model error differs between the programmes (Supplementary Table S7). The largest difference appeared at the 65 Canadian CALM site-years, where the model underestimated the mean ALT of 89.1 cm (mean bias −23.0 cm). Harmonised measurement protocols or a programme term in the model are expected to reduce this error. The climate input is monthly, and the soil covariates are static. Daily climate

and remotely sensed snow and soil moisture are expected to explain more of the residual. The Kudryavtsev model, which includes snow and moisture terms, was more accurate than the recalibrated Stefan model in the Lena Delta and Canada under spatial-block validation, so an anchor that accounts for moisture may reduce the error further. Active layers thicker than about 3 m lie outside the training range, and the comparison of a single-year map with observations rests on few cells in the Lena Delta and Canada. Both ranges can be extended by adding labels from other regions.

Conclusions

This study proposed a physics-residual workflow in which a gradient-boosting learner corrects a Stefan model recalibrated with local observations. Using site-year data in which every observation is paired with the climate of its year, we evaluated the workflow in three permafrost regions under five validation designs and with label numbers from 25 to all labels, and compared it with five published products.

The workflow showed three improvements over existing methods. First, it improved on the physical model in regions with diverse soils and terrain. In the Lena Delta and Canada, the learner represented the effects of soil and terrain that one regional coefficient cannot represent. In Alaska, where the recalibrated Stefan model alone is accurate, the workflow kept that accuracy, and with local labels it was at least as accurate as the recalibrated Stefan model at every label number. Second, it was more accurate than the published products far from their training sites. In the Lena Delta, the products overestimated the thin active layers by about 20 cm, whereas the mean bias of the physics-residual model was +2.8 cm. Third, it was more stable than direct ML. The accuracy of direct ML decreased greatly in regions withheld from training, whereas the physics-residual model stayed close to the more accurate of the recalibrated Stefan model and direct ML in every region and validation design. The workflow suits ALT prediction, for which field observations are sparse, and it extends the combination of physical models and machine learning recommended for Alaska by Gautam et al. [23] to other regions and to small numbers of labels.

The yearly 1 km maps of 2011–2024, with the area of applicability and prediction intervals, provide the thaw depth of each year where there are no observations. With climate data alone, the workflow can produce a Stefan-based map for any permafrost region, and the map can then be corrected with local labels. The evaluation procedure, which consists of region hold-out and site-year scoring, can be used as a common standard for comparing ALT products.

Data availability

All input data are publicly available. ALT observations: ABoVE soil moisture and active-layer thickness, version 2, at https://doi.org/10.3334/ORNLDAAC/2369 (ORNL DAAC [6]), ALLena thaw-depth data at https://doi.org/10.1594/PANGAEA.973813 (PANGAEA [7]) and GTN-P CALM at https://doi.org/10.1594/PANGAEA.972777 (PANGAEA [5]). Climate: ERA5-Land monthly averaged data (Copernicus Climate Data Store [26]). Covariates: Copernicus DEM GLO-30 (ESA [27], accessed through AWS Open Data), SoilGrids 2.0 [28], ESA CCI permafrost extent v4.0 (CEDA [31]), JRC Global Surface Water v1.4 [52], and MODIS MOD13Q1 and MOD10A2 version 6.1 (NASA, accessed through AppEEARS [29, 30]); for the sensitivity tests, NCSCD v2 (Bolin Centre [36, 37]), raster CAVM (Mendeley Data [38]) and ESA WorldCover 2021 v200 (Zenodo [39]). Products: Wei et al. [32] at https://doi.org/10.5281/zenodo.21667583 (Zenodo, version 2), Liu et al. [33] at https://doi.org/10.5281/zenodo.10070610 (Zenodo), ESA CCI ALT v4 and v5 (CEDA [14, 15]), Aalto et al. [34] at https://doi.org/10.5061/dryad.886pr72 (Dryad) and Yi and Kimball [35] at https://doi.org/10.3334/ORNLDAAC/1760 (ORNL DAAC). The site-year label table, the yearly climate-input table, the static covariate table, the out-of-fold predictions of all methods and the yearly 1 km maps of Alaska, the Lena Delta and north-western Canada (2011–2024) generated in this study are available from the corresponding author on reasonable request.

References

1. Hjort, J. *et al.* Degrading permafrost puts Arctic infrastructure at risk by mid-century. *Nat. Commun.* **9**, 5147 (2018). DOI 10.1038/s41467-018-07557-4.
2. Schuur, E. A. G. *et al.* Climate change and the permafrost carbon feedback. *Nature* **520**, 171–179 (2015). DOI 10.1038/nature14338.
3. Biskaborn, B. K. *et al.* Permafrost is warming at a global scale. *Nat. Commun.* **10**, 264 (2019). DOI 10.1038/s41467-018-08240-4.
4. Nelson, F. E. *et al.* Estimating active-layer thickness over a large region: Kuparuk River basin, Alaska, U.S.A. *Arct. Alp. Res.* **29**, 367–378 (1997). DOI 10.1080/00040851.1997.12003258.
5. Streletskiy, D. A. *et al.* GTN-P CALM: 35 years of active layer thickness (ALT) across latitudinal and elevational gradients in the Northern Hemisphere. *PANGAEA* https://doi.org/10.1594/PANGAEA.972777 (2025).

6. Moore, M. A. *et al.* ABoVE: soil moisture and active layer thickness in Alaska, USA and Canada, 2005–2024, version 2. *ORNL Distributed Active Archive Center* https://doi.org/10.3334/ORNLDAAC/2369 (2025).

7. Veremeeva, A. *et al.* ALLena: thaw depth measurements of the active layer in the Lena River Delta region from 1998 to 2022, Northeastern Siberia. *PANGAEA* https://doi.org/10.1594/PANGAEA.973813 (2025).

8. Stefan, J. Ueber die Theorie der Eisbildung, insbesondere über die Eisbildung im Polarmeere. *Ann. Phys.* **278**, 269–286 (1891). DOI 10.1002/andp.18912780206.

9. Riseborough, D., Shiklomanov, N., Etzelmüller, B., Gruber, S. & Marchenko, S. Recent advances in permafrost modelling. *Permafr. Periglac. Process.* **19**, 137–156 (2008). DOI 10.1002/ppp.615.

10. Roberts, D. R. *et al.* Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* **40**, 913–929 (2017). DOI 10.1111/ecog.02881.

11. Ploton, P. *et al.* Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nat. Commun.* **11**, 4540 (2020). DOI 10.1038/s41467-020-18321-y.

12. Meyer, H. & Pebesma, E. Machine learning-based global maps of ecological variables and the challenge of assessing them. *Nat. Commun.* **13**, 2208 (2022). DOI 10.1038/s41467-022-29838-9.

13. Wadoux, A. M. J.-C., Heuvelink, G. B. M., de Bruin, S. & Brus, D. J. Spatial cross-validation is not the right way to evaluate map accuracy. *Ecol. Model.* **457**, 109692 (2021). DOI 10.1016/j.ecolmodel.2021.109692.

14. Westermann, S. *et al.* ESA Permafrost Climate Change Initiative (Permafrost_cci): permafrost active layer thickness for the Northern Hemisphere, v4.0. *NERC EDS Centre for Environmental Data Analysis* https://doi.org/10.5285/d34330ce3f604e368c06d76de1987ce5 (2024).

15. Westermann, S. *et al.* ESA Permafrost Climate Change Initiative (Permafrost_cci): permafrost active layer thickness for the Northern Hemisphere, v5.0. *NERC EDS Centre for Environmental Data Analysis* https://doi.org/10.5285/a6fbedd8ee5b472c8e84e55f746c1704 (2025).

16. Liu, Z. *et al.* Widespread deepening of the active layer in northern permafrost regions from 2003 to 2020. *Environ. Res. Lett.* **19**, 014020 (2024). DOI 10.1088/1748-9326/ad0f73.

17. Wei, Y., Wu, Z., Wang, J., Shen, H. & Chen, Y. A 1 km resolution dataset of Northern Hemisphere permafrost active layer thickness (2000–2024). Preprint at https://doi.org/10.5194/essd-2026-447 (2026).

18. Aalto, J., Karjalainen, O., Hjort, J. & Luoto, M. Statistical forecasting of current and future circum-Arctic ground temperatures and active layer thickness. *Geophys. Res. Lett.* **45**, 4889–4898 (2018). DOI 10.1029/2018GL078007.

19. Ran, Y. *et al.* New high-resolution estimates of the permafrost thermal state and hydrothermal conditions over the Northern Hemisphere. *Earth Syst. Sci. Data* **14**, 865–884 (2022). DOI 10.5194/essd-14-865-2022.

20. Willard, J., Jia, X., Xu, S., Steinbach, M. & Kumar, V. Integrating scientific knowledge with machine learning for engineering and environmental systems. *ACM Comput. Surv.* **55**, 66 (2022). DOI 10.1145/3514228.

21. Read, J. S. *et al.* Process-guided deep learning predictions of lake water temperature. *Water Resour. Res.* **55**, 9173–9190 (2019). DOI 10.1029/2019WR024922.

22. Jia, X. *et al.* Physics-guided machine learning for scientific discovery: an application in simulating lake temperature profiles. *ACM/IMS Trans. Data Sci.* **2**, 20 (2021). DOI 10.1145/3447814.

23. Gautam, S., Mishra, U., Scott, S. N. & Lara, M. J. Machine learning and process-based modeling of spatiotemporal changes in active layer thickness across Alaska. *Sci. Rep.* **15**, 42420 (2025). DOI 10.1038/s41598-025-26586-w.

24. Muñoz-Sabater, J. *et al.* ERA5-Land: a state-of-the-art global reanalysis dataset for land applications. *Earth Syst. Sci. Data* **13**, 4349–4383 (2021). DOI 10.5194/essd-13-4349-2021.

25. Brown, J., Hinkel, K. M. & Nelson, F. E. The Circumpolar Active Layer Monitoring (CALM) program: research designs and initial results. *Polar Geogr.* **24**, 166–258 (2000). DOI 10.1080/10889370009377698.

26. Copernicus Climate Change Service. ERA5-Land monthly averaged data from 1950 to present. *Copernicus Climate Change Service (C3S) Climate Data Store (CDS)* https://doi.org/10.24381/cds.68d2bb30 (2019).

27. European Space Agency & Airbus. Copernicus DEM. *European Space Agency* https://doi.org/10.5270/ESA-c5d3d65 (2022).

28. Poggio, L. *et al.* SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. *SOIL* **7**, 217–240 (2021). DOI 10.5194/soil-7-217-2021.

29. Didan, K. MODIS/Terra vegetation indices 16-day L3 global 250m SIN grid V061. *NASA EOSDIS Land Processes DAAC* https://doi.org/10.5067/MODIS/MOD13Q1.061 (2021).

30. Hall, D. K. & Riggs, G. A. MODIS/Terra snow cover 8-day L3 global 500m SIN grid, version 61. *NASA National Snow and Ice Data Center DAAC* https://doi.org/10.5067/MODIS/MOD10A2.061 (2021).

31. Westermann, S. *et al.* ESA Permafrost Climate Change Initiative (Permafrost_cci): permafrost extent for the Northern Hemisphere, v4.0. *NERC EDS Centre for Environmental Data Analysis* https://doi.org/10.5285/93444bc1c4364a59869e004bf9bfd94a (2024).

32. Wei, Y. & Chen, Y. A 1 km resolution dataset of Northern Hemisphere permafrost active layer thickness (2000–2024), version 2. *Zenodo* https://doi.org/10.5281/zenodo.21667583 (2026).

33. Liu, Z. Gridded active layer thickness across northern permafrost regions from 2003 to 2020. *Zenodo* https://doi.org/10.5281/zenodo.10070610 (2023).

34. Aalto, J., Karjalainen, O., Hjort, J. & Luoto, M. Data from: statistical forecasting of current and future circum-Arctic ground temperatures and active layer thickness. *Dryad* https://doi.org/10.5061/dryad.886pr72 (2019).

35. Yi, Y. & Kimball, J. S. ABoVE: active layer thickness from remote sensing permafrost model, Alaska, 2001–2015. *ORNL Distributed Active Archive Center* https://doi.org/10.3334/ORNLDAAC/1760 (2020).

36. Hugelius, G. *et al.* The Northern Circumpolar Soil Carbon Database: spatially distributed datasets of soil coverage and soil carbon storage in the northern permafrost regions. *Earth Syst. Sci. Data* **5**, 3–13 (2013). DOI 10.5194/essd-5-3-2013.

37. Hugelius, G. *et al.* A new data set for estimating organic carbon storage to 3 m depth in soils of the northern circumpolar permafrost region. *Earth Syst. Sci. Data* **5**, 393–402 (2013). DOI 10.5194/essd-5-393-2013.

38. Raynolds, M. K. *et al.* A raster version of the Circumpolar Arctic Vegetation Map (CAVM). *Remote Sens. Environ.* **232**, 111297 (2019). DOI 10.1016/j.rse.2019.111297.

39. Zanaga, D. *et al.* ESA WorldCover 10 m 2021 v200. *Zenodo* https://doi.org/10.5281/zenodo.7254221 (2022).

40. Kudryavtsev, V. A., Garagula, L. S., Kondrat'yeva, K. A. & Melamed, V. G. *Osnovy Merzlotnogo Prognoza [Fundamentals of Frost Forecasting in Geological Engineering Investigations]* (Moscow State University, 1974). In Russian. English translation: CRREL Draft Translation 606, US Army Cold Regions Research and Engineering Laboratory, Hanover, NH, 1977.

41. Romanovsky, V. E. & Osterkamp, T. E. Thawing of the active layer on the coastal plain of the Alaskan Arctic. *Permafr. Periglac. Process.* **8**, 1–22 (1997). DOI 10.1002/(SICI)1099-1530(199701)8:1<1::AID-PPP243>3.0.CO;2-U.

42. Shiklomanov, N. I. & Nelson, F. E. Active-layer mapping at regional scales: a 13-year spatial time series for the Kuparuk region, north-central Alaska. *Permafr. Periglac. Process.* **13**, 219–230 (2002). DOI 10.1002/ppp.425.

43. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. & Gulin, A. CatBoost: unbiased boosting with categorical features. *Adv. Neural Inf. Process. Syst.* **31**, 6639–6649 (2018).

44. Grinsztajn, L., Oyallon, E. & Varoquaux, G. Why do tree-based models still outperform deep learning on typical tabular data? *Adv. Neural Inf. Process. Syst.* **35**, 507–520 (2022).

45. Gorishniy, Y., Rubachev, I., Khrulkov, V. & Babenko, A. Revisiting deep learning models for tabular data. *Adv. Neural Inf. Process. Syst.* **34**, 18932–18943 (2021).

46. Gorishniy, Y., Kotelnikov, A. & Babenko, A. TabM: advancing tabular deep learning with parameter-efficient ensembling. *Int. Conf. Learn. Represent.* (2025).

47. Holzmüller, D., Grinsztajn, L. & Steinwart, I. Better by default: strong pre-tuned MLPs and boosted trees on tabular data. *Adv. Neural Inf. Process. Syst.* **37**, 26577–26658 (2024).

48. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **30**, 4765–4774 (2017).

49. Efron, B. & Tibshirani, R. J. *An Introduction to the Bootstrap* (Chapman and Hall, New York, 1993).

50. Chen, R. H. *et al.* ABoVE: active layer thickness from airborne L- and P-band SAR, Alaska, 2017, ver. 3. *ORNL Distributed Active Archive Center* https://doi.org/10.3334/ORNLDAAC/2004 (2022).

51. Obu, J. *et al.* Northern Hemisphere permafrost map based on TTOP modelling for 2000–2016 at 1 km^2 scale. *Earth-Sci. Rev.* **193**, 299–316 (2019). DOI 10.1016/j.earscirev.2019.04.023.

52. Pekel, J.-F., Cottam, A., Gorelick, N. & Belward, A. S. High-resolution mapping of global surface water and its long-term changes. *Nature* **540**, 418–422 (2016). DOI 10.1038/nature20584.

53. Meyer, H. & Pebesma, E. Predicting into unknown space? Estimating the area of applicability of spatial prediction models. *Methods Ecol. Evol.* **12**, 1620–1633 (2021). DOI 10.1111/2041-210X.13650.

54. Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J. & Wasserman, L. Distribution-free predictive inference for regression. *J. Am. Stat. Assoc.* **113**, 1094–1111 (2018). DOI 10.1080/01621459.2017.1307116.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2026-25484554). This work was also supported by the Basic Research Project of the Korea Institute of Geoscience and Mineral Resources, "Complex Disaster-Resilient Safety Network Innovation Program" (2710110524), funded by the Ministry of Science and ICT of Korea.

Author contributions statement

S.B. designed the study, compiled the data, performed the analyses and wrote the manuscript. S.H.K. and Y.T.J. contributed to the interpretation of the remote-sensing covariates and the permafrost field data and revised the manuscript. Y.C. supervised the study and revised the manuscript. All authors approved the final manuscript.

Additional information

Supplementary information: Supplementary Methods, Supplementary Tables S1–S7 and Supplementary Figures S1–S12 accompany this paper.

Competing interests: The authors declare no competing interests.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation.pdf