Title

Machine learning models inaccurately predict current and future high-latitude C balances

Journal

Environmental Research Letters, 18(1)

ISSN

1748-9318

Authors

Shirley, Ian A

Mekonnen, Zelalem A

Grant, Robert F

et al.

Publication Date

2023

DOI

10.1088/1748-9326/acacb2

ENVIRONMENTAL RESEARCH
LETTERS

LETTER • OPEN ACCESS

Machine learning models inaccurately predict current and future high-latitude C balances

IOP Publishing Environ. Res. Lett. 18 (2023) 014026 https://doi.org/10.1088/1748-9326/acacb2

ENVIRONMENTAL RESEARCH
LETTERS

OPEN ACCESS

RECEIVED
1 September 2022

REVISED
23 November 2022

ACCEPTED FOR PUBLICATION
19 December 2022

PUBLISHED
6 January 2023

LETTER

Machine learning models inaccurately predict current and future high-latitude C balances

Ian A Shirley[1,2,*], **Zelalem A Mekonnen**[1], **Robert F Grant**[3], **Baptiste Dafflon**[1] and **William J Riley**[1]

[1] Climate and Ecosystem Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, United States of America
[2] Department of Physics, University of California-Berkeley, Berkeley 94720-3114, CA, United States of America
[3] Department of Renewable Resources, University of Alberta, Edmonton, Canada
[*] Author to whom any correspondence should be addressed.

E-mail: IShirley@lbl.gov

Keywords: machine learning, upscaling, forecasting, independent evaluation, carbon cycle, high-latitudes, climate change

Supplementary material for this article is available online

Abstract

The high-latitude carbon (C) cycle is a key feedback to the global climate system, yet because of system complexity and data limitations, there is currently disagreement over whether the region is a source or sink of C. Recent advances in big data analytics and computing power have popularized the use of machine learning (ML) algorithms to upscale site measurements of ecosystem processes, and in some cases forecast the response of these processes to climate change. Due to data limitations, however, ML model predictions of these processes are almost never validated with independent datasets. To better understand and characterize the limitations of these methods, we develop an approach to independently evaluate ML upscaling and forecasting. We mimic data-driven upscaling and forecasting efforts by applying ML algorithms to different subsets of regional process-model simulation gridcells, and then test ML performance using the remaining gridcells. In this study, we simulate C fluxes and environmental data across Alaska using *ecosys*, a process-rich terrestrial ecosystem model, and then apply boosted regression tree ML algorithms to training data configurations that mirror and expand upon existing AmeriFLUX eddy-covariance data availability. We first show that a ML model trained using *ecosys* outputs from currently-available Alaska AmeriFLUX sites incorrectly predicts that Alaska is presently a modeled net C source. Increased spatial coverage of the training dataset improves ML predictions, halving the bias when 240 modeled sites are used instead of 15. However, even this more accurate ML model incorrectly predicts Alaska C fluxes under 21st century climate change because of changes in atmospheric CO_2, litter inputs, and vegetation composition that have impacts on C fluxes which cannot be inferred from the training data. Our results provide key insights to future C flux upscaling efforts and expose the potential for inaccurate ML upscaling and forecasting of high-latitude C cycle dynamics.

1. Introduction

High-latitude ecosystems play a key role in the global carbon (C) cycle because it is the fastest warming region on the planet (Serreze *et al* 2009) with large stocks of C stored in permafrost soils (Hugelius *et al* 2014). However, estimates of current and future high-latitude C balance are highly uncertain, as highlighted by a current mismatch between process-models (which estimate that the high-latitudes are currently a C sink; McGuire *et al* 2012) and observation-based estimates (some of which estimate that high-latitudes are currently a C source; Commane *et al* 2017, Natali *et al* 2019). This uncertainty is driven by high spatial and temporal variability (Grant *et al* 2017, Uhlemann *et al* 2021), a complex and interacting set of environmental drivers and feedbacks (Arora *et al* 2019), and limited data due to harsh environmental conditions and remoteness of northern lands.

Direct measurements of ecosystem C exchanges are performed using eddy-covariance (EC), chambers, and long-term plot based observations, but these methods can only be applied at local scales and provide spatially sparse estimates. Uses of this data to inform regional estimates of current-day C fluxes are typically divided into two approaches. Process-based terrestrial ecosystem models, which simulate physical, chemical, and biological responses to driving environmental variables, use site data for calibration and to validate model predictions. Upscaling methods, in contrast, use the data to generate statistical or machine learning (ML) models which are then applied to predict across a larger study domain. ML models typically outperform statistical regression-based approaches in predictive power due to their ability to capture complexity, non-linearities, and interaction effects (Virkkala *et al* 2021). In recent years, many studies have applied ML models to upscale C fluxes at regional (Natali *et al* 2019, Peltola *et al* 2019, Reitz *et al* 2021, Virkkala *et al* 2021, Abbasian *et al* 2022) and global (Jung *et al* 2020, Zeng *et al* 2020) scales. ML approaches have also been applied to upscale other ecosystem processes including vegetation dynamics (Pearson *et al* 2013, Bastin *et al* 2019), crop yields (Crane-Droesch 2018), and changes in soil C stocks (Mishra *et al* 2021, Naidu and Bagchi 2021).

ML model performance is strongly dependent on the size and spatial distribution of the training dataset, particularly as the complexity of the modeled system increases (Raudys and Jain 1991). High-latitude ecosystems are highly complex systems with tightly coupled C, heat, water, and nutrient cycles characterized by strong heterogeneity, feedback loops, and interaction effects (Nobrega and Grogan 2008, Jorgenson *et al* 2010, Belshe *et al* 2012, Keuper *et al* 2012, Becker *et al* 2016, Cable *et al* 2016, Finger *et al* 2016, Grant *et al* 2017, Jafarov *et al* 2018, Arora *et al* 2019, Waldrop *et al* 2021, Mekonnen *et al* 2021b). However, data availability at high-latitudes is very limited. Although Alaska is the high-latitude region with the highest density of EC flux towers (25 AmeriFLUX towers were active at some point during the period 2010–2020), the footprint of each flux tower is only ∼1 km^2. Therefore, these EC flux towers monitor only ∼0.002% of the Alaskan land surface (148 1346 km^2; Tramontana *et al* 2016). Performance of ML models of highly complex ecosystem processes trained on such limited data may suffer from underspecification (D'Amour *et al* 2020), shortcut learning (Geirhos *et al* 2020), and other structural mismatches between available data and underlying dynamics (Arjovsky *et al* 2020). Since these effects cannot be quantified using the training dataset, commonly employed techniques like *k*-fold cross-validation (*k*CV) may lead to overestimation of model performance. These issues may lead to overconfidence in the ML model predictions, particularly since they are derived from data.

Some ML upscaling studies (Pearson *et al* 2013, Bastin *et al* 2019, Natali *et al* 2019, Naidu and Bagchi 2021) use ML models generated using training data from current climate conditions to forecast responses of ecosystem processes to decades of climate change. This approach is attractive because while ML models are challenging to generate they are easy to use for predictions. However, many factors that drive and interact with ecosystem processes will change significantly under future climate conditions (e.g. atmospheric CO_2 concentration, air and soil temperatures, nutrient and water availability, vegetation composition), leading to an expected degradation of ML model performance. Validation of ML model performance under future climate conditions is not possible today, and given the low interpretability of typical ML models, it is not clear how strongly ML model performance will be affected by these types of future changes.

In this study, we develop an approach to characterize the limitations of ML upscaling and forecasting induced by limited data availability and climate-induced changes to ecosystem processes (figure 1). We use outputs from *ecosys*, a process-rich ecosystem model that has been extensively tested against EC fluxes under diverse arctic conditions (Grant *et al* 2009, 2015, Chang *et al* 2019, Mekonnen *et al* 2019, Riley *et al* 2021, Shirley *et al* 2022), to train and evaluate the performance of boosted regression tree (BRT) ML models across Alaska. We first examine the impact of variation in spatial and temporal training data coverage on ML model predictions of microbial respiration (R_h), net primary productivity (NPP), and net ecosystem exchange (NEE). Then, we evaluate the ability of the highest performing ML model to forecast the response of these C fluxes to climate change throughout the 21st century. Finally, we use convergent cross-mapping (CCM) to identify and rank the drivers of ML model bias in R_h and NPP at the end of the century. Since the internal complexity of a process-model is much smaller than that of real-world ecosystem processes and the simulated data is not affected by noise or measurement error, the performance of these ML models is significantly better than if they were trained on real-world datasets. However, this exercise allows us to estimate the 'best-case' performance of ML models used to upscale and forecast high-latitude C balances.

2. Data and methods

2.1. The *ecosys* model

Ecosys is a process-rich, mechanistic, hourly time-step ecosystem model with coupled C, energy, nutrient, and water cycles. The model is forced with hourly meteorological inputs, and calculates energy balance

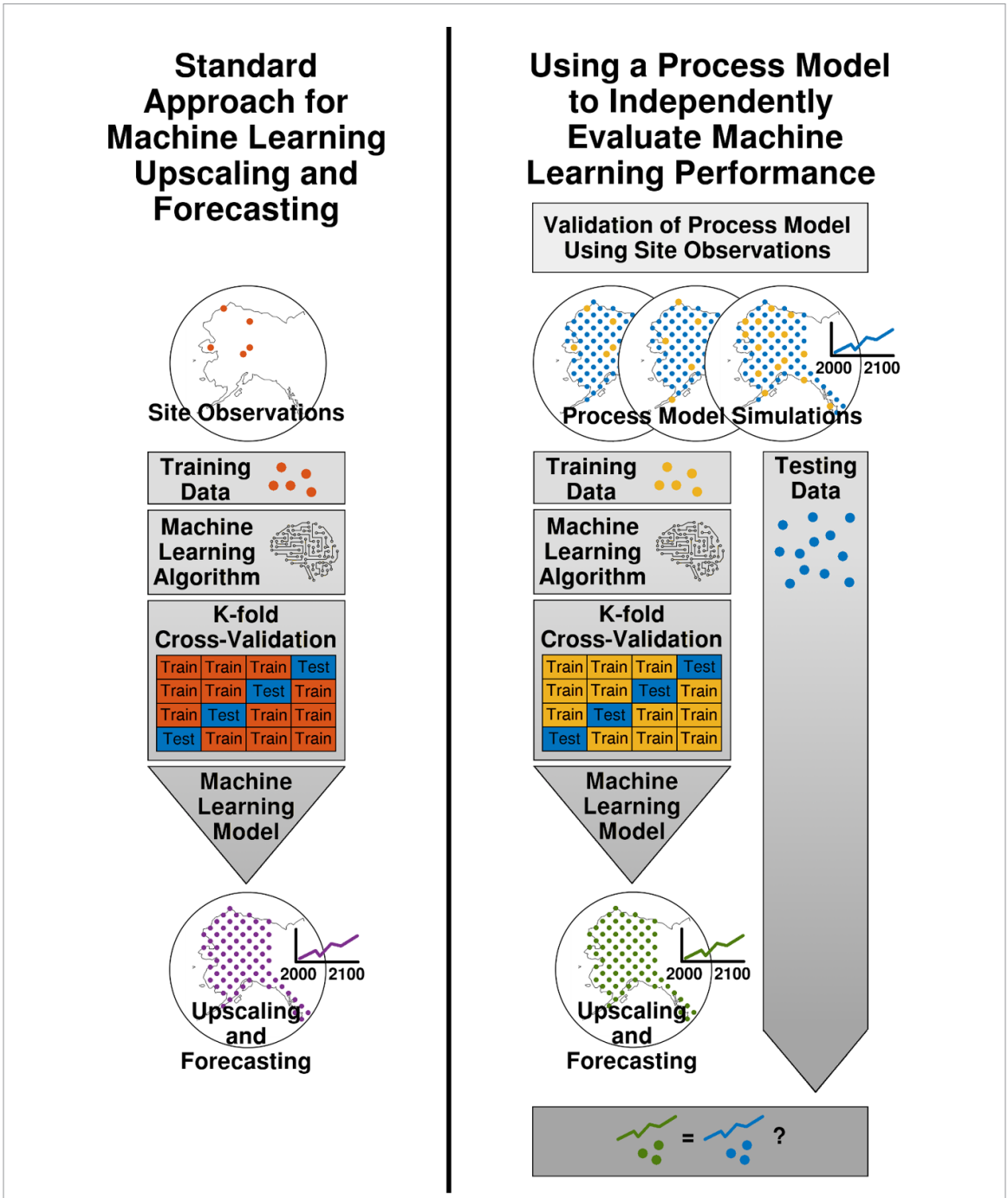

Figure 1. Regional process model simulations can be used to independently evaluate ML upscaling and forecasting. The standard methodology to upscale and forecast site observations of ecosystem processes is outlined in the left panel. Our approach to independently evaluate ML performance is shown on the right. In this approach, a process model that has been validated using the available site observations is used to simulate current and future ecosystem processes across a region of interest. Different subsets of the simulated data are then used to generate ML predictions across space and time that can be independently evaluated using the remaining simulated data.

using first-order closure schemes. Five Alaska plant functional types (PFTs; deciduous shrub, evergreen shrub, sedge, moss, lichen) compete in multi-layer canopy and soil profiles for light, water, and nutrients. Each PFT fixes CO_2 according to the Farquhar biogeochemical growth model (Farquhar *et al* 1980). Canopy stomatal conductance optimizes chloroplast CO_2 concentration unless limited by water availability. In the soil and litter layers, oxidation of dissolved organic C, which is produced via hydrolysis of organic matter pools (woody litter, non-woody litter, particulate organic matter, humus, microbial biomass), drives microbial growth and metabolism. Plant nutrient availability is regulated by stoichiometry of microbial biomass and soil organic matter. A full description of algorithms, and parameters used in *ecosys* can be found in the supplementary material of Mekonnen *et al* (2019).

The model is forced using hourly air temperature, precipitation, incoming shortwave radiation, relative

humidity, and wind speed from the North American Regional Reanalysis (Wei *et al* 2014) across a 0.25° × 0.25° grid that covers Alaska. Future climate anomalies for years 2018–2100 were taken from the CCSM4 climate model forced by Representative Concentration Pathway 8.5 (RCP8.5). Initial soil properties for each grid cell were taken from the Unified North America Soil Map (Liu *et al* 2013) and the Northern Circumpolar Soil Carbon Database (Hugelius *et al* 2013). The frequency of stand-replacing fire events under current climate conditions was calculated from the LANDFIRE product (Rollins 2009), and increased throughout the 21st century by a fixed amount that relates changes in environmental conditions under the RCP8.5 climate scenario with increases in lightning ignition as described in Veraverbeke *et al* (2017).

Ecosys has been extensively tested at many sites and against remote sensed observations across high-latitude ecosystems. A detailed description of *ecosys* performance at site and regional scales can be found in the supplementary material of Shirley *et al* (2022). Recently, *ecosys* has also been shown to accurately represent site observations of NEE at eight of the EC towers located in Alaska (figure S1, Shirley *et al* 2022).

2.2. BRTs and *k*CV

BRT ML models are linear combinations of decision trees that have been iteratively fit to reduce a loss function. BRT models are able to capture non-linear response curves and interaction effects between variables (Elith *et al* 2008) and were used to upscale C fluxes in Natali *et al* (2019) and Virkkala *et al* (2021). We implemented our BRT models in R using the 'gbm' package (Greenwell *et al* 2020) with a Gaussian error distribution, bag fraction of 0.5, tree complexity of 5, and a starting learning rate of 0.1 which was switched to 0.3 if the optimal tree number exceeded 10 000. The optimal tree number was identified using *k*CV. In *k*CV, the training dataset is first split into *k* groups (here, ten groups are used). Then, *k* different ML models are developed leaving out one group as a test set. Performance metrics (deviance, correlation coefficients) are calculated using the test set for each model, and summary statistics are calculated from the mean performance of the *k* models. The summary statistics are used to select the optimal tree number, and are also used to estimate ML predictive performance on independent datasets. We explored the sensitivity of ML model performance to the tree complexity and learning rate hyperparameters, and found that the impact of these parameters on model performance is smaller than the impact of improvements in spatial coverage (figure S2).

2.3. Convergence cross-mapping

Convergence cross-mapping (CCM) is a method of causal inference that uses nonlinear state space reconstruction to test for causation in weakly coupled dynamical systems (Sugihara *et al* 2012). CCM has been used to infer causal relationships in several climate, earth science, and ecological studies (Sugihara *et al* 2012, van Nes *et al* 2015, Wang *et al* 2018, Díaz *et al* 2022). In CCM, reconstructed state space of one variable is used to predict the reconstructed state space of a second variable. Cross-map skill is the correlation between the predicted and actual reconstructed state space, with high skill indicating a strong causative effect.

2.4. Data coverage and experimental design

In this study, we create BRT models to upscale and forecast ten-daily R_h, NPP, and NEE simulated by *ecosys* across Alaska at 0.25° resolution. The predictor variables for the BRT models (table 1) were chosen to represent a realistic upscaling effort that might be attempted with current publicly available data.

Alaska is one of the most well-studied high-latitude regions, but spatial coverage of AmeriFLUX towers is low. All 25 Alaska AmeriFLUX towers active between years 2010 and 2019 are located in 15 of the 4319 *ecosys* 0.25° × 0.25° grid cells, and are not optimally distributed to capture spatial and temporal variability. We used spatially constrained *k*-means clustering of simulated environmental data (SKATER algorithm; AssunÇão *et al* 2006) to define 15 regions in Alaska and found that AmeriFLUX towers occupy only 6 of these regions. The number of active Alaska AmeriFLUX sites increased throughout the decade from 10 sites in 2010 to 18 sites in 2019, but year-round data coverage is not available at all sites. Coverage is high during June and July but drops to less than 50% in winter (figure 2).

We trained BRT models of each C flux using seven training data configurations to evaluate how training dataset spatial and temporal coverage impacts ML model performance. The AmeriFLUX (AF) training dataset is intended to represent the current availability of EC flux tower data in Alaska. For this dataset, *ecosys* outputs were included in the training dataset when an AmeriFLUX tower was active within that gridcell. The AmeriFLUX full coverage (AFfc) training dataset includes *ecosys* outputs for each of the 15 gridcells that contain an AmeriFLUX tower but assumes continuous data coverage for years 2010–2019. The km15 dataset optimizes the locations of 15 training gridcells by including the centroid gridcell of the 15 regions identified using spatially constrained *k*-means clustering (figure 2). Following Hoffman *et al* (2013), we chose the centroids by minimizing the distance between gridcells in each cluster and the cluster centroid. For each of the four remaining training datasets, we doubled the number of gridcells included in the training dataset, choosing the locations using *k*-means clustering as above. These training datasets are further described in table S1. We

Table 1. Variables used to train the ML model. Name, description, units, and timescale are given for each variable used to train the ML models.

Name	Description	Units	Timescale	Data source
Air T	Daily mean air temperature	°C	10-Daily	NARR (Wei *et al* 2014)
SW_{in}	Daily mean incoming shortwave radiation	W m^{-2}	10-Daily	NARR (Wei *et al* 2014)
Soil T—10 cm	Daily mean soil temperature at 10 cm	°C	10-Daily	*ecosys* output
Soil T—2 m	Daily mean soil temperature at 2 m	°C	10-Daily	*ecosys* output
SSM	Surface soil moisture; Daily mean soil water content 0–5 cm	m^3 m^{-3}	10-Daily	*ecosys* output
RZSM	Root-zone soil moisture; Daily mean soil water content 0–100 cm	m^3 m^{-3}	10-Daily	*ecosys* output
SOC	Soil organic carbon in top 30 cm of soil	gC m^{-2}	Yearly	*ecosys* output
LAI	Ecosystem leaf area index	—	10-Daily	*ecosys* output
Sand	Soil sand content	kg kg^{-1}	Fixed	Unified North America Soil Map (Liu *et al* 2013)
Silt	Soil silt content	kg kg^{-1}	Fixed	Unified North America Soil Map (Liu *et al* 2013)
Tundra/boreal	Tundra or boreal region	—	Fixed	

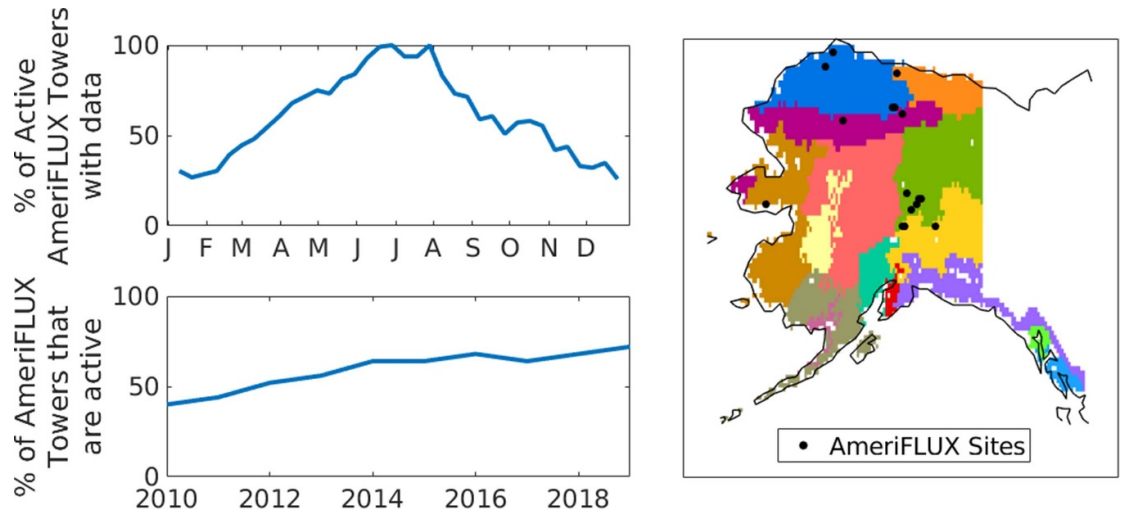

Figure 2. Alaskan AmeriFLUX sites exhibit uneven spatial and temporal coverage. (top left) Temporal coverage of active AmeriFLUX sites throughout the year. (bottom left) Percent of Alaskan AmeriFLUX sites that are active in each year. (right) Locations of 15 Alaskan AmeriFLUX sites with data for years 2010–2019 (black dots). Ecoregions identified by spatially constrained k-means clustering of simulated environmental data (SKATER algorithm) are shown in randomly assigned colors.

used each model to extrapolate C fluxes across Alaska, and excluded gridcells used in the training dataset for evaluation of model performance.

We then evaluated the ability of ML models to forecast C fluxes. The model trained on the largest dataset (km240) was used to predict Alaska R_h, NPP, and NEE throughout the 21st century. Sources of bias between the ML model predictions and *ecosys* simulations were evaluated using CCM. We explored the extent to which changes in atmospheric CO_2 concentration, air and soil temperature, deciduous shrub NPP dominance, and fire introduced bias into ML predictions of NPP, and the extent to which changes in litter C, soil temperature, deciduous shrub NPP dominance, and fire introduced bias into ML predictions of R_h.

3. Results and discussion

3.1. ML model trained with current AmeriFLUX availability and coverage predicts incorrect sign of present-day Alaska net C exchange

When data is limited, kCV is typically used to evaluate ML model performance. Here we find that the predictive power of the ML models trained using existing AmeriFLUX sites and coverage (AF) appears to be excellent when evaluated using kCV. kCV correlation coefficients for each C flux (R_h: $r = 0.95$; NPP: $r = 0.92$; NEE: $r = 0.86$) are much higher than estimated in ML C flux upscaling studies that use real data (e.g. Natali *et al* (2019)—NEE, non-growing season only: $r = 0.75$; Virkkala *et al* (2021)—NEE: $r = 0.27$; Tramontana *et al* (2016)—NEE: $r = 0.68$). Our ML

model performs better than ML models trained on real data because training data used in this study are generated by *ecosys*, which, while complex and process-rich, is simpler than real ecosystems and is free from noise, bias, and data gaps. When outputs from the entire modeled domain are used for validation the apparent performance of the AF model decreases significantly. Correlation coefficients for each flux are substantially lower (R_h: $r = 0.75$; NPP: $r = 0.77$; NEE: $r = 0.68$), and mean absolute bias (MAB) is high relative to the Alaska mean for each C flux (figure 3). This result demonstrates that *k*CV methods can give unreliably high confidence in the performance of ML models used to upscale or spatially extrapolate outside of the training set.

When extrapolated across Alaska, the AF model incorrectly predicts the sign of present-day net C exchange simulated by *ecosys*. Whereas Alaska is a slight sink of C according to *ecosys* (-19.7 gC m^2 yr^{-1}), the AF model predicts that it is a strong source of C (60.7 gC m^2 yr^{-1}). The AF prediction of Alaska means R_h is close to the target value and MAB for R_h is much lower than for NPP with this training data. These results imply that the mismatch between *ecosys* and AF predictions of present-day Alaska NEE is primarily due to large underestimation of plant productivity, particularly throughout southern Alaska (figure S3).

3.2. Increased spatial coverage of training data improves ML predictions of present-day Alaska net C exchange

AmeriFLUX coverage in Alaska is both spatially and temporally incomplete, but ML model predictions of Alaska net C exchange improve more via training site redistribution than increased temporal coverage. Surprisingly, NPP and NEE MAB increase when the full time series from each AmeriFLUX site is included in the training data, and the ML model predicts that Alaska is an even stronger net C source (figure 3). On the other hand, when training data is taken from the 15 identified eco-region centroids, instead of from AmeriFLUX sites, NPP and NEE MAB decrease significantly, correlation coefficients increase, and the ML model correctly predicts that Alaska is a slight net sink of C (figure 3).

As the number of training data sites increases from 15 to 240, ML model performance improves and correlation coefficients for each flux increase and approach the values estimated by *k*CV. MAB for the model trained with 15 sites is nearly twice as large as for the model trained with 240 sites for each flux (figure 3). Additionally, the spatial distribution of bias changes as more sites are added. With fewer sites, the bias is regional with spatially coherent patches that either overestimate or underestimate C fluxes. As the number of sites increases, positive and negative biases decrease and become more evenly distributed across Alaska (figure 4). Finally, the ML model distribution for each flux converges on the target distribution simulated by *ecosys* as the number of sites increases above 100 (figure 3).

3.3. ML model performance degrades throughout the century

In this section we investigate how the performance of the ML model trained on 240 sites (km240) changes throughout the 21st century. We chose this model because it has the largest training dataset and has excellent performance under current climate conditions. For the years 2010–2019, ML estimates of Alaska mean annual C fluxes agree very well with *ecosys* (R_h: 197.6 vs 194.4 gC m^2 yr^{-1}; NPP: 214.3 vs 214.0 gC m^2 yr^{-1}; NEE: -19.6 vs -19.9 gC m^2 yr^{-1}) and the model successfully captures the seasonal dynamics of each C flux (figure S3). Further, the relative influence of each training variable for the ML models is consistent with expectations and processes included in *ecosys*. The models for NPP and NEE depend most strongly on air and soil temperature, radiation, and leaf area index (LAI), whereas the model for R_h is dependent primarily on soil temperature and LAI (which can be interpreted as a proxy for plant litter C; figure S4). Two hundred and forty sites across Alaska is extremely optimistic since installation and maintenance of EC flux towers is expensive. However, we use this ideal scenario to set an upper boundary on the ability of ML models trained with real-world data to forecast ecosystem responses to climate change.

We find that the km240 ML model incorrectly predicts changes in ecosystem processes induced by climate change over the 21st century under an RCP8.5 scenario and that its performance degrades compared to the underlying model on which it was based. The end-of-century fluxes predicted by *ecosys* are large but realistic (∼300–650 gC m^{-2} yr^{-1}). Even larger NPP values have been measured under current climate conditions in Pacific Northwest and British Columbia forests (Jassal *et al* 2007, Schwalm *et al* 2007). *Ecosys* projected a two-fold increase in NEE MAB, a three-fold increase in R_h MAB, and a four-fold increase in NPP MAB by year 2100 (figure S5). ML model predictions of annual mean Alaska R_h are reasonably similar to *ecosys* through year 2040, but by the end of the century are underestimated by 104 gC m^2 yr^{-1}. The performance of the NPP ML model is even worse: its estimates of annual mean Alaska NPP diverge from *ecosys* by year 2030, leading to an underestimation of 204 gC m^2 yr^{-1} at the end of the century. For both R_h and NPP, the largest increase in bias occurs during spring (figure S3). While the ML model is able to capture an increase in end-of-century growing season

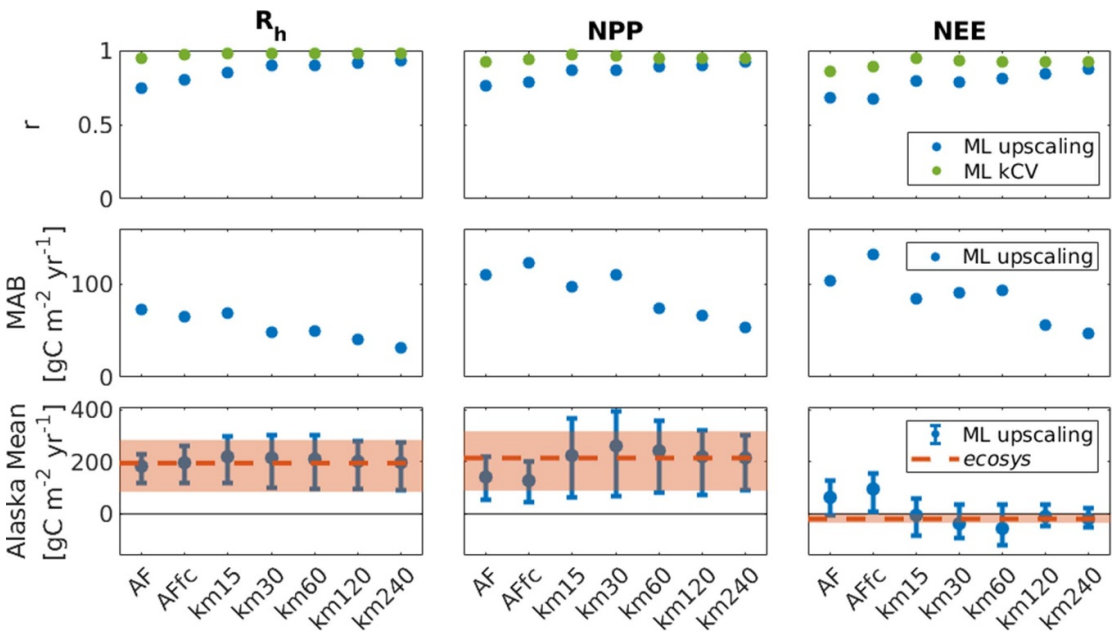

Figure 3. ML estimation of *ecosys* carbon fluxes improves with increased spatial coverage of training data. Performance of ML predictions is evaluated for R_h (left column), NPP (middle column), and NEE (right column). Correlation coefficients (top row) are shown as evaluated by comparison of ML upscaling to *ecosys* simulations (blue) and *k*CV using the training dataset (green). Mean absolute bias (MAB) for each flux is shown in the middle row (blue). C fluxes averaged across Alaska (bottom row) are shown for ML upscaling (blue) and *ecosys* simulations (target, red dashed lines). The 25th–75th percentile range of Alaska C fluxes is also shown for the ML upscaling (error bars) and *ecosys* simulations (shaded regions). Descriptions of the training datasets used for each model can be found in table S1.

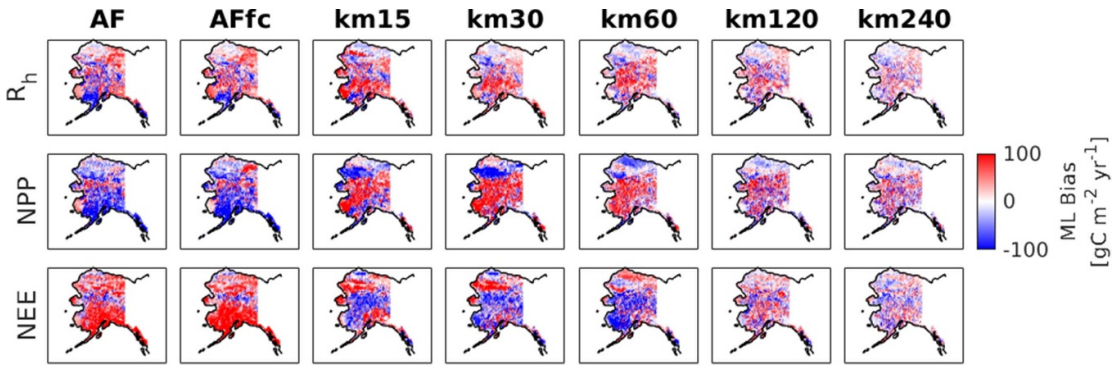

Figure 4. ML model bias decreases and becomes less clustered with increased spatial coverage of training data. Biases in R_h (top row), NPP (middle row), and NEE (bottom row) relative to *ecosys* simulations are shown across Alaska for ML models trained with different *ecosys* datasets. Descriptions of the training datasets used for each model can be found in table S1.

length (because it was trained on ten-daily data), it is unable to capture increases in peak C fluxes during the growing season (figure S3).

This degradation of ML model performance leads to a striking discrepancy between *ecosys* and ML model estimates of end-of-century Alaska net C exchange. Whereas *ecosys* predicts that Alaska C sink strength will steadily increase throughout the century, the km240 ML model predicts that Alaska will be net C neutral from mid-century and onwards (figure 5) and the AF ML model predicts that Alaska will remain a C sink until 2100 (figure S6). The km240 ML model predicts that 46% of grid cells will be C sources in the 2090s, whereas only 7% of *ecosys* grid cells are identified as sources. This result demonstrates that even an ideal ML model trained and evaluated on ideal data is unable to correctly predict the sign of net C exchange simulated by a process model in Alaska after a century of climate change.

3.4. ML model cannot predict ecosystem responses to changing CO_2 atmospheric concentrations and vegetation structure

Over the 21st century, climate change will induce large and complex changes to soil-plant-atmosphere interactions. Process models like *ecosys* are designed

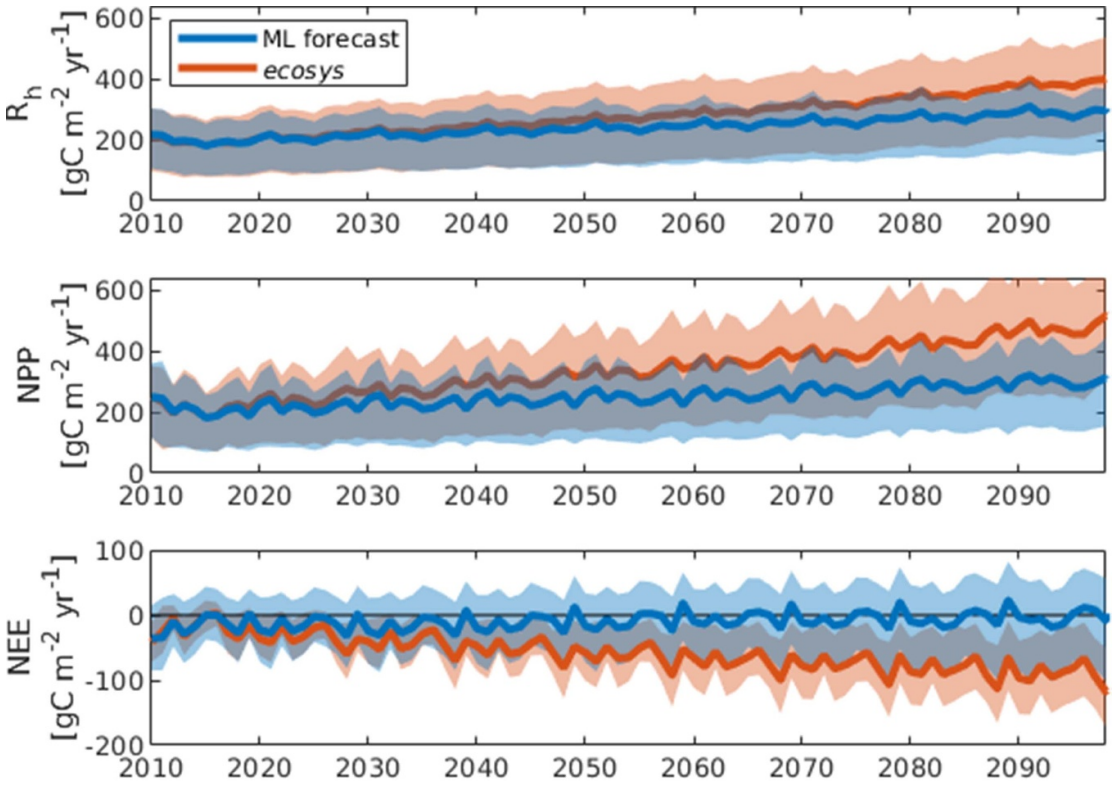

Figure 5. Performance of ML models degrades throughout the century. Alaska mean annual R_h, NPP, and NEE are shown for *ecosys* simulations (blue) and ML model trained on data from 240 sites (km240) for 2010–2019 (red). Shaded areas represent 25th–75th percentile ranges.

to simulate ecosystem responses to these changes, but ML models can only make predictions based on relationships contained in the training data. To quantify which 21st century changes in soil-plant-atmosphere interactions are primarily responsible for the observed degradation in ML model performance, we apply convergence cross-mapping (CCM), which uses state space reconstruction to quantify causal relationships between non-stationary and non-linear time series (Methods). Here, we explore the extent to which changes in atmospheric CO_2, litter inputs, vegetation composition, air and soil temperatures, and fire frequency generate biases in ML predictions of NPP and R_h.

We find that increasing atmospheric CO_2 is the primary driver of ML underestimation of 21st century NPP. Increases in ML NPP bias are strongly coupled to rising CO_2 in 86% of grid cells (figure 6). Elevated CO_2 increases photosynthetic fixation rates through changes to carboxylation, photorespiration, and water and nitrogen use efficiency (Ainsworth and Rogers 2007). These processes are represented in *ecosys*, and have been tested against data from free air CO_2 enrichment experiments (Grant 2013). Because training data is generated under present day CO_2 concentrations, ML techniques are unable to capture changes to the relationships between environmental variables and NPP under elevated atmospheric CO_2.

Since increases in NPP generate increases in plant litter, underestimation of litter C follows from the NPP underestimation by the ML model. Root and shoot litter is highly labile, so increases in litter C typically lead to increases in microbial respiration in field studies (Adamczyk *et al* 2020) and in the *ecosys* model (Grant *et al* 2020). Indeed, we find that litter C is the leading driver of ML underestimation of 21st century R_h, with increases in ML R_h bias strongly tied to increases in litter C in 40% of grid cells (figure 7). In *ecosys*, increases in R_h drive increases in symbiotic and non-symbiotic N_2 fixation that in turn support the large increases in NPP.

Shifting high-latitude vegetation composition is a secondary but strong control on both NPP and R_h ML biases. Increases in growth and abundance of deciduous trees and shrubs have been observed in high-latitude ecosystems in recent decades (Tape *et al* 2006). *Ecosys* projects that vegetation composition will continue to change throughout the 21st century, impacting C cycling via changes in phenology, litter quality, nutrient acquisition and partitioning, and surface energy budgets (Mekonnen et al 2019, 2021a). The inability of the ML model to capture these changes leads to underestimation of end-of-century NPP (R_h) in 30% (34%) of grid cells. We also find that warming air and soil temperatures and increasing fire frequency are weaker drivers of ML bias for both NPP and R_h (figures 6 and 7).

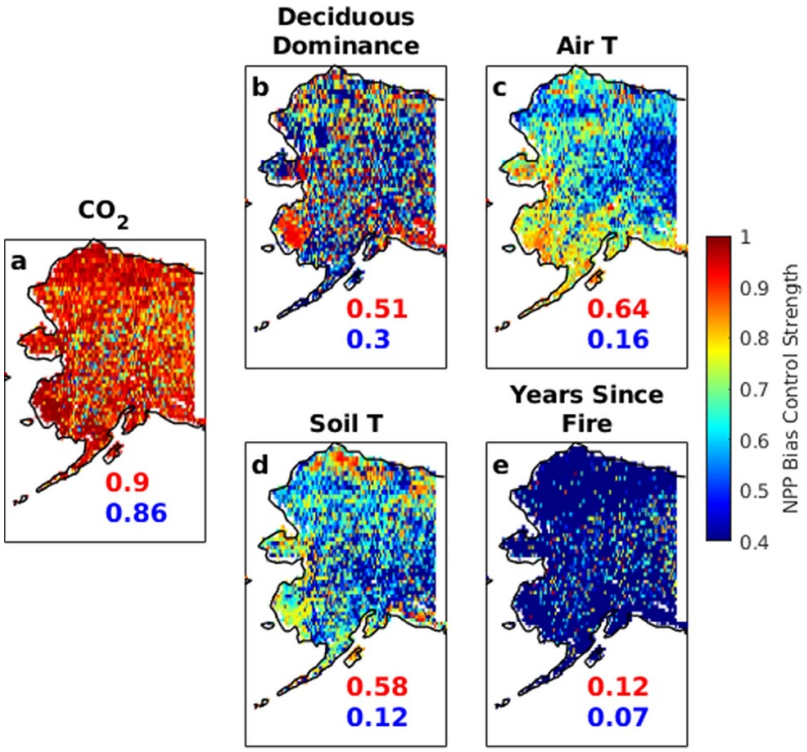

Figure 6. Atmospheric CO_2 concentration is the dominant control of bias in ML predictions of NPP throughout the century. The following drivers of ML model bias in NPP throughout the 21st century are identified and ranked using CCM: (a) atmospheric CO_2 concentration, (b) NPP dominance of deciduous PFT, (c) mean annual air temperature, (d) mean annual temperature in the top 25 cm of soil, (e) and years since the last fire. Mean bias control strength for each driver is shown in red. Drivers are ranked by the fraction of grid cells with bias control strength greater than 0.8 (shown in blue).

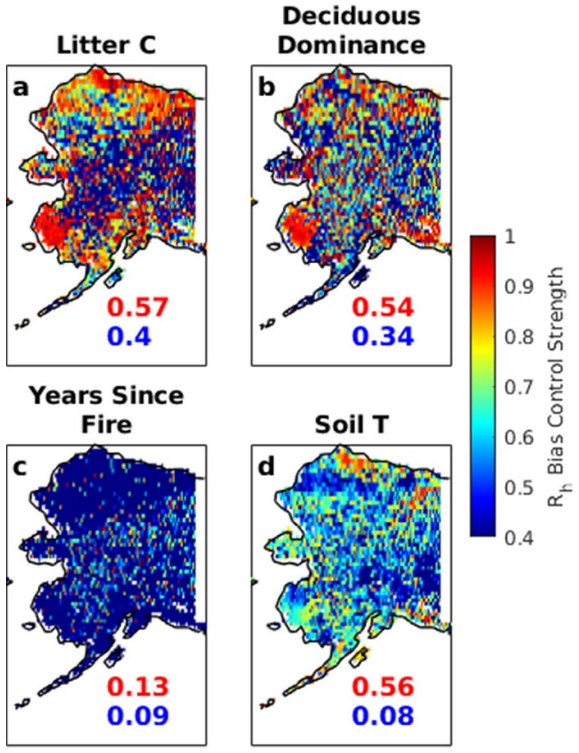

Figure 7. Plant litter carbon is the dominant control of bias in ML predictions of R_h throughout the century. The following drivers of ML model bias in R_h throughout the 21st century are identified and ranked using CCM: (a) shoot and root litter carbon, (b) NPP dominance of deciduous PFT, (c) years since the last fire, and (d) mean annual temperature in the top 25 cm of soil. Mean bias control strength for each driver is shown in red. Drivers are ranked by the fraction of grid cells with bias control strength greater than 0.8 (shown in blue).

4. Conclusion

We developed a method to independently evaluate ML upscaling and forecasting of ecosystem C cycle processes. We first used a well-tested mechanistic ecosystem model (*ecosys*) to simulate current and future R_h, NPP, and NEE at 0.25° resolution across Alaska. Different subsets of the *ecosys* gridcells were then used to train BRT ML models for these C fluxes, and the remaining gridcells were used to evaluate ML predictions under current and future climates. This approach represents a best-case ML model development scenario since (a) the process-model complexity, although substantial, is lower than real-world ecosystems and (b) the process model results used to train and test the ML model are free of noise, bias, and gaps.

We found that the ability of ML models to upscale and forecast high-latitude C fluxes is limited by data availability and future changes to ecosystem processes. Cross-validation methods, though widely used in ML applications, give poor indication of true predictive skill when training datasets do not provide adequate coverage of the prediction space. Regarding upscaling under current conditions, the ML model trained with *ecosys* simulations at existing AmeriFLUX sites predicts an opposite sign of the Alaska C balance. This result mirrors the current mismatch between ecosystem model and ML-based estimates of high-latitude C balances and suggests that sampling biases in generation of ML models, rather than incomplete ecosystem model process representation, may be at fault. We also found that increased spatial coverage of the training dataset (well beyond the current Alaska AmeriFLUX sites and beyond what is practical with current funding) significantly improves ML upscaling. Our results highlight the importance of intentional site selection for training data collection and that a substantial increase in optimally-located high-latitude EC flux tower site coverage is needed to produce accurate ML estimates of large-scale net C exchanges.

Regarding forecasting, we found that the ML model has poor predictive capability at multi-decadal scales. Even using the ML model trained with more than ten times the existing EC flux tower sites, changes in CO_2 concentrations, litter C, and vegetation structure that cannot be captured by the training data led to large Alaska C flux prediction biases under 21st century climate change. We therefore discourage using ML to upscale ecosystem processes and emphasize that changes in forcing and ecosystem properties can lead to inaccurate ML forecasting even under best-case conditions.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://ngee-arctic.ornl.gov/data/. Data will be available from 30 January 2023.

Acknowledgments

This research was supported by the Director, Office of Science, Office of Biological and Environmental Research of the US Department of Energy under Contract No. DE-AC02-05CH11231 to Lawrence Berkeley National Laboratory as part of the Next-Generation Ecosystem Experiments in the Arctic (NGEE-Arctic) project.

ORCID iDs

Ian A Shirley https://orcid.org/0000-0002-2229-1414
Zelalem A Mekonnen https://orcid.org/0000-0002-2647-0671
Robert F Grant https://orcid.org/0000-0002-8890-6231
Baptiste Dafflon https://orcid.org/0000-0001-9871-5650
William J Riley https://orcid.org/0000-0002-4615-2304

References

Abbasian H, Solgi E, Hosseini S M and Kia S H 2022 Modeling terrestrial net ecosystem exchange using machine learning techniques based on flux tower measurements *Ecol. Model.* **466** 109901

Adamczyk M, Perez-Mon C, Gunz S and Frey B 2020 Strong shifts in microbial community structure are associated with increased litter input rather than temperature in High Arctic soils *Soil Biol. Biochem.* **151** 108054

Ainsworth E A and Rogers A 2007 The response of photosynthesis and stomatal conductance to rising [CO_2]: mechanisms and environmental interactions *Plant Cell Environ.* **30** 258–70

Arjovsky M, Bottou L, Gulrajani I and Lopez-Paz D 2020 Invariant risk minimization (arXiv:1907.02893v3)

Arora B, Wainwright H M, Dwivedi D, Vaughn L J S, Curtis J B, Torn M S, Dafflon B and Hubbard S S 2019 Evaluating temporal controls on greenhouse gas (GHG) fluxes in an Arctic tundra environment: an entropy-based approach *Sci. Total Environ.* **649** 284–99

AssunÇão R M, Neves M C, Câmara G and Da Costa Freitas C 2006 Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees *Int. J. Geogr. Inf. Sci.* **20** 797–811

Bastin J-F, Finegold Y, Garcia C, Mollicone D, Rezende M, Routh D, Zohner C M and Crowther T W 2019 The global tree restoration potential *Science* **365** 76–79

Becker M S, Jonathan Davies T, Pollard W H and Cornelissen H 2016 Ground ice melt in the high Arctic leads to greater ecological heterogeneity *J. Ecol.* **104** 114–24

Belshe E F, Schuur E A G, Bolker B M and Bracho R 2012 Incorporating spatial heterogeneity created by permafrost thaw into a landscape carbon estimate *J. Geophys. Res.* **117** G01026

Cable W L, Romanovsky V E and Torre Jorgenson M 2016 Scaling-up permafrost thermal measurements in western Alaska using an ecotype approach *Cryosphere* **10** 2517–32

Chang K-Y, Riley W J, Crill P M, Grant R F, Rich V I and Saleska S R 2019 Large carbon cycle sensitivities to climate across a permafrost thaw gradient in subarctic Sweden *The Cryosphere* **13** 647–63

Commane R et al 2017 Carbon dioxide sources from Alaska driven by increasing early winter respiration from Arctic tundra Proc. Natl Acad. Sci. USA 114 5361–6
Crane-Droesch A 2018 Machine learning methods for crop yield prediction and climate change impact assessment in agriculture Environ. Res. Lett. 13 114003
D'Amour A et al 2020 Underspecification presents challenges for credibility in modern machine learning (arXiv:2011.03395)
Díaz E, Adsuara J E, Martínez Á M, Piles M and Camps-Valls G 2022 Inferring causal relations from observational long-term carbon and water fluxes records Sci. Rep. 12 1610
Elith J, Leathwick J R and Hastie T 2008 A working guide to boosted regression trees J. Anim. Ecol. 77 802–13
Farquhar G D, von Caemmerer S and Berry J A 1980 A biochemical model of photosynthetic CO_2 assimilation in leaves of C_3 species Planta 149 78–90
Finger R A, Turetsky M R, Kielland K, Ruess R W, Mack M C, Euskirchen E S and Wurzburger N 2016 Effects of permafrost thaw on nitrogen availability and plant–soil interactions in a boreal Alaskan lowland J. Ecol. 104 1542–54
Geirhos R, Jacobsen J-H, Michaelis C, Zemel R, Brendel W, Bethge M and Wichmann F A 2020 Shortcut learning in deep neural networks Nat. Mach. Intell. 2 665–73
Grant R F 2013 Modelling changes in nitrogen cycling to sustain increases in forest productivity under elevated atmospheric CO_2 and contrasting site conditions Biogeosciences 10 7703–21
Grant R F, Barr A G, Black T A, Margolis H A, Dunn A L, Metsaranta J, Wang S, McCaughey J H and Bourque C A 2009 Interannual variation in net ecosystem productivity of Canadian forests as affected by regional weather patterns—a fluxnet-Canada synthesis Agric. For. Meteorol. 149 2022–39
Grant R F, Dyck M, Puurveen D and Naeth M A 2020 Nitrogen and phosphorus control carbon sequestration in agricultural ecosystems: modelling carbon, nitrogen, and phosphorus balances at the Breton plots with ecosys under historical and future climates Can. J. Soil Sci. 100 408–29
Grant R F, Humphreys E R and Lafleur P M 2015 Ecosystem $CO_2$and CH_4 exchange in a mixed tundra and a fen within a hydrologically diverse Arctic landscape: 1. Modeling versus measurements J. Geophys. Res. 120 1366–87
Grant R F, Mekonnen Z A, Riley W J, Arora B and Torn M S 2017 Mathematical modelling of Arctic polygonal tundra with ecosys: 2. Microtopography determines how CO_2 and CH_4 exchange responds to changes in temperature and precipitation J. Geophys. Res. 122 3174–87
Greenwell B, Boehmke B, Cunningham J and Developers G B M 2020 gbm: generalized boosted regression models (R package) (available at: https://CRAN.R-project.org/package=gbm)
Hoffman F M, Kumar J, Mills R T and Hargrove W W 2013 Representativeness-based sampling network design for the state of Alaska Landsc. Ecol. 28 1567–86
Hugelius G et al 2013 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
Hugelius G et al 2014 Estimated stocks of circumpolar permafrost carbon with quantified uncertainty ranges and identified data gaps Biogeosciences 11 6573–93
Jafarov E E, Coon E T, Harp D R, Wilson C J, Painter S L, Atchley A L and Romanovsky V E 2018 Modeling the role of preferential snow accumulation in through talik development and hillslope groundwater flow in a transitional permafrost landscape Environ. Res. Lett. 13 105006
Jassal R S, Andrew Black T, Cai T, Morgenstern K, Li Z, Gaumont-Guay D and Nesic Z 2007 Components of ecosystem respiration and an estimate of net primary productivity of an intermediate-aged Douglas-fir stand Agric. For. Meteorol. 144 44–57
Jorgenson M T, Torre Jorgenson M, Romanovsky V, Harden J, Shur Y, O'Donnell J, Schuur E A G, Kanevskiy M and Marchenko S 2010 Resilience and vulnerability of permafrost to climate change. This article is one of a selection of papers from the dynamics of change in Alaska's boreal forests: resilience and vulnerability in response to climate warming Can. J. For. Res. 40 1219–36
Jung M et al 2020 Scaling carbon fluxes from eddy covariance sites to globe: synthesis and evaluation of the FLUXCOM approach Biogeosciences 17 1343–65
Keuper F, Bodegom P M, Dorrepaal E, Weedon J T, Hal J, Logtestijn R S P and Aerts R 2012 A frozen feast: thawing permafrost increases plant-available nitrogen in subarctic peatlands Glob. Change Biol. 18 1998–2007
Liu S, Wei Y, Post W M, Cook R B, Schaefer K and Thornton M M 2013 The unified North American soil map and its implication on the soil organic carbon stock in North America Biogeosciences 10 2915–30
McGuire A D et al 2012 An assessment of the carbon balance of Arctic tundra: comparisons among observations, process models, and atmospheric inversions Biogeosciences 9 3185–204
Mekonnen Z A et al 2021a Arctic tundra shrubification: a review of mechanisms and impacts on ecosystem carbon balance Environ. Res. Lett. 16 053001
Mekonnen Z A, Riley W J, Grant R F, Salmon V G, Iversen C M, Biraud S C, Breen A L and Lara M J 2021b Topographical controls on hillslope-scale hydrology drive shrub distributions on the Seward Peninsula, Alaska J. Geophys. Res. Biogeosci. 126 e2020JG005823
Mekonnen Z A, Riley W J, Randerson J T, Grant R F and Rogers B M 2019 Expansion of high-latitude deciduous forests driven by interactions between climate warming and fire Nat. Plants 5 952–8
Mishra U et al 2021 Spatial heterogeneity and environmental predictors of permafrost region soil organic carbon stocks Sci. Adv. 7 eaaz5236
Naidu D G T and Bagchi S 2021 Greening of the earth does not compensate for rising soil heterotrophic respiration under climate change Glob. Change Biol. 27 2029–38
Natali S M et al 2019 Large loss of CO in winter observed across the northern permafrost region Nat. Clim. Change 9 852–7
Nobrega S and Grogan P 2008 Landscape and ecosystem-level controls on net carbon dioxide exchange along a natural moisture gradient in Canadian low Arctic tundra Ecosystems 11 377–96
Pearson R G, Phillips S J, Loranty M M, Beck P S A, Damoulas T, Knight S J and Goetz S J 2013 Shifts in Arctic vegetation and associated feedbacks under climate change Nat. Clim. Change 3 673–7
Peltola O et al 2019 Monthly gridded data product of northern wetland methane emissions based on upscaling eddy covariance observations Earth Syst. Sci. Data 11 1263–89
Raudys S J and Jain A K 1991 Small sample size effects in statistical pattern recognition: recommendations for practitioners IEEE Trans. Pattern Anal. Mach. Intell. 13 252–64
Reitz O, Graf A, Schmidt M, Ketzler G and Leuchner M 2021 Upscaling net ecosystem exchange over heterogeneous landscapes with machine learning J. Geophys. Res. 126 e2020JG005814
Riley W J, Mekonnen Z A, Tang J, Zhu Q, Bouskill N J and Grant R F 2021 Non-growing season plant nutrient uptake controls Arctic tundra vegetation composition under future climate Environ. Res. Lett. 16 074047
Rollins M G 2009 LANDFIRE: a nationally consistent vegetation, wildland fire, and fuel assessment Int. J. Wildland Fire 18 235
Schwalm C R, Andrew Black T, Morgenstern K and Humphreys E R 2007 A method for deriving net primary productivity and component respiratory fluxes from tower-based eddy covariance data: a case study using a 17-year data record from a Douglas-fir chronosequence Glob. Change Biol. 13 370–85

Serreze M C, Barrett A P, Stroeve J C, Kindig D N and Holland M M 2009 The emergence of surface-based Arctic amplification *The Cryosphere* **3** 11–19

Shirley I A, Mekonnen Z A, Grant R F, Dafflon B, Hubbard S S and Riley W J 2022 Rapidly changing high-latitude seasonality: implications for the 21st century carbon cycle in Alaska *Environ. Res. Lett.* **17** 014032

Sugihara G, May R, Ye H, Hsieh C-H, Deyle E, Fogarty M and Munch S 2012 Detecting causality in complex ecosystems *Science* **338** 496–500

Tape K, Sturm M and Racine C 2006 The evidence for shrub expansion in Northern Alaska and the Pan-Arctic *Glob. Change Biol.* **12** 686–702

Tramontana G *et al* 2016 Predicting carbon dioxide and energy fluxes across global FLUXNET sites with regression algorithms *Biogeosciences* **13** 4291–313

Uhlemann S, Dafflon B, Peterson J, Ulrich C, Shirley I, Michail S and Hubbard S S 2021 Geophysical monitoring shows that spatial heterogeneity in thermohydrological dynamics reshapes a transitional permafrost system *Geophys. Res. Lett.* **48** e2020GL091149

van Nes E H, Scheffer M, Brovkin V, Lenton T M, Ye H, Deyle E and Sugihara G 2015 Causal feedbacks in climate change *Nat. Clim. Change* **5** 445–8

Veraverbeke S, Rogers B M, Goulden M L, Jandt R R, Miller C E, Wiggins E B and Randerson J T 2017 Lightning as a major driver of recent large fire years in North American boreal forests *Nat. Clim. Change* **7** 529–34

Virkkala A-M *et al* 2021 Statistical upscaling of ecosystem CO fluxes across the terrestrial tundra and boreal domain: regional patterns and uncertainties *Glob. Change Biol.* **27** 4040–59

Waldrop M P *et al* 2021 Carbon fluxes and microbial activities from boreal peatlands experiencing permafrost thaw *J. Geophys. Res.* **126** e2020JG005869

Wang Y, Yang J, Chen Y, De Maeyer P, Li Z and Duan W 2018 Detecting the causal effect of soil moisture on precipitation using convergent cross mapping *Sci. Rep.* **8** 12171

Wei Y *et al* 2014 The North American carbon program multi-scale synthesis and terrestrial model intercomparison project—part 2: environmental driver data *Geosci. Model Dev.* **7** 2875–93

Zeng J, Matsunaga T, Tan Z-H, Saigusa N, Shirai T, Tang Y, Peng S and Fukuda Y 2020 Global terrestrial carbon fluxes of 1999–2019 estimated by upscaling eddy covariance data with a random forest *Sci. Data* **7** 1–11