Check for updates

JGR Biogeosciences

RESEARCH ARTICLE
10.1029/2025JG009307

Informing Robust Functional Relationship Benchmarks: An Evaluation of the Temperature Sensitivity of Ecosystem Respiration Across the Arctic-Boreal Region

Jeralyn Poe[1], Deborah Huntzinger[1,2,3], William J. Riley[4], Jon M. Wells[5], Edward A. G. Schuur[3], Christopher Schwalm[6], Logan T. Berner[1], Heidi Rodenhizer[6], Nicholas J. Bouskill[7], Victor Brovkin[8], Eleanor J. Burke[9], Philippe Ciais[10], Goran Georgievski[8,11,12], Adrian Gustafson[13], David M. Lawrence[14], Andrew H. MacDougall[15], Zelalem A. Mekonnen[4], Joe R. Melton[16], Gesa Meyer[16], Alexandra Pongracz[13], Chunjing Qiu[17], Benjamin N. Sulman[18], Sean C. Swenson[14], Jing Tao[4], David Wårlind[13], Yi Xi[10], Fengming Yuan[18], Qing Zhu[4], and Christina Schädel[6]

[1]School of Informatics, Computing, and Cyber Systems, Northern Arizona University, Flagstaff, AZ, USA, [2]School of Earth and Sustainability, Northern Arizona University, Flagstaff, AZ, USA, [3]Center for Ecosystem Science and Society, Northern Arizona University, Flagstaff, AZ, USA, [4]Climate and Ecosystem Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA, [5]Department of Soil and Environmental Sciences, University of Wisconsin-Madison, Madison, WI, USA, [6]Woodwell Climate Research Center, Falmouth, MA, USA, [7]Department of Biological and Ecological Engineering, Oregon State University, Corvallis, OR, USA, [8]Department of Climate Dynamics, Max Planck Institute for Meteorology, Hamburg, Germany, [9]Met Office Hadley Centre, Exeter, UK, [10]Laboratoire des Sciences du Climat et de l'Environnement, CEA-CNRS-UVSQ, Université Paris-Saclay, Gif-sur-Yvette, France, [11]Earth Resilience Science Unit, Potsdam Institute for Climate Impact Research (PIK), Member of the Leibniz Association, Potsdam, Germany, [12]Department of Integrative Earth System Science, Max Planck Institute of Geoanthropology, Jena, Germany, [13]Department of Earth and Environmental Sciences, Lund University, Lund, Sweden, [14]CGD, NSF National Center for Atmospheric Research, Boulder, CO, USA, [15]Department of Earth & Environmental Science, St. Francis Xavier University, Antigonish, NS, Canada, [16]Climate Research Division, Environment and Climate Change Canada, Victoria, BC, Canada, [17]Research Center for Global Change and Ecological Forecasting, School of Ecological and Environmental Sciences, Institute of Eco-Chongming, East China Normal University, Shanghai, China, [18]Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA

Key Points:

- The range in inferred model skill is substantially impacted by the choices applied in constructing functional relationship benchmarks
- The inferred performance of a given model is most sensitive to the number of daily observations and temporal extent of the benchmark data set
- These results can guide the development of consistent and robust functional relationships for future model evaluation studies

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

J. Poe,
jmp838@nau.edu

Citation:

Poe, J., Huntzinger, D., Riley, W. J., Wells, J. M., Schuur, E. A. G., Schwalm, C., et al. (2026). Informing robust functional relationship benchmarks: An evaluation of the temperature sensitivity of ecosystem respiration across the Arctic-Boreal region. *Journal of Geophysical Research: Biogeosciences*, *131*, e2025JG009307. https://doi.org/10.1029/2025JG009307

Received 28 JUL 2025
Accepted 17 APR 2026

© 2026. The Author(s).
This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

Abstract

During land model development, simulated carbon dynamics are often benchmarked against observational data sets to evaluate model performance. Functional relationship benchmarks are the relationship between a driving variable (e.g., temperature) and a response variable (e.g., ecosystem respiration) and are a promising tool for assessing model performance by evaluating modeled sensitivities to changing environmental conditions. However, observed functional relationships can be influenced by choices made during data collection and throughout the benchmarking process, impacting the inferred skill of land models. To avoid misrepresenting a model's true performance, it is necessary to systematically evaluate best practices when constructing functional relationship benchmarks. We developed a set of guidelines for constructing functional relationship benchmarks, considering the choice of data set, number of daily observations, temporal extent, and temporal resolution across Alaska and Canada over a 20-year period from 2001 to 2020. The temperature sensitivity of ecosystem respiration from observations, evaluated through an apparent Q_{10}, is highly variable both spatially and as a result of the data processing approach applied in the benchmark formation. When benchmarking 13 models from the Warming Permafrost Model Intercomparison Project (WrPMIP), the range in inferred model skill is substantially impacted by the choices applied in constructing functional relationship benchmarks. The inferred performance of a given model is most sensitive to the number of daily observations and temporal extent, followed by choice of benchmark data set and temporal averaging. Results from this analysis can guide the development of consistent and robust functional relationships for future model evaluation studies.

Plain Language Summary

Land models are often compared against data sets to evaluate model performance, referred to as model benchmarking. However, data sets themselves can contain errors and vary based on the measurement approach, and choices made during the processing of data can influence benchmarks derived from observational data products. Here, we evaluated how ecosystem respiration responds to temperature across Alaska and Canada, and how this relationship can be influenced by the choice of data set, number of data points throughout the month, number of years, and averaging over time. We found that the

relationship between ecosystem respiration and temperature varies widely according to these choices, particularly the number of daily observations and temporal extent. When applying different subsets of the data to test model performance, we found that these choices produced a large range in model scores compared with those when applying the full data set. As a result, we developed a set of best practices that can serve as a guide during model benchmarking and can provide a more accurate sense of model performance.

1. Introduction

Future estimates of ecosystem carbon storage as a result of increasing temperature and carbon dioxide (CO_2) concentrations depend on reliable projections from process-based land models (Fisher et al., 2014; Piao et al., 2013; Sitch et al., 2015). Producing reliable model projections is particularly important in the Arctic-Boreal region, where temperatures are rising at a faster rate than the rest of the globe (Rantanen et al., 2022). Rapid warming is driving permafrost thaw and subsequent decomposition of previously frozen soil organic carbon, which is releasing additional CO_2 and methane (CH_4) into the atmosphere (Schuur et al., 2022; Xue et al., 2016). This additional CO_2 and CH_4 released from permafrost thaw could cause additional warming through the permafrost carbon climate feedback (MacDougall et al., 2012; Schaefer et al., 2014; Schuur et al., 2008, 2015). However, land models often perform poorly at capturing observed permafrost carbon dynamics due to the heterogeneous nature of permafrost degradation and complex interdependent processes that both impact and are impacted by permafrost thaw (Burke et al., 2020; McGuire et al., 2018; Schädel et al., 2024). Evaluating and ultimately improving model performance of simulated carbon cycle processes, and how modeled carbon dynamics respond to changes in environmental conditions can be supported by benchmarking models with observational data products, referred to as model benchmarking (Fisher et al., 2018).

During model benchmarking, assessing simulated carbon fluxes such as ecosystem respiration (Reco) against observations can help establish whether or not a model accurately captures the magnitude and seasonality of fluxes. However, such comparisons do not indicate whether the model includes an accurate representation of processes (e.g., correct flux sensitivity to warming; Huntzinger et al., 2020; Luo et al., 2012; Schwalm et al., 2019). For example, a model could arrive at the correct magnitude of net carbon flux for a given site due to compensating errors in carbon fixation and Reco. Another model benchmarking approach is to evaluate how modeled fluxes respond to environmental forcing factors or simulated variables within the model. These types of benchmarks are referred to as functional relationships or functional benchmarks (Luo et al., 2015; Randerson et al., 2009). Functional benchmarks are constructed by developing an empirical relationship between two variables, such as Reco as a function of temperature or precipitation, or photosynthesis rate as a function of light levels (Collier et al., 2018). Applying functional relationships as benchmarks can help link model structure, such as process representation and parameterization, to model skill and go beyond what can be learned from only evaluating modeled flux or stock magnitudes (Huntzinger et al., 2020). For example, functional relationships can be used to constrain simulated carbon dynamics by revealing mismatches in modeled sensitivities of carbon fluxes to changing environmental conditions (McGuire et al., 2016).

While benchmarking through functional relationships can help identify model strengths and deficiencies regarding specific simulated processes (Bouskill et al., 2014; Schwalm et al., 2019), prior research has demonstrated that choices made in the implementation of benchmarks can influence a model's inferred performance. For example, the choice of data set used to construct benchmarks can have more impact on inferred model skill than the choice of model itself (Schwalm et al., 2013). In addition, when constructing functional relationships, the spatial and temporal resolution of the observational data as well as averaging methods applied prior to calculating the relationships can strongly influence the shape of the response (Poe et al., 2025). However, in situ data can be challenging to consistently collect, often leaving temporal data gaps and irregular collection periods for entire seasons, particularly in remote environments such as the Arctic-Boreal region (Berner et al., 2024; Pallandt et al., 2022). This, in turn, can potentially impact the information content of the resulting relationships. Therefore, the choices made in the benchmarking process and the impact of these choices, both on the information content of the benchmark and on inferences of model skill or performance, require further evaluation.

One functional relationship that can be particularly informative for model development of permafrost carbon dynamics is the temperature sensitivity of Reco. In the Arctic-Boreal region, the response of Reco to temperature

21698961, 2026, 5, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025JG009307 by Helmholtz-Zentrum Potsdam GFZ, Wiley Online Library on [29/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

21698961, 2026, 5, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025JG009307 by Helmholtz-Zentrum Potsdam GFZ, Wiley Online Library on [29/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

impacts the amount of CO_2 that will be released from permafrost ecosystems as they warm (Torn et al., 2025), and needs to be accurately captured in land models to predict when and how much carbon will be released from permafrost thaw under a warming climate (Koven et al., 2017). Since Reco includes heterotrophic respiration (Rh) and above and below ground autotrophic respiration (Ra), its temperature sensitivity can be confounded by multiple above and below ground factors (Qin et al., 2021). Soil respiration can be directly measured through dark or opaque chambers that eliminate photosynthesis; however, soil respiration measurements are often lacking in high latitude regions (Hursh et al., 2017). More common measurement methods include eddy covariance flux towers and transparent chambers that directly measure net ecosystem exchange (NEE), from which Reco can then be partitioned. The partitioning method also results in an observation-based product rather than a direct measurement (Zhan et al., 2022), therefore introducing another layer of uncertainty into benchmarks derived from these products. Specifically, some common eddy covariance partitioning measurements impose a functional temperature response relationship on the estimated respiration (Stoy et al., 2006). Upscaling and machine learning methods are also commonly used to derive Reco and its components from site-level observations that better match the spatial scales of land models (e.g., Virkkala et al., 2024); however, these are geostatistical upscaled products with inherent uncertainties (Hashimoto et al., 2015; Tang et al., 2019). The various spatial scales at which Reco is estimated (i.e., chamber, flux tower, and upscaled) further impact the inferred functional relationship depending on the data product applied (Reichstein & Beer, 2008).

To simulate the relationship between respiration and temperature, land models often scale decomposition using an Arrhenius-based model or, more commonly, a Q_{10} function (Sun et al., 2023; Todd-Brown et al., 2013; Varney et al., 2022). Q_{10} is defined as the magnitude of respiration increase for every 10°C increase in temperature (Davidson & Janssens, 2006; Lloyd & Taylor, 1994). The intrinsic Q_{10} applied by many land models is often a fixed value across space (Lawrence et al., 2015; Todd-Brown et al., 2018), which then interacts with other factors such as soil moisture, freeze-thaw transitions, and implicitly represented microbial processes to display an apparent Q_{10} (Tang & Riley, 2015). Due to these environmental interactions, the apparent Q_{10} is often a different value than a model's intrinsic Q_{10} that is implemented as a parameter (Reichstein & Beer, 2008; Ren et al., 2020; Wu et al., 2021). However, the apparent Q_{10} can be a useful metric in model evaluation to quantify how well models are capturing the overall sensitivity of respiration to changes in temperature (Sun et al., 2023). In model benchmarking, an apparent Q_{10} can be applied to quantitatively compare the shape of the functional relationship between models and observations (Huntzinger et al., 2020).

As functional relationships have become a more common benchmarking approach in model development, it is crucial to understand what factors can impact their information content and predictive power. In particular, a systematic evaluation of the choices in constructing functional relationships is needed, including what quality of data is needed to construct robust functional relationships, and subsequently how these choices can impact inferred model skill. Here, we evaluate the functional relationship between soil temperature and Reco in ecosystems across Alaska and Canada using three scales of data products, including chambers, flux towers, and upscaled gridded data products. We then quantify the functional relationships from models and observations through an apparent Q_{10}. Specifically, we evaluate how the choice of observational data product, number of daily observations, temporal averaging, and spatial and temporal extent impact the functional relationship and resulting Q_{10}. Finally, we assess how inferred model skill is influenced by the choices made when constructing functional benchmarks using simulation results from the Warming Permafrost Model Intercomparison Project (WrPMIP).

2. Methods

2.1. Observations

We investigated the impacts of the choices made when constructing functional relationships and resulting apparent Q_{10}, specifically: (a) the impact that spatial scale and choice of data set can have on the functional relationship, (b) the impact of the number of daily mean observations on the shape of the functional relationship, (c) the extent that functional relationships can be applied across space and time, and (d) the impact of temporal averaging on the shape of the functional relationship (Table S1 in Supporting Information S1). We constructed functional relationships between soil temperature and Reco using observed or observationally derived datasets across three spatial scales: <1 m^2 using autochamber measurements, $\sim$1 km^2 using flux tower data, and up to $0.5° \times 0.5°$ (ca. 50×50 km) using upscaled or reanalysis gridded data products that match the spatial resolution of many land models. Here, we evaluated Reco instead of its above or below ground respiration components, as this

21698961, 2026, 5, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025JG009307 by Helmholtz-Zentrum Potsdam GFZ, Wiley Online Library on [29/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

is the variable available from the flux tower and chamber measurements, as well as an output regularly produced by land models.

First, we evaluated autochamber measurements from the Carbon in Permafrost Experimental Heating Research (CiPEHR) site located along a permafrost thaw gradient near Healy, Alaska (Rodenhizer et al., 2023). The Arctic-Boreal flux (ABCFlux) v2 database also contains observations from over 100 chamber sites across the Arctic-Boreal region (Virkkala et al., 2025). However, we were limited to only evaluating chamber measurements at CiPEHR because other chambers across the region were either (a) not co-located with a flux tower, preventing analysis across spatial scale, (b) had measurements that were collected at different years from the corresponding flux tower, or (c) did not contain concurrent Reco and soil temperature measurements. At CiPEHR, experimental warming was performed using open top chambers to increase air and surface soil temperatures in the summer and snow fences to increase soil temperatures in the winter (Natali et al., 2011). We evaluated the chamber measurements from the six control plots that were associated with each snow fence where no additional warming was applied. At each plot, CiPEHR produced half-hourly autochamber measurements of NEE from 2008 to 2021 (Celis et al., 2017). Autochamber Reco measurements at CiPEHR were partitioned from NEE using night time measurements when photosynthetically active radiation (PAR) was <5 μmol m^{-2} s^{-1}, and then extrapolated to daytime by fitting a relationship between soil temperature and Reco at 10 cm depth (Natali et al., 2011). We only evaluated Reco when PAR was <5 μmol m^{-2} s^{-1} and between the hours of 18:00-6:00 Alaska Standard Time (Nagano et al., 2018) to reduce the impact from daytime flux estimates that may be sensitive to the extrapolation approach.

Second, from the AmeriFlux database (Chu et al., 2023), we selected 16 flux towers across Alaska and Canada with four or more years of soil temperature and Reco growing season measurements during the time period 2001–2020 (Table 1). These 16 sites cover a variety of Arctic vegetation types including shrublands, wetlands, and boreal forests across Alaska and Canada, enabling the evaluation of how functional relationships vary across space and ecosystem type (Table 1, Figure S1 in Supporting Information S1). At each site, soil temperatures were collected at various depths, and here we analyzed soil temperatures at a depth of 5 cm - the deepest measurement consistently collected across all chamber and tower sites. At each flux tower, NEE was measured at a half-hourly temporal resolution and then partitioned into its Reco and gross primary production component fluxes. To minimize the impact that the partitioning could have on results, we used non-gap filled data when available and night time measurements where photosynthetic photon flux density (PPFD) was <10 μmol m^{-2} s^{-1} based on standard practice (Euskirchen et al., 2014). We further selected measurements between the hours of 18:00-6:00 to account for any low PPFD values that may have occurred during daytime hours (Nagano et al., 2018). For this analysis, we used daily averaged measurements from flux towers and chambers, with the exception of when evaluating the impacts of temporal averaging on functional relationships, as daily averaged data was more comparable with the model outputs.

Third, we applied several gridded data products that report Reco, including FLUXCOM-X (Nelson et al., 2023) and an upscaled carbon flux data set specific to the Arctic-Boreal region (Virkkala et al., 2024). FLUXCOM-X and Virkkala et al., 2024 both upscaled flux tower measurements using machine learning and satellite observations but did so globally at about 5-km resolution and 1-km resolution for the pan-Arctic, respectively. We applied reformatted versions of these data products at 0.5° × 0.5° that better match the spatial resolution from the models. We used monthly soil temperature from the MERRA-2 (GMAO, 2015) and ERA-5 (EMCWF, 2019) reanalysis products (Table S2 in Supporting Information S1). MERRA-2 and ERA-5 are both global reanalysis data products, and here we evaluated soil temperatures at 10 cm for MERRA-2 and 7 cm for ERA-5, as these were the shallowest soil temperature estimates available.

2.2. Models

We evaluated the performance of 13 WrPMIP models (Table S3 in Supporting Information S1) when benchmarked against functional relationships derived from different observational data products and benchmark construction choices. WrPMIP is a model intercomparison effort to assess how modeled permafrost carbon dynamics respond to a warmer climate. Models in WrPMIP simulated permafrost carbon dynamics under warmer conditions by replicating the air and soil warming observed at experimental warming sites across the pan-Arctic. The warmed simulations then allow for modeled carbon flux dynamics to be compared to those observed across

Journal of Geophysical Research: Biogeosciences

10.1029/2025JG009307

Table 1
Site Descriptions of Eddy Covariance Flux Tower Sites Used in the Analysis

Site ID	Site name	Latitude	Longitude	Temporal coverage	Amount of daily data (%)	Landcover type	Reference
CA-Man	Manitoba–Northern Old Black Spruce	55.8796	−98.4808	2001–2006	95.10	Evergreen Needleleaf Forest	Amiro (2023)
CA-Oas	Saskatchewan–Western Boreal, Mature Aspen	53.6289	−106.1978	2001–2010	81.63	Deciduous Broadleaf Forest	Black (2016a, 2016b)
CA-Obs	Saskatchewan–Western Boreal, Mature Black Spruce	53.9872	−105.1178	2001–2010	59.54	Evergreen Needleleaf Forest	Black (2016a, 2016b)
CA-Ojp	Saskatchewan–Western Boreal, Mature Jack Pine	53.9163	−104.692	2001–2010	67.84	Evergreen Needleleaf Forest	Black (2019)
CA-SCB	Scotty Creek Bog	61.3089	−121.2984	2014–2019	80.39	Permanent Wetlands	Sonnentag and Quinton (2021)
CA-SCC	Scotty Creek Landscape	61.3079	−121.2992	2013–2016	69.77	Evergreen Needleleaf Forest	Sonnentag and Quinton (2025)
CA-SJ2	Saskatchewan–Western Boreal, Jack Pine forest harvested in 2002	53.945	−104.649	2004–2008	70.20	Evergreen Needleleaf Forest (2002 harvest)	Barr and Black (2018)
US-BZF	Bonanza Creek Rich Fen	64.7013	−148.3121	2011–2012; 2017–2020	97.93	Permanent Wetlands	Euskirchen (2025a, 2025b)
US-BZS	Bonanza Creek Black Spruce	64.6963	−148.3235	2011–2012; 2017–2020	96.95	Evergreen Needleleaf Forest	Euskirchen (2025a, 2025b)
US-EML	Eight Mile Lake Permafrost thaw gradient	63.8784	−149.2536	2015–2019	82.48	Open Shrublands	Bracho et al. (2021)
US-Fcr	Cascaden Ridge Fire Scar	65.3968	−148.9348	2011–2014	94.12	Open Shrublands (2010 fire)	Ueyama et al. (2023)
US-Ich	Imnavait Creek Watershed Heath Tundra	68.6068	−149.2958	2007–2009; 2012–2020	59.69	Open Shrublands	Euskirchen (2025a, 2025b)
US-Ics	Imnavait Creek Watershed Wet Sedge Tundra	68.6058	−149.311	2007–2008; 2011; 2014–2020	58.10	Permanent Wetlands	Euskirchen (2025a, 2025b)
US-Prr	Poker Flat Research Range Black Spruce Forest	65.1237	−147.4876	2011–2020	89.74	Evergreen Needleleaf Forest	Iwahana et al. (2023)
US-Rpf	Poker Flat Research Range: Succession from fire scar to deciduous forest	65.1198	−147.429	2009–2012; 2014–2020	89.07	Deciduous Broadleaf Forest (2004 fire)	Ueyama et al. (2025a, 2025b)
US-Uaf	University of Alaska, Fairbanks	64.8663	−147.8555	2003–2020	95.68	Evergreen Needleleaf Forest	Ueyama et al. (2025a, 2025b)

Note. Amount of daily data column indicates the percentage of Reco and soil temperature measurements at 5 cm depth available during the growing season.

21698961, 2026, 5, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025JG009307 by Helmholtz-Zentrum Potsdam GFZ, Wiley Online Library on [29/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

21698961, 2026, 5, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025JG009307 by Helmholtz-Zentrum Potsdam GFZ, Wiley Online Library on [29/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

experimental warming sites. Results from WrPMIP can further inform permafrost representations within models, particularly through benchmarking and evaluation of model performance (Poe et al., 2024).

The land models in WrPMIP simulated the permafrost region of 55°N–85°N and performed three pan-Arctic simulations—control, open top chamber, and snow fence—by aiming to reproduce the amount of warming observed at a collection of open top chamber and snow fence experiments from the Tundra Flux Database (Maes et al., 2024). Warming targets were set between 0.5 and 1°C of summer air warming for open top chamber simulations and 1.5°C–3°C of winter soil warming at a depth of 10 cm for snow fence simulations across the pan-Arctic. Each model was forced with CRU-JRA climatology—a combination of Climate Research Unit monthly meteorological data (Harris et al., 2020) and the Japanese Reanalysis (Kobayashi et al., 2015). Models were spun up to 1901 by recycling 1901–1920 climate forcing until reaching stable conditions. Models then performed a transient run from 1901 to 2020, and the warming simulations were run from 2000 to 2021, matching the time period of data collection from warming experiments in the Tundra Flux Database. The last 20 years of the transient run (pre-warming) were used as the control simulation for each model.

Pan-Arctic air warming was achieved in the models by increasing the aerodynamic resistance of thermal and water vapor transfer in the canopy, which traps heat and increases air temperatures. Soil warming was achieved in models by reducing the thermal conductivity of snow, which causes the snow pack to better insulate the soil and less heat to be lost to the atmosphere in the winter (Schadel et al., 2018). The reduction of thermal conductivity of snow was then turned off during each growing season so that the timing of snow melt and summertime soil moisture from excess snow was not impacted. Here, we applied the control simulations from 2001 to 2020, each with a monthly temporal resolution and a regridded spatial resolution of $0.5° \times 0.5°$. Each model in WrPMIP has different soil layer depths, which could impact the results when comparing model outputs. As a result, we extrapolated the available soil layers in each model to match those from CLM5.0, with 25 soil layers that extend to a depth of ∼50 m (Lawrence et al., 2019). We then used the soil layer with a depth of 4 cm, as it is closest to the 5 cm soil depth measured at the chambers and flux towers.

The experimental warming sites that the WrPMIP models replicated were located from 55°N to 78°N, covering both tundra and boreal regions. However, all of these experimental warming sites were located in either tundra, grassland, or high elevation ecosystems. When benchmarking models at these sites, the warming responses could potentially be misrepresented if the site locations are characterized by boreal forest landcover types in the models. As a result, the boreal forest region in the CLM5, CLM5-ExIce, and ELM-ECA simulations was replaced with C3 grass to correspond with the tundra ecosystem type from the warming experiments (Schädel et al., 2018).

2.3. Functional Relationship Determination

We fitted functional relationships between soil temperature at a depth of 5 cm and Reco using co-located observations from each flux tower site. For the upscaled data sets and models, we fitted functional relationships by selecting the grid cell that contained each flux tower site. We focused on data during the growing season (defined here as May–September) due to data availability. We then applied an exponential fit to Reco as a function of soil temperature:

$$\text{Reco} = a\, e^{bT} \tag{1}$$

where a and b are fitted parameters that vary at each site according to the exponential fit and T is the soil temperature in °C at each site. More specifically, a is the intercept of the functional relationship and b determines the steepness of the exponential curve. Fitting an exponential curve allows for the extraction of an apparent Q_{10} to quantify the amount of respiration increase for every 10°C increase in temperature. Quantifying the Reco response to temperature then allows for a direct comparison between the apparent Q_{10} from observations and land models at a given location. For the rest of the paper, when discussing Q_{10}, we refer to apparent Q_{10}, rather than intrinsic Q_{10}, unless otherwise noted. We used the equation below to calculate an apparent Q_{10} from each observed and modeled functional relationship, where b is the same parameter as Equation 1.

$$Q_{10} = e^{10b} \tag{2}$$

21698961, 2026, 5, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025JG009307 by Helmholtz-Zentrum Potsdam GFZ, Wiley Online Library on [29/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

2.4. Analysis Approach

When testing the impact of the choice of data set on the functional relationship and Q_{10}, we were limited to the evaluation of chambers against flux towers and upscaled data sets at one location—CiPEHR—because of its extensive chamber measurements. However, the CiPEHR experiment is located within ~0.5 km of the US-EML flux tower, providing an opportunity to evaluate the potential uncertainty in functional relationships from measurements at different spatial scales that are in close proximity (Celis et al., 2017). At each flux tower location, we then selected the gridcell from the four upscaled data products (Table S2 in Supporting Information S1) that contained the latitude and longitude of the flux towers. Additionally, when evaluating the choice of data set, we focused on the years that were consistent across all data set types at each site.

The flux towers used in this analysis measure half-hourly NEE, which can be partitioned into Reco, spanning several years to decades, and are located in various ecosystem types across Alaska and Canada. Therefore, they provide an opportunity to further evaluate the variability in functional relationships across space and time. We used data from 16 flux towers to investigate how apparent Q_{10} varies across space, and how the number of daily observations, temporal extent, and temporal resolution influence the shape of functional relationships and the resulting Q_{10}. Using the nearly continuous growing season data record from these towers (Table 1), we first examined the number of daily observations necessary to ensure that the shape and magnitude of the functional relationship is accurate, that is, within the 95% confidence interval (CI) of the Q_{10} calculated with the full-length data set. To do so, we randomly selected 1 through 30 daily averaged data points per month over the full data record at each site to replicate potential data gaps common in high latitude site-level data. In cases where a particular month contained full days of missing data, we evaluated up to the full extent available. We performed this random selection for 500 iterations on each subset of data points to capture all possible ranges of the functional relationship, fitting an exponential curve to all 500 iterations across the 30 sets. With each of these relationships constructed with randomized data, we then calculated an apparent Q_{10} and evaluated the percentage of times that the randomized Q_{10} fell within the 95% confidence interval (CI) of the baseline Q_{10} that used all of the daily data points.

We examined the temporal extent that functional relationships can be applied across time by constructing functional relationships and extracting a Q_{10} over consecutive 1-year subsets, with increments increasing up to the full time period available at each flux tower site. To examine the influence of temporal averaging on the functional relationship, we averaged the baseline half-hourly flux tower measurements up to daily, weekly, monthly, and growing season (May–September) means, and calculated individual exponential relationships for Reco as a function of soil temperature at each temporal resolution. We then applied the results from each of these permutations as benchmarks to examine the influence that these various choices can have on the inferred model skill.

2.5. Calculation of Inferred Model Skill

The relationship between soil temperature and Reco, quantified through an apparent Q_{10}, can be applied as a benchmark for models to measure mismatches in temperature sensitivity against data products. However, if the Q_{10} from the observational benchmark varies as a result of choices applied to construct the benchmark (e.g., temporal averaging or number of daily observations), then inferences of model skill could be misrepresented. We investigated how inferred model skill is impacted by the benchmarking choices from the various functional relationships constructed in Section 2.4. Here, we calculated a single Q_{10} for each model at the $0.5° \times 0.5°$ gridcell that contains the latitude and longitude of each of the 16 flux tower sites. Additionally, we evaluated the impacts of benchmarking choices on model skill over the time periods that overlapped with each flux tower site. We then compared the difference between the apparent Q_{10} from models and the Q_{10} from various data products and choice categories (Table S4 in Supporting Information S1), calculating multiple inferred model scores from different permutations of the benchmark data sets.

We calculated the differences in apparent Q_{10} between models and observations, scaled from 0 to 1, as a measure of inferred model skill. We chose this scoring metric to normalize large biases in Q_{10} between models and observations, where scores closer to 1 indicate higher model agreement with the Q_{10} from observations. The scaling of the skill score followed the methods for calculating a functional relationship score from the International Land Model Benchmarking (ILAMB) project (Collier et al., 2018), except we used Q_{10} in the metric calculation instead of root mean square error:

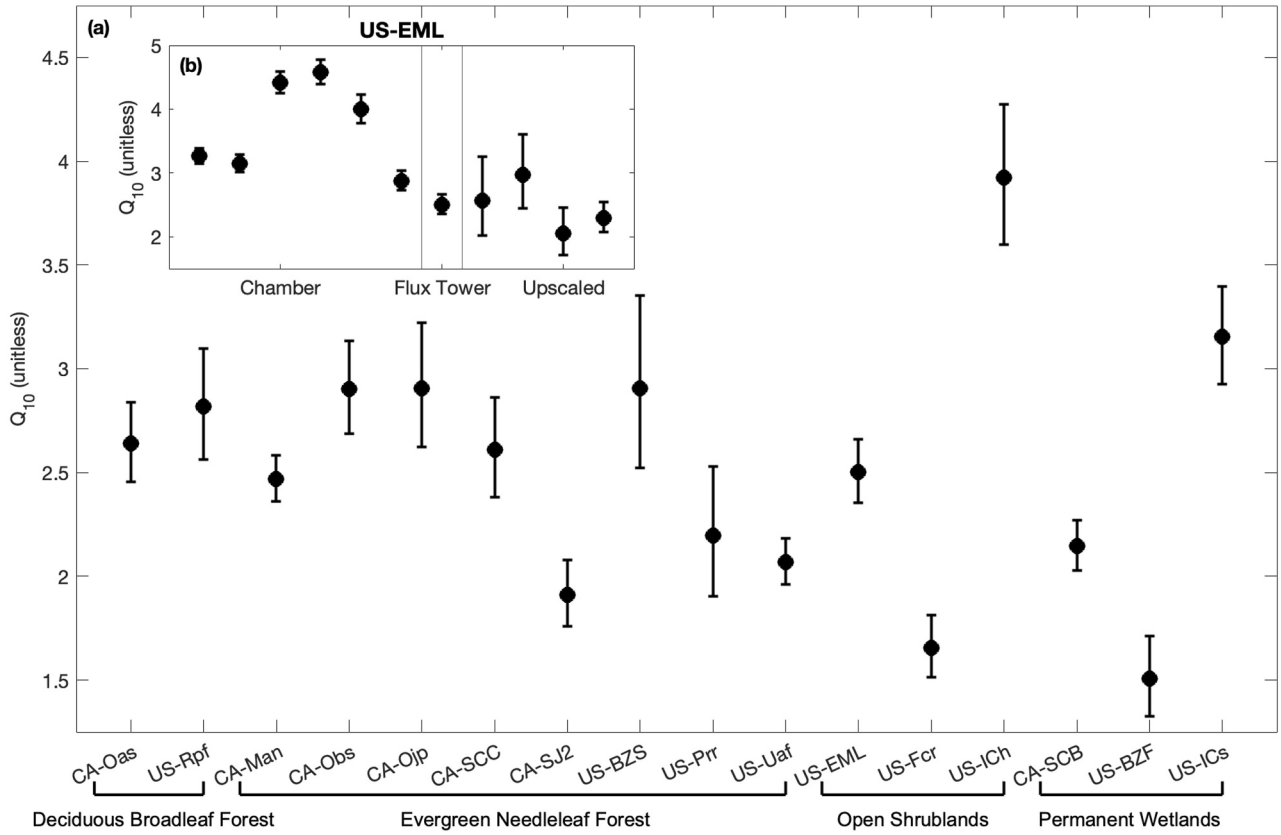


Figure 1. Spread in Q_{10} at (a) 16 flux tower sites across Alaska and Canada and (b) the location of US-EML according to the choice and spatial scale of data set with its close proximity to CiPEHR. In (a), Q_{10} values were calculated from daily growing season measurements for the full temporal extent at each flux tower (varying time periods from 2001 to 2020). In (b), chamber Q_{10} values were calculated from six control chamber plots at CiPEHR, and the flux tower Q_{10} was calculated at US-EML. The upscaled category consists of four Q_{10} values calculated from FLUXCOM-X with MERRA-2 and ERA-5 (left) and from Virkkala et al. (2024) with MERRA-2 and ERA-5 (right). Vertical bars indicate the 95% CI of each Q_{10}.

$$S_{\text{skill}} = e^{-\frac{\|\text{mod}-\text{obs}\|}{\|\text{mod}\|}} \tag{3}$$

where mod is the apparent Q_{10} from the model, obs is the apparent Q_{10} from the observational data set, and ‖ is the Euclidean norm. When evaluating the inferred model skill, skill scores that remain consistent within each choice category indicate that these choices have a low impact on model skill inference, while skill scores that have more variation within a category indicate a larger impact on model skill inference.

3. Results and Discussion

3.1. Choice of Benchmark Data Set

To evaluate how the choice of data set impacts the functional relationship, we initially focused on the location of CiPEHR/US-EML, as this location has multiple data products available at various spatial scales (i.e., chamber, flux tower, upscaled). The Q_{10} from these various observational approaches ranged from 2.3 to 4.6 (Figure 1b). At the chamber scale, the Q_{10} across plots ranged from 2.9 to 4.6. The Q_{10} at the flux tower scale for US-EML was 2.5. At the gridded scale, the Q_{10} ranged from 2.3 to 3.1. The variability in Q_{10} across chamber plots at CiPEHR was large but not unexpected given the landscape heterogeneity. These differences in Q_{10} across the chambers are linked to plot location across a gradient of permafrost thaw (Schuur et al., 2021). Here, active layer thickness, soil moisture, and soil temperature varied across plots (Mauritz et al., 2017; Natali et al., 2015), thus impacting apparent Q_{10}. The remaining 15 locations across the study area also exhibited large variability across spatial scale and choice of data set (Figure S2 in Supporting Information S1). At many of these sites, the Q_{10} values from flux towers often fell within the 95% CI of the Q_{10} from at least one of the upscaled products. However, the Q_{10} values from the upscaled

21698961, 2026, 5, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025JG009307 by Helmholtz-Zentrum Potsdam GFZ, Wiley Online Library on [29/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

21698961, 2026, 5, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025JG009307 by Helmholtz-Zentrum Potsdam GFZ, Wiley Online Library on [29/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

products themselves also varied from one another, where the Q_{10} values from FLUXCOM-X were often higher than those from Virkkala et al., 2024. Overall, these Q_{10} values are in range with previous Q_{10} values reported across the permafrost region, often between 1 and 10 (Kim et al., 2016; Wilkman et al., 2018; Wu et al., 2021).

This variability in Q_{10} across scales has also been observed in previous studies (e.g., Reichstein & Beer, 2008; Wilkman et al., 2018) due to different confounding factors at increasing spatial scales, different observational approaches, and partitioning methods. Even though site-level data may provide more direct observations at a specific location, upscaled data sets may often be a more accurate benchmark for models because the spatial scales are more comparable with similar amounts of landscape heterogeneity. Additionally, even though each location exhibited different Q_{10} values according to the approach and spatial scale evaluated, applying multiple data products can create a higher likelihood of encompassing the full variability from different observational approaches. As a result, when performing model benchmarking, multiple data sets should be applied as benchmarks and ideally over multiple spatial scales (Luo et al., 2012; Schwalm et al., 2013). In particular, benchmarking models against site-level data is important to ensure that finer scale heterogeneity is also being evaluated, in addition to applying benchmark data sets with spatial scales that match those of models (e.g., 0.5° × 0.5°).

3.2. Variability in Q_{10} Across Space

The Q_{10} values calculated from daily observations at the 16 flux tower sites across Alaska and Canada varied from 1.5 to 3.9 (Figure 1a). There were four site clusters within these selected sites – Scotty Creek (CA-SCB and CA-SCC), Bonanza Creek (US-BZF and US-BZS), Imnavait Creek (US-ICh and US-ICs), and Poker Flat Research Range (US-Prr and US-Rpf). At each of these site clusters, flux towers were located within 3 km of each other but often exhibited Q_{10} values that were substantially different from one another. For example, US-BZF exhibited a Q_{10} of 1.5, while US-BZS exhibited a Q_{10} of 2.9 even though the sites were <1 km apart. These flux towers in close proximity might exhibit large variations in Q_{10} because they are often intentionally set up in heterogeneous areas with different landcover types (e.g., Imnavait Creek) or along moisture gradients (e.g., Bonanza Creek). The landcover type at US-BZF was described as a permanent wetland, while US-BZS was described as an evergreen needleleaf forest (Euskirchen, 2025a, 2025b). Wetlands and areas with high amounts of soil moisture tend to exhibit lower sensitivities of Reco to temperature due to reduced oxygen availability for microbial activity (Liu et al., 2024; Meyer et al., 2018), explaining the lower Q_{10} at US-BZF. This spatial heterogeneity has a strong influence on the functional relationships across the Arctic-Boreal region and should be closely considered when applying functional relationships outside of a site's spatial extent or extrapolating to larger domains.

The highest Q_{10} across sites occurred at US-ICs and US-ICh, sites at the highest latitude and with the lowest mean annual temperature (Table 1). Across wide regions, Q_{10} is known to generally increase with latitude and as mean annual temperature decreases (Koven et al., 2017; Li et al., 2020; Ren et al., 2020). The temperature sensitivity of respiration has also been observed to vary across space depending on factors such as soil properties, soil moisture, C:N, vegetation productivity, and active layer thickness (Han & Jin, 2018; Liu et al., 2024; Meyer et al., 2018; Wang et al., 2020). However, these environmental variables are not always collected at sites, making it difficult to draw conclusions about factors that could be driving differences in Q_{10} across space. Additionally, due to data availability, this analysis did not take into account how the Q_{10} could be affected by a lack of non-growing season data that is common in high latitude ecosystems. The apparent Q_{10} of an ecosystem can be strongly impacted by low temperatures and freeze-thaw processes (Hamdi et al., 2013; Li et al., 2019), where the transition from frozen to thawed soil can drive a large change in respiration over a small change in temperature. As a result, the Q_{10} is often higher in the non-growing season (Wang et al., 2020), and Q_{10} values are often calculated and evaluated separately by season (Suseela et al., 2012; Wang et al., 2014).

At individual flux tower sites, apparent Q_{10} also varied with the depth of temperature evaluated, typically being lowest when calculated with air temperature and highest with increasing soil depths from 2 to 10 cm (Figure S3 in Supporting Information S1). However, Q_{10} did not consistently increase with soil depth at all sites. Additionally, soil temperatures did not consistently decrease with depth across sites (Figure S4 in Supporting Information S1), indicating that whether Q_{10} increases or decreases with soil depth likely also varies across space depending on site-level characteristics such as thaw depth. Several studies in permafrost regions found a higher Q_{10} in deeper soil layers compared to surface soil layers due to factors such as carbon input, C:N, and pH (Li et al., 2020; Ren et al., 2020); however, these studies evaluated depths that extended past the 10 cm evaluated here. Together, these findings highlight that the soil depth at which temperature measurements are taken should be closely considered

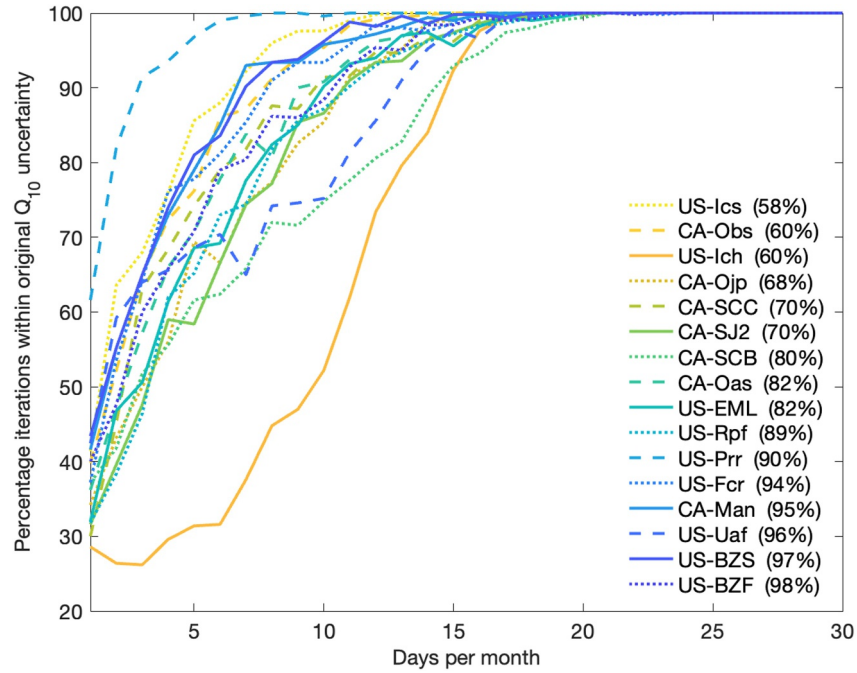


Figure 2. Percentage of the 500 random iterations that fall within the 95% CI of the baseline daily Q_{10} calculated with one through 30 days per month across the full temporal extent of each site. The point at which each line levels out at 100% is the minimum number of days needed to be collected each month to completely fall within the 95% CI of the Q_{10} calculated with all data. Each line indicates a flux tower, and lines are color coded by the percentage of available data each site had to begin with from Table 1. Sites with the lowest amount of available data to begin with are colored yellow, and sites with the highest amount of available data are colored dark blue.

when comparing apparent Q_{10} across studies or in model benchmarking (Reichstein & Beer, 2008). Additionally, land models should simulate soil temperature at multiple depths while also accounting for likely Q_{10} variation with soil depth (Li et al., 2020; Ren et al., 2020).

3.3. Number of Daily Observations

Across each of the 16 flux towers, the number of daily mean observations necessary for the Q_{10} to fall within the 95% CI of the Q_{10} constructed with all data points was approximately 18 days per month, ranging from 8 to 21 days (Figure 2). For example, at US-EML, the functional relationships from 500 random iterations of selecting 1 day per month during the growing season resulted in a wide spread of functional relationships, but that spread rapidly decreased as more days were sampled out of the month (Figure S5 in Supporting Information S1). At US-EML, the spread of functional relationships and resulting Q_{10} values were fully captured within the 95% CI of the original relationship when at least 20 days out of the month were sampled. Across sites, the Q_{10} calculated with only one random daily mean measurement per month was captured within the 95% CI of the baseline Q_{10} approximately 38% of the time. When sampling 15 daily measurements per month, the likelihood of the Q_{10} being captured in the 95% CI of the Q_{10} using the full data set rose to approximately 98% (Figure 2). As a result, a Q_{10} that is calculated with data containing gaps spanning half a month or more is at risk of misrepresenting the true apparent Q_{10} at a given site.

Since individual flux towers did not have 100% of the daily data available to begin with (Table 1), the apparent Q_{10} constructed with the full extent of data available could still potentially be inaccurate. However, the number of days sampled to be completely captured by the 95% CI of the full response was generally consistent across sites, regardless of how much data each site had to begin with. This is most clearly seen in Figure 2, where sites with less daily data to begin with reached 100% of their iterations captured by the baseline Q_{10} at the same rate as sites with more data to begin with. Since data collection at high latitude sites is often limited to a few years and large gaps between measurements are common (Pallandt et al., 2022), these results can also serve as an indicator for the minimum number of days that site level data should be collected to capture the full temperature sensitivity of Reco.

21698961, 2026, 5, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025JG009307 by Helmholtz-Zentrum Potsdam GFZ, Wiley Online Library on [29/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

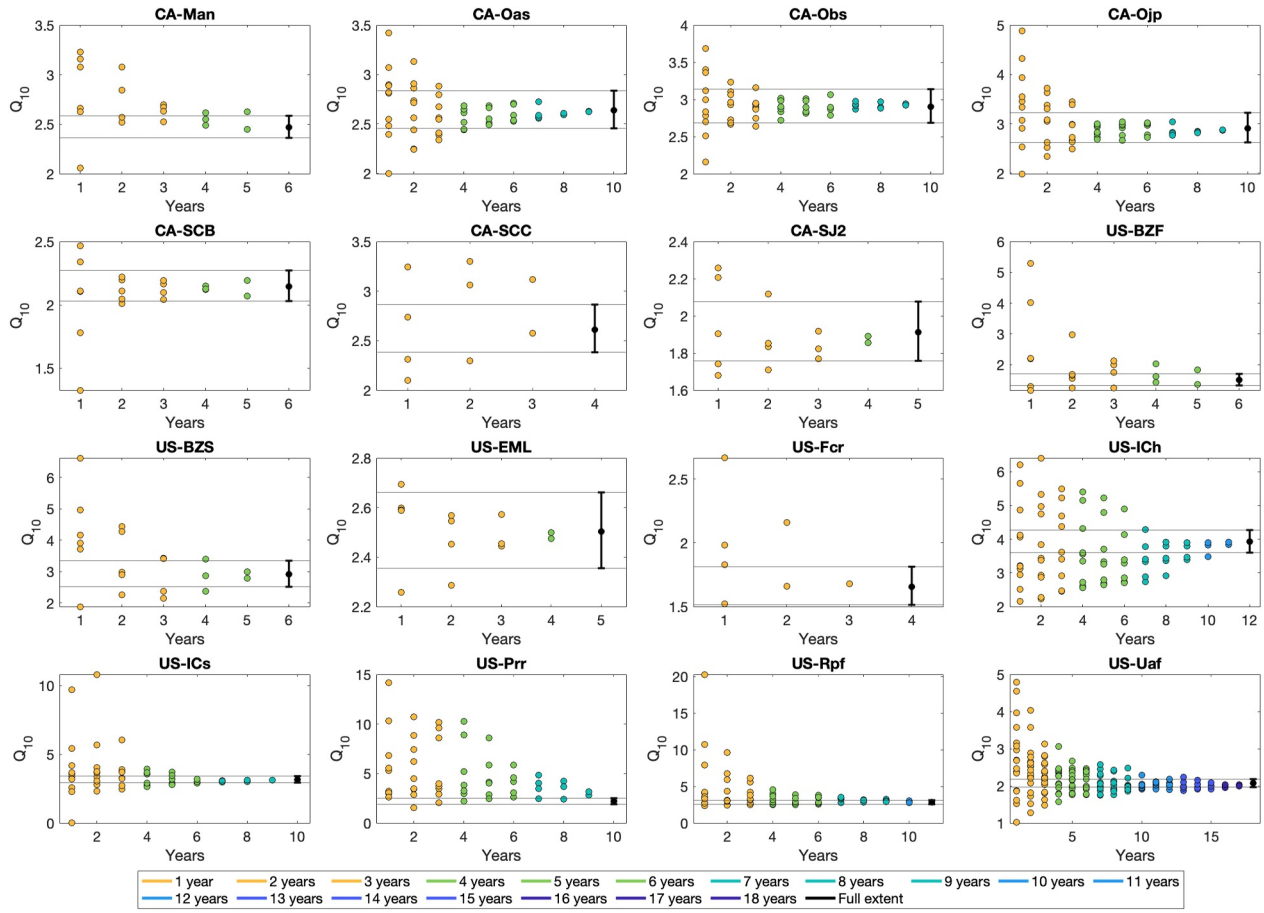


Figure 3. Variability in Q_{10} at each flux tower site based on the length of the temporal coverage subset. Colored dots indicate Q_{10} values created from subsets of 1 year (lighter colors) up to subsets of years closer to the full temporal extent (darker colors). Black dots represent the Q_{10} from the full temporal extent at each site (ranging from 4 to 18 years), with horizontal black lines indicating the 95% CI. Note that the Q_{10} values are on different scales at each site.

3.4. Temporal Extent

The record lengths among flux tower sites ranged from 4 to 18 years (mean = 8 years), where the relationship between Reco and soil temperature displayed strong year-to-year variability at each flux tower site. Here, the functional relationship constructed from any 1 year of growing season data often did not fall within the 95% CI of the relationship constructed with the full temporal extent (Figure 3; Figure S6 in Supporting Information S1). At 11 sites, at least one Q_{10} from individual years fell within the 95% CI of the Q_{10} calculated from the full extent; however, five sites (CA-Man, US-BZF, US-BZS, US-Prr, and US-Uaf) had no Q_{10} from individual years fall within this 95% CI (Figure 3). The year-to-year fluctuations in the relationship between soil temperature and Reco could be due to annual variations in each ecosystem's environmental conditions such as seasonal temperature, soil moisture, and quality of soil organic carbon (Chen et al., 2010; Han & Jin, 2018; Lefèvre et al., 2014). This range in Q_{10} decreased as more years were sampled, where half of the sites had all Q_{10} values fall within the 95% CI of the Q_{10} constructed from the full temporal extent within 5 years (Figure 3). However, this varied by site, with sites taking between three and 17 years for the Q_{10} from all subsets of data to fall within the 95% CI of the Q_{10} from the full temporal extent. The Q_{10} at each site generally remaining consistent after several years of measurements (i.e., at least 5 years) demonstrates the power of functional relationships to be applied across longer periods of time.

At several sites such as US-Prr, US-Rpf, and US-Uaf, there was substantial inter-annual variability in the shape and magnitude of the relationships (Figures S6n–S6p in Supporting Information S1). These variations could potentially be due to significant environmental changes such as disturbance at these sites. For example, US-Rpf experienced a severe wildfire in 2004 (Ueyama et al., 2019), but, on the other hand, provides the

21698961, 2026, 5, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025JG009307 by Helmholtz-Zentrum Potsdam GFZ, Wiley Online Library on [29/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

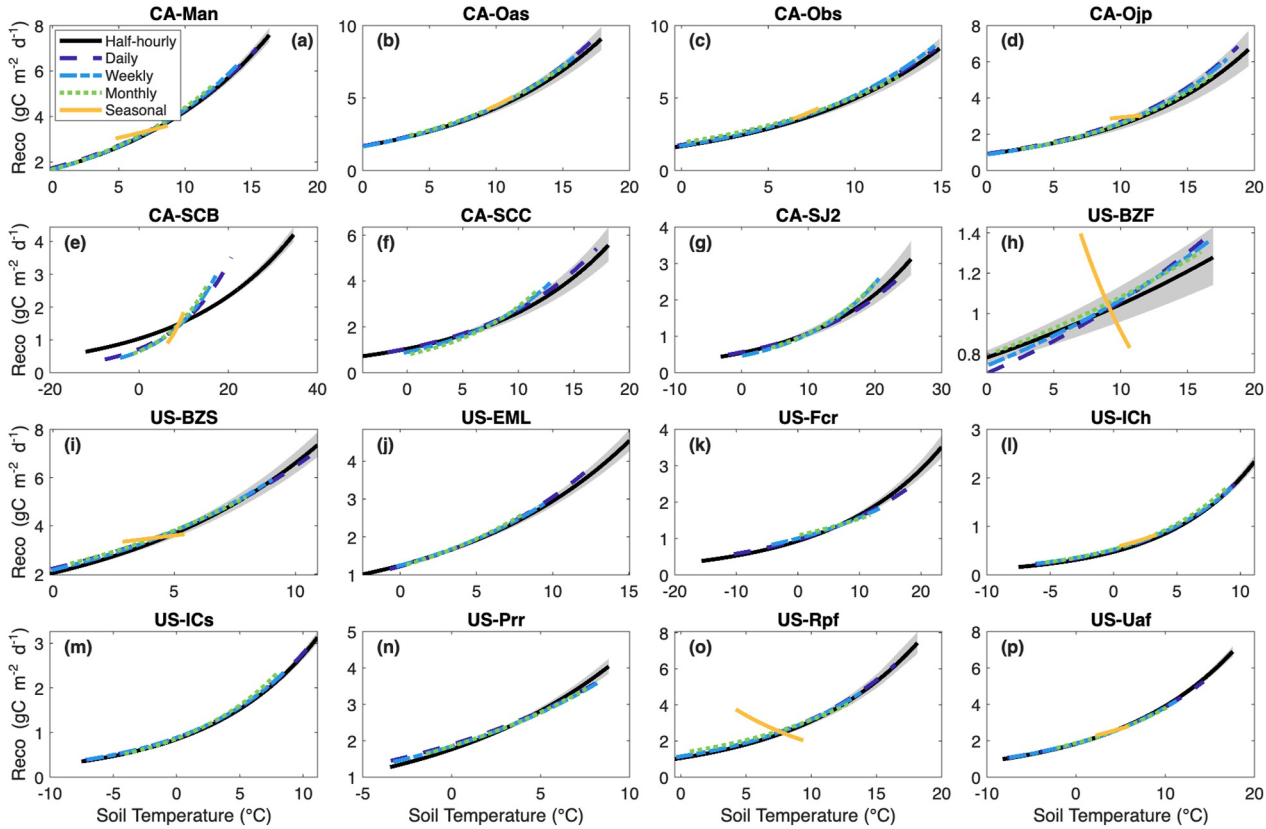


Figure 4. Impacts of temporal averaging on the relationship between soil temperature and Reco at 16 flux tower sites in panels (a)–(p). Black lines indicate the response of Reco to soil temperature from the baseline half-hourly data, with the shaded gray regions indicating the 95% CI. Colored lines indicate relationships constructed from temporally averaged data. Note that soil temperature and Reco are on different scales at each site.

opportunity to evaluate post-fire impacts with measurements from 2009 to 2020. At US-Rpf, Reco fluxes were substantially lower in the early succession period following the fire (2009–2012), after which Reco began to increase following vegetation regrowth (Ueyama et al., 2019; Figure S7 in Supporting Information S1). This increase in Reco is also seen in the functional relationships between the two recovery periods, where there was an upward shift in the magnitude of the long-term relationship (Figure S8 in Supporting Information S1). Furthermore, the short-term relationship exhibited a higher Q_{10} of 4.0 (3.6, 4.5) in the first four measurement years during early succession following the wildfire (2009–2012) compared to a Q_{10} of 3.6 (3.3, 3.8) in mid-succession years (2014–2020, Figure S8 in Supporting Information S1). While most sites exhibit some inter-annual variability in Q_{10} that averages out with a long enough measurement period, extreme events can alter the Q_{10} at sites for at least several years. At these locations, the Q_{10} may not be able to be as widely applied across time, making it important to consider recent disturbances when constructing functional relationships to benchmark land models.

3.5. Temporal Averaging

The functional relationships after averaging flux tower data up to daily, weekly, and monthly means resulted in a similar shape and magnitude as the baseline relationship using half-hourly data (Figure 4). However, the R^2 from each relationship was oftentimes highest when constructed with data averaged over monthly timescales and lowest over seasonal timescales (Table S5 in Supporting Information S1). The weaker relationship at half-hourly resolutions is due to noise that can occur at sub-daily timesteps (Niu et al., 2011) that averages out over daily to monthly timesteps. Several sites had consistently low R^2 values (<0.4) across all timescales (Table S5 in Supporting Information S1). While the weaker relationship found at some sites indicates that other environmental factors (e.g., soil moisture) have an impact on the functional relationship, these functional relationships are still an

21698961, 2026, 5, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025JG009307 by Helmholtz-Zentrum Potsdam GFZ, Wiley Online Library on [29/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

21698961, 2026, 5, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025JG009307 by Helmholtz-Zentrum Potsdam GFZ, Wiley Online Library on [29/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

important benchmark for models to track the full ecosystem response to temperature at a given site. The weakest relationship between soil temperature and Reco occurred over seasonal averages, where averaging reduced the number of seasonal observations to oftentimes five data points or less. In several cases, this resulted in Q_{10} values and a 95% CI that are beyond what is physically possible. We therefore excluded these values from the analysis (Table S5 in Supporting Information S1 in Supporting Information S1). At other sites with more data points at the seasonal timescale (>10), lower R^2 values are likely the result of less variability in the factors controlling the relationship over longer timescales (Johnston et al., 2021).

Although the correlation became stronger up to monthly timesteps, the range of temperatures underlying the functional relationship became narrower until the seasonal mean, where it failed to capture the full response of Reco to temperature observed at each site (Figure 4). Here, the temperature ranges from the monthly and seasonal mean relationships were reduced by an average of 6.1°C and 15.4°C, respectively, compared to the baseline half-hourly relationships across all sites (Figure 4). As a result of the narrow temperature range and weak correlation at seasonal timesteps, we excluded seasonally averaged functional relationships from the rest of the analysis.

Although temporal averaging from half-hourly to monthly reduced the range of observations, the Q_{10} generally remained the same across sites. Temporal averaging from half-hourly to monthly altered the Q_{10} by as little as 0.1 (e.g., CA-Oas) and up to 1.16 (e.g., CA-SCC; Table S5 in Supporting Information S1). The sites with the largest differences in Q_{10} after averaging (Figures 4e–4g) also had some of the shortest temporal extents, indicating that functional relationships from sites with fewer years of data are more likely to be impacted by temporal averaging. Often in model benchmarking there are computational constraints associated with evaluating models or data at their native temporal resolutions (e.g., evaluating modeled carbon fluxes at monthly temporal resolution rather than finer scale; Xia et al., 2020). Additionally, many benchmark data sets are upscaled products that are at daily or monthly temporal resolutions. These results indicate that daily or monthly averages of modeled or observed outputs should not significantly impact the Q_{10}, but seasonal averages will likely result in a different Q_{10}.

3.6. Influence of Benchmarking Choices on Inferred Model Skill

The Q_{10} from WrPMIP models ranged from 1.2 to 5.5 (mean = 2.4) across flux tower sites (Figure S9 in Supporting Information S1). Models both underestimated and overestimated the daily baseline Q_{10} that was calculated from direct observations at each site (Figure S9 in Supporting Information S1). These differences in Q_{10} between models and observations translated to baseline model skill scores between 0.31 and 1.0 (mean = 0.78; Table S6 in Supporting Information S1). We then used these initial model scores as a reference when evaluating the impacts of benchmarking choices on inferred model skill. The impacts of the four benchmarking choice categories (Table S4 in Supporting Information S1) led to large variations in inferred model skill within individual models (Figure 5; Figures S10, and S11 in Supporting Information S1). The largest range in inferred model skill at each site typically occurred according to the number of daily observations used to construct the relationship, where the median variation in inferred model skill was 0.63 across models and sites (Figure 5). The temporal extent and choice of data set (i.e., chamber, flux tower, and upscaled) used to construct the relationship also resulted in a large range of model scores, with median ranges of 0.49 and 0.27, respectively. Temporal averaging of flux tower data from half-hourly to monthly generally resulted in a low range of model scores, with a median range of 0.08.

Although the number of daily observations and temporal extent typically had the largest impact on inferred model skill, this ranking varied based on the site evaluated. For example, at US-EML, the choice of data set had a larger impact on inferred model skill than the temporal extent (Figure S10 in Supporting Information S1). This could be because the choice of data set category at US-EML included chamber, flux tower, and upscaled observations, which displayed a wide range of Q_{10} values according to the spatial scale. These results from US-EML emphasize the impact that the choice of data set can have on the functional relationship and resulting Q_{10}, particularly when there is a large mismatch in spatial scale between the benchmark data set and model. Additionally, the impact of temporal extent had variable impacts on inferred model skill depending on the site evaluated (Figure S10 in Supporting Information S1). For example, the temporal extent at US-Prr and US-Rpf resulted in a high impact on inferred model skill, with ranges in scores up to nearly 1, while sites such as US-EML and CA-SJ2 produced low ranges in scores of around 0.1 and 0.2, respectively.

These wide ranges in inferred model skill depending on the number of daily observations and temporal extent emphasize the impact of choices made during the construction of the benchmark itself on inferred model skill. The

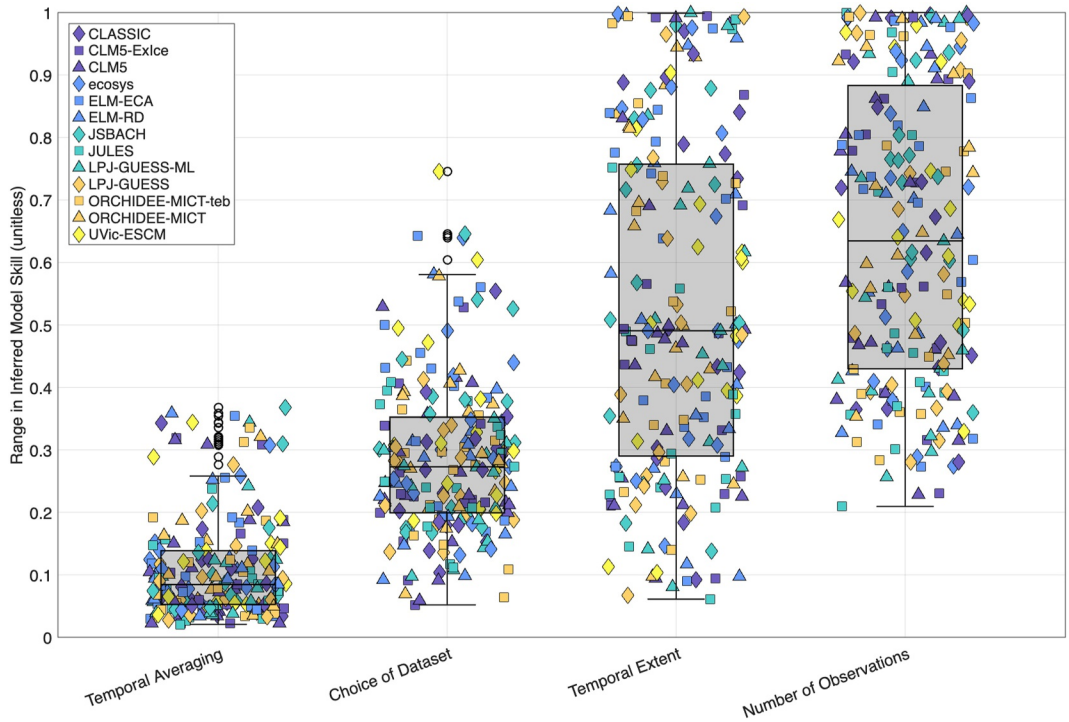


Figure 5. Range in inferred model skill through the Q_{10} scoring metric across all models and sites. The range in inferred model skill varies according to benchmarking choices for temporal averaging, choice of data set (i.e., chamber, flux tower, and upscaled), temporal extent, and number of daily observations. The solid line within each boxplot represents the median range in the inferred model skill and box edges represent the interquartile range.

four choice categories examined (Table S4 in Supporting Information S1) all produced a range in inferred model skill, but the coarsest form of the data as a benchmark resulted in both the highest and lowest inferred model scores (Figure S11 in Supporting Information S1). For example, applying data with a limited number of daily observations can result in inferring both higher and lower estimates of model performance than a model's true skill. This range in inferred model skill can be considerably reduced by carefully choosing benchmark data sets that are spatially and temporally representative of the area being evaluated. The large range in inferred model skill according to the choice of data set is largely due to the heterogeneity of the landscape and is difficult to avoid during model benchmarking. However, these variations across space should still be closely considered, and multiple benchmarks should be applied during model evaluation to account for this variability.

While models both underestimated and overestimated the baseline Q_{10} from flux tower sites across Alaska and Canada (Figure S9 in Supporting Information S1), it should be noted that this analysis was focused on the best practices in constructing functional relationships as benchmarks and how choices made in the benchmarking process can impact model skill. Understanding what causes models to either over or underestimate apparent Q_{10} across the Arctic-Boreal region will be crucial for accurately simulating permafrost carbon dynamics. Further analysis of how well models capture these functional relationships, particularly models in WrPMIP that are focused on the impacts of warming on the Arctic-Boreal region, is the subject of future work. Instead, results from this analysis provide a set of guidelines for reducing the range of inferred model skill that can be introduced by uncertainties in the application of observational benchmarks.

Beyond applications for benchmarking and model development, these results have clear implications for observationalists. Because of historical spatiotemporal scarcity in high-latitude observations (Bouskill et al., 2020; Euskirchen et al., 2017), our results underscore the requirement for co-located, high-frequency measurements. Specifically, inferring ecosystem responses to environmental change necessitates the simultaneous acquisition of fluxes and critical environmental variables (e.g., multi-depth temperature) to establish robust functional relationships to constrain model simulations. Consistent sampling frequency across both diurnal/monthly and inter-annual timescales is also essential to eliminate temporal bias and ensure long-term data fidelity.

21698961, 2026, 5, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025JG009307 by Helmholtz-Zentrum Potsdam GFZ, Wiley Online Library on [29/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

4. Conclusions

Functional relationships are an informative approach to evaluate model performance; however, the shape and information content of functional relationships can be influenced by the choices made during the benchmarking process. We systematically evaluated the impacts that these choices have on the shape of the functional relationship and apparent Q_{10}, and based on our findings, we recommend the following best practices for constructing functional relationship benchmarks:

(a) Q_{10} varied substantially across space, highlighting the location dependency of functional relationships. Gaining a clearer understanding of what drives Q_{10} to vary across space requires more environmental variables to be consistently collected at each site along with carbon fluxes. (b) Functional relationships should be constructed with at least half of daily observations throughout each month to have the same shape and magnitude as functional relationships constructed with all available data. (c) Functional relationships and the resulting Q_{10} can fluctuate substantially from year to year and should be constructed with at least 2 years of data to begin to capture these variations. (d) The choice of Reco and temperature data sets can have a large impact on the functional relationship and Q_{10} depending on the spatial scale and approach considered. As a result, multiple data products at multiple spatial scales should be applied when constructing functional benchmarks to evaluate inferred model skill. (e) The impact of temporal averaging on Q_{10} is low when using averaged data up to monthly temporal resolution. However, in order to capture the full temperature range of the functional relationship, data should be applied at the finest temporal resolution available.

The choices made when constructing functional relationships during the benchmarking process can have a substantial influence on the inferred model performance. For our case study on apparent Q_{10} across the Arctic-Boreal region, inferred model skill is most impacted by the number of daily observations used and the temporal extent, and is least impacted by temporal averaging and the choice of data set used to construct the functional benchmark. This analysis provides a guide for future benchmarking studies to reduce the impact of choices made when collecting and applying data as benchmarks, thereby reducing the range in inferred model skill.

Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

Availability Statement

The data products from this study are publicly available. Chamber data were obtained from the CiPEHR data set (Rodenhizer et al., 2023). Eddy covariance sites with corresponding citations are included in Table 1, and are available through Ameriflux (https://ameriflux.lbl.gov/). Upscaled Reco data products were obtained from FLUXCOM-X (Nelson et al., 2023; Virkkala et al., 2024), and upscaled soil temperature data products were obtained from ERA-5 (ECMWF, 2019) and MERRA-2 (GMAO, 2015). Reco and soil temperature model outputs were obtained from WrPMIP (Wells et al., 2026).

Acknowledgments
This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant 1938054. C.S., D.H., J.M.W., and E.A.G.S received funding from the US Department of Energy (DOE) Grant DE-SC0022116. W.J.R., Z.A.M., J. T., and Q.Z. were supported by the Reducing Uncertainties in Biogeochemical Interactions through Synthesis and Computation (RUBISCO) Scientific Focus Area, Office of Biological and Environmental Research of the U.S. Department of Energy Office of Science. Lawrence Berkeley National Laboratory (LBNL) is managed by the University of California for the U.S. Department of Energy under Contract DE-AC02-05CH11231. J.T. was funded by the Next-Generation Ecosystem Experiments in the Arctic (NGEE-Arctic) project, supported by the Biological and Environmental Research (BER) Program in the U.S. Department of Energy Office of Science under contract DE-AC02-05CH11231. L.T.B. was supported by the NASA Arctic Boreal Vulnerability Experiment (ABoVE) under Grant 80NSSC22K1247. B.N.S. and F.Y. are supported by the Next Generation Ecosystem Experiments Arctic (NGEE Arctic) project of the U.S. Department of Energy Office of Science Office of Biological and Environmental Research. V.B. was supported by the European Research Council under the European Union's Horizon 2020 Research and Innovation programme as part of the Q-Arctic project (Grant agreement number 951288). G.G. acknowledges funding by the German Federal Ministry of Education and Research project Methane MOMENT (BMBF # 03F0931F). D.W. acknowledges financial support from the Strategic Research Area MERGE (Modeling the Regional and Global Earth System - www.merge.lu.se) and the European Union's Horizon Europe research and innovation programme under GreenFeedBack (Grant agreement number 101056921). A.H.M.D. acknowledges support from the Natural Science and Engineering Research Council of Canada and the Digital Alliance of Canada.

References

Amiro, B. (2023). AmeriFlux BASE CA-Man Manitoba - Northern old Black spruce (former BOREAS Northern study area), ver. 3-5, AmeriFlux AMP [Dataset]. https://doi.org/10.17190/AMF/1245997

Barr, A., & Black, A. T. (2018). AmeriFlux BASE CA-SJ2 Saskatchewan - Western Boreal, Jack Pine forest harvested in 2002, ver [Dataset]. *AmeriFlux AMP*. https://doi.org/10.17190/AMF/1436321

Berner, L. T., Orndahl, K. M., Rose, M., Tamstorf, M., Arndal, M. F., Alexander, H. D., et al. (2024). The Arctic plant aboveground biomass synthesis dataset. *Scientific Data*, *11*(1), 305. https://doi.org/10.1038/s41597-024-03139-w

Black, A. T. (2016a). AmeriFlux BASE CA-Oas Saskatchewan - Western boreal, mature aspen, ver. 1-1, AmeriFlux AMP [Dataset]. https://doi.org/10.17190/AMF/1375197

Black, A. T. (2016b). AmeriFlux BASE CA-Obs Saskatchewan - Western boreal, mature Black spruce, ver. 1-1, AmeriFlux AMP [Dataset]. https://doi.org/10.17190/AMF/1375198

Black, A. T. (2019). AmeriFlux BASE CA-Ojp Saskatchewan - Western boreal, mature Jack Pine [Dataset]. *Ver 2-5, AmeriFlux AMP*. https://doi.org/10.17190/AMF/1375199

Bouskill, N. J., Riley, W. J., & Tang, J. Y. (2014). Meta-analysis of high-latitude nitrogen-addition and warming studies implies ecological mechanisms overlooked by land models. *Biogeosciences*, *11*(23), 6969–6983. https://doi.org/10.5194/bg-11-6969-2014

Bouskill, N. J., Riley, W. J., Zhu, Q., Mekonnen, Z. A., & Grant, R. F. (2020). Alaskan carbon-climate feedbacks will be weaker than inferred from short-term experiments. *Nature Communications*, *11*(1), 5798. https://doi.org/10.1038/s41467-020-19574-3

Bracho, R., Celis, G., Rodenhizer, H., See, C., & Schuur, E. A. (2021). AmeriFlux BASE US-EML Eight Mile Lake Permafrost thaw gradient, Healy Alaska [Dataset]. *Ver. 4-5, AmeriFlux AMP*. https://doi.org/10.17190/AMF/1418678

21698961, 2026, 5, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025JG009307 by Helmholtz-Zentrum Potsdam GFZ, Wiley Online Library on [29/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Burke, E. J., Zhang, Y., & Krinner, G. (2020). Evaluating permafrost physics in the Coupled Model Intercomparison Project 6 (CMIP6) models and their sensitivity to climate change. *The Cryosphere*, *14*(9), 3155–3174. https://doi.org/10.5194/tc-14-3155-2020

Celis, G., Mauritz, M., Bracho, R., Salmon, V. G., Webb, E. E., Hutchings, J., et al. (2017). Tundra is a consistent source of CO2 at a site with progressive permafrost thaw during 6 years of chamber and eddy covariance measurements. *Journal of Geophysical Research: Biogeosciences*, *122*(6), 1471–1485. https://doi.org/10.1002/2016JG003671

Chen, B., Liu, S., Ge, J., & Chu, J. (2010). Annual and seasonal variations of Q10 soil respiration in the sub-alpine forests of the Eastern Qinghai-Tibet Plateau, China. *Soil Biology and Biochemistry*, *42*(10), 1735–1742. https://doi.org/10.1016/j.soilbio.2010.06.010

Chu, H., Christianson, D. S., Cheah, Y. W., Pastorello, G., O'Brien, F., Geden, J., et al. (2023). AmeriFlux BASE data pipeline to support network growth and data sharing. *Scientific Data*, *10*(1), 614. https://doi.org/10.1038/s41597-023-02531-2

Collier, N., Hoffman, F. M., Lawrence, D. M., Keppel-Aleks, G., Koven, C. D., Riley, W. J., et al. (2018). The international land model benchmarking (ILAMB) System: Design, theory, and implementation. *Journal of Advances in Modeling Earth Systems*, *10*(11), 2731–2754. https://doi.org/10.1029/2018MS001354

Davidson, E. A., & Janssens, I. A. (2006). Temperature sensitivity of soil carbon decomposition and feedbacks to climate change. *Nature*, *440*(7081), 165–173. https://doi.org/10.1038/nature04514

European Centre for Medium-Range Weather Forecasts (ECMWF). (2019). ERA5 reanalysis (0.25 degree latitude-longitude grid) [Dataset]. *Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory*. https://doi.org/10.5065/BH6N-5N20

Euskirchen, E. (2025a). AmeriFlux BASE US-BZF Bonanza Creek rich fen [Dataset]. *Ver. 5-5, AmeriFlux AMP*. https://doi.org/10.17190/AMF/1756433

Euskirchen, E. (2025b). AmeriFlux BASE US-BZS Bonanza Creek Black spruce [Dataset]. *Ver. 4-5, AmeriFlux AMP*. https://doi.org/10.17190/AMF/1756434

Euskirchen, E., Shaver, G., & Bret-Harte, S. (2025a). AmeriFlux BASE US-ICh Imnavait Creek watershed Heath Tundra [Dataset]. *Ver. 6-5, AmeriFlux AMP*. https://doi.org/10.17190/AMF/1246133

Euskirchen, E., Shaver, G., & Bret-Harte, S. (2025b). AmeriFlux BASE US-ICs Imnavait Creek watershed Wet Sedge tundra [Dataset]. *Ver. 10-5, AmeriFlux AMP*. https://doi.org/10.17190/AMF/1246130

Euskirchen, E. S., Bret-Harte, M. S., Shaver, G. R., Edgar, C. W., & Romanovsky, V. E. (2017). Long-term release of carbon dioxide from Arctic Tundra ecosystems in Alaska. *Ecosystems*, *20*(5), 960–974. https://doi.org/10.1007/s10021-016-0085-9

Euskirchen, E. S., Edgar, C. W., Turetsky, M. R., Waldrop, M. P., & Harden, J. W. (2014). Differential response of carbon fluxes to climate in three peatland ecosystems that vary in the presence and stability of permafrost. *Journal of Geophysical Research: Biogeosciences*, *119*(8), 1576–1595. https://doi.org/10.1002/2014JG002683

Fisher, J. B., Hayes, D. J., Schwalm, C. R., Huntzinger, D. N., Stofferahn, E., Schaefer, K., et al. (2018). Missing pieces to modeling the Arctic-Boreal puzzle. *Environmental Research Letters*, *13*(2), 020202. https://doi.org/10.1088/1748-9326/AA9D9A

Fisher, J. B., Huntzinger, D. N., Schwalm, C. R., & Sitch, S. (2014). Modeling the terrestrial biosphere. *Annual Review of Environment and Resources*, *39*(1), 91–123. https://doi.org/10.1146/ANNUREV-ENVIRON-012913-093456

Global Modeling and Assimilation Office (GMAO). (2015). MERRA-2 tavgM_2d_Ind_Nx: 2d,Monthlymean,Time-Averaged,Single-Level, Assimilation,Land surface diagnostics V5.12.4 [Dataset]. *Goddard Earth Sciences Data and Information Services Center (GES DISC)*. https://doi.org/10.5067/8S35XF81C28F

Hamdi, S., Moyano, F., Sall, S., Bernoux, M., & Chevallier, T. (2013). Synthesis analysis of the temperature sensitivity of soil respiration from laboratory studies in relation to incubation methods and soil conditions. *Soil Biology and Biochemistry*, *58*, 115–126. https://doi.org/10.1016/j.soilbio.2012.11.012

Han, M., & Jin, G. (2018). Seasonal variations of Q10 soil respiration and its components in the temperate forest ecosystems, northeastern China. *European Journal of Soil Biology*, *85*, 36–42. https://doi.org/10.1016/j.ejsobi.2018.01.001

Harris, I., Osborn, T. J., Jones, P., & Lister, D. (2020). Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset. *Scientific Data*, *7*(1), 109. https://doi.org/10.1038/s41597-020-0453-3

Hashimoto, S., Carvalhais, N., Ito, A., Migliavacca, M., Nishina, K., & Reichstein, M. (2015). Global spatiotemporal distribution of soil respiration modeled using a global database. *Biogeosciences*, *12*(13), 4121–4132. https://doi.org/10.5194/bg-12-4121-2015

Huntzinger, D. N., Schaefer, K., Schwalm, C., Fisher, J. B., Hayes, D., Stofferahn, E., et al. (2020). Evaluation of simulated soil carbon dynamics in Arctic-Boreal ecosystems. *Environmental Research Letters*, *15*(2), 025005. https://doi.org/10.1088/1748-9326/AB6784

Hursh, A., Ballantyne, A., Cooper, L., Maneta, M., Kimball, J., & Watts, J. (2017). The sensitivity of soil respiration to soil temperature, moisture, and carbon supply at the global scale. *Global Change Biology*, *23*(5), 2090–2103. https://doi.org/10.1111/gcb.13489

Iwahana, G., Kobayashi, H., Ikawa, H., & Suzuki, R. (2023). AmeriFlux BASE US-Prr poker flat research range Black spruce Forest [Dataset]. *Ver. 4-5, AmeriFlux AMP*. https://doi.org/10.17190/AMF/1246153

Johnston, A. S. A., Meade, A., Ardö, J., Arriga, N., Black, A., Blanken, P. D., et al. (2021). Temperature thresholds of ecosystem respiration at a global scale. *Nature Ecology & Evolution*, *5*(4), 487–494. https://doi.org/10.1038/s41559-021-01398-z

Kim, Y., Lee, B. Y., Suzuki, R., & Kushida, K. (2016). Spatial characteristics of ecosystem respiration in three tundra ecosystems of Alaska. *Polar Science*, *10*(3), 356–363. https://doi.org/10.1016/j.polar.2015.12.001

Kobayashi, S., Ota, Y., Harada, Y., Ebita, A., Moriya, M., Onoda, H., et al. (2015). The JRA-55 reanalysis: General specifications and basic characteristics. *Journal of the Meteorological Society of Japan*, *93*(1), 5–48. https://doi.org/10.2151/jmsj.2015-001

Koven, C. D., Hugelius, G., Lawrence, D. M., & Wieder, W. R. (2017). Higher climatological temperature sensitivity of soil carbon in cold than warm climates. *Nature Climate Change*, *7*(11), 817–822. https://doi.org/10.1038/nclimate3421

Lawrence, D. M., Fisher, R. A., Koven, C. D., Oleson, K. W., Swenson, S. C., Bonan, G., et al. (2019). The community land model version 5: Description of new features, benchmarking, and impact of forcing uncertainty. *Journal of Advances in Modeling Earth Systems*, *11*(12), 4245–4287. https://doi.org/10.1029/2018MS001583

Lawrence, D. M., Koven, C. D., Swenson, S. C., Riley, W. J., & Slater, A. G. (2015). Permafrost thaw and resulting soil moisture changes regulate projected high-latitude CO2 and CH4 emissions. *Environmental Research Letters*, *10*(9), 094011. https://doi.org/10.1088/1748-9326/10/9/094011

Lefèvre, R., Barré, P., Moyano, F. E., Christensen, B. T., Bardoux, G., Eglin, T., et al. (2014). Higher temperature sensitivity for stable than for labile soil organic carbon - Evidence from incubations of long-term bare fallow soils. *Global Change Biology*, *20*(2), 633–640. https://doi.org/10.1111/gcb.12402

Li, J., Pei, J., Pendall, E., Fang, C., & Nie, M. (2019). Spatial heterogeneity of temperature sensitivity of soil respiration: A global analysis of field observations. *Soil Biology and Biochemistry*, *141*, 107675. https://doi.org/10.1016/j.soilbio.2019.107675

21698961, 2026, 5, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025JG009307 by Helmholtz-Zentrum Potsdam GFZ, Wiley Online Library on [29/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Li, J., Pei, J., Pendall, E., Reich, P. B., Noh, N. J., Li, B., et al. (2020). Rising temperature May trigger deep soil carbon loss across Forest ecosystems. *Advanced Science*, *7*(19), 2001242. https://doi.org/10.1002/advs.202001242

Liu, H., Rezanezhad, F., Zhao, Y., He, H., Van Cappellen, P., & Lennartz, B. (2024). The apparent temperature sensitivity (Q10) of peat soil respiration: A synthesis study. *Geoderma*, *443*, 116844. https://doi.org/10.1016/j.geoderma.2024.116844

Lloyd, J., & Taylor, J. A. (1994). On the temperature dependence of soil respiration. In *Functional Ecology*, https://doi.org/10.2307/2389824

Luo, Y., Keenan, T. F., & Smith, M. (2015). Predictability of the terrestrial carbon cycle. *Global Change Biology*, *21*(5), 1737–1751. https://doi.org/10.1111/GCB.12766

Luo, Y. Q., Randerson, J. T., Abramowitz, G., Bacour, C., Blyth, E., Carvalhais, N., et al. (2012). A framework for benchmarking land models. *Biogeosciences*, *9*(10), 3857–3874. https://doi.org/10.5194/BG-9-3857-2012

MacDougall, A. H., Avis, C. A., & Weaver, A. J. (2012). Significant contribution to climate warming from the permafrost carbon feedback. *Nature Geoscience*, *5*(10), 719–721. https://doi.org/10.1038/ngeo1573

Maes, S. L., Dietrich, J., Midolo, G., Schwieger, S., Kummu, M., Vandvik, V., et al. (2024). Environmental drivers of increased ecosystem respiration in a warming tundra. *Nature*, *629*(8010), 105–113. https://doi.org/10.1038/s41586-024-07274-7

Mauritz, M., Bracho, R., Celis, G., Hutchings, J., Natali, S. M., Pegoraro, E., et al. (2017). Nonlinear CO2 flux response to 7 years of experimentally induced permafrost thaw. *Global Change Biology*, *23*(9), 3646–3666. https://doi.org/10.1111/GCB.13661

McGuire, A. D., Koven, C., Lawrence, D. M., Clein, J. S., Xia, J., Beer, C., et al. (2016). Variability in the sensitivity among model simulations of permafrost and carbon dynamics in the permafrost region between 1960 and 2009. *Global Biogeochemical Cycles*, *30*(7), 1015–1037. https://doi.org/10.1002/2016GB005405

McGuire, A. D., Lawrence, D. M., Koven, C., Clein, J. S., Burke, E., Chen, G., et al. (2018). Dependence of the evolution of carbon dynamics in the northern permafrost region on the trajectory of climate change. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(15), 3882–3887. https://doi.org/10.1073/PNAS.1719903115

Meyer, N., Welp, G., & Amelung, W. (2018). The temperature sensitivity (Q10) of soil respiration: Controlling factors and spatial prediction at regional scale based on environmental soil classes. *Global Biogeochemical Cycles*, *32*(2), 306–323. https://doi.org/10.1002/2017GB005644

Nagano, H., Ikawa, H., Nakai, T., Matsushima-Yashima, M., Kobayashi, H., Kim, Y., & Suzuki, R. (2018). Extremely dry environment down-regulates nighttime respiration of a black spruce forest in Interior Alaska. *Agricultural and Forest Meteorology*, *249*, 297–309. https://doi.org/10.1016/j.agrformet.2017.11.001

Natali, S. M., Schuur, E. A. G., Mauritz, M., Schade, J. D., Celis, G., Crummer, K. G., et al. (2015). Permafrost thaw and soil moisture driving CO2 and CH4 release from upland tundra. *Journal of Geophysical Research: Biogeosciences*, *120*(3), 525–537. https://doi.org/10.1002/2014JG002872

Natali, S. M., Schuur, E. A. G., Trucco, C., Hicks Pries, C. E., Crummer, K. G., & Baron Lopez, A. F. (2011). Effects of experimental warming of air, soil and permafrost on carbon balance in Alaskan tundra. *Global Change Biology*, *17*(3), 1394–1407. https://doi.org/10.1111/J.1365-2486.2010.02303.X

Nelson, J. A., Walther, S., Jung, M., Gans, F., Kraft, B., Weber, U., et al. (2023). FLUXOM-X-BASE [Dataset]. *COS*. https://doi.org/10.18160/5NZG-JMJE

Niu, S., Luo, Y., Fei, S., Montagnani, L., Bohrer, G., Janssens, I. A., et al. (2011). Seasonal hysteresis of net ecosystem exchange in response to temperature change: Patterns and causes. *Global Change Biology*, *17*(10), 3102–3114. https://doi.org/10.1111/j.1365-2486.2011.02459.x

Pallandt, M. M. T. A., Kumar, J., Mauritz, M., Schuur, E. A. G., Virkkala, A. M., Celis, G., et al. (2022). Representativeness assessment of the Pan-Arctic eddy covariance site network and optimized future enhancements. *Biogeosciences*, *19*(3), 559–583. https://doi.org/10.5194/BG-19-559-2022

Piao, S., Sitch, S., Ciais, P., Friedlingstein, P., Peylin, P., Wang, X., et al. (2013). Evaluation of terrestrial carbon cycle models for their response to climate variability and to CO2 trends. *Global Change Biology*, *19*(7), 2117–2132. https://doi.org/10.1111/gcb.12187

Poe, J., Huntzinger, D., Collier, N., Schwalm, C., Wells, J., Schädel, C., et al. (2025). Impacts of benchmarking choices on inferred model skill of the Arctic–Boreal terrestrial carbon cycle. *Environmental Research: Ecology*, *4*(1), 015007. https://doi.org/10.1088/2752-664X/adacee

Poe, J., Wells, J., Schädel, C., Huntzinger, D. N., & Riley, W. J. (2024). Simulating arctic carbon emissions in a warming world. *Eos*, *105*. https://doi.org/10.1029/2024EO240433

Qin, S., Kou, D., Mao, C., Chen, Y., Chen, L., & Yang, Y. (2021). Temperature sensitivity of permafrost carbon release mediated by mineral and microbial properties. In *Science Advances*. https://doi.org/10.1126/sciadv.abe3596

Randerson, J. T., Hoffman, F. M., Thornton, P. E., Mahowald, N. M., Lindsay, K., Lee, Y. H., et al. (2009). Systematic assessment of terrestrial biogeochemistry in coupled climate-carbon models. *Global Change Biology*, *15*(10), 2462–2484. https://doi.org/10.1111/j.1365-2486.2009.01912.x

Rantanen, M., Karpechko, A. Y., Lipponen, A., Nordling, K., Hyvärinen, O., Ruosteenoja, K., et al. (2022). The Arctic has warmed nearly four times faster than the globe since 1979. *Communications Earth & Environment*, *3*(1), 1–10. https://doi.org/10.1038/s43247-022-00498-3

Reichstein, M., & Beer, C. (2008). Soil respiration across scales: The importance of a model-data integration framework for data interpretation. *Journal of Plant Nutrition and Soil Science*, *171*(3), 344–354. https://doi.org/10.1002/jpln.200700075

Ren, S., Ding, J., Yan, Z., Cao, Y., Li, J., Wang, Y., et al. (2020). Higher temperature sensitivity of soil C release to atmosphere from Northern permafrost soils as indicated by a meta-analysis. *Global Biogeochemical Cycles*, *34*(11), e2020GB006688. https://doi.org/10.1029/2020GB006688

Rodenhizer, H. G., Mauritz, M., Taylor, M., Ledman, J., Natali, S., Schuur, E., & Bonanza, C. (2023). Eight mile Lake research watershed, carbon in permafrost experimental heating research (CiPEHR): Compiled half-hourly dataset, 2009-2021 ver 2 [Dataset]. *Environmental Data Initiative*. https://doi.org/10.6073/pasta/b418701259e674a8a89c8a2ba4ff6d6f

Schädel, C., Koven, C. D., Lawrence, D. M., Celis, G., Garnello, A. J., Hutchings, J., et al. (2018). Divergent patterns of experimental and model-derived permafrost ecosystem carbon dynamics in response to Arctic warming. *Environmental Research Letters*, *13*(10), 105002. https://doi.org/10.1088/1748-9326/AAE0FF

Schädel, C., Rogers, B. M., Lawrence, D. M., Koven, C. D., Brovkin, V., Burke, E. J., et al. (2024). Earth system models must include permafrost carbon processes. In *Nature Climate Change* (Vol. 14(2), pp. 114–116). Nature Research. https://doi.org/10.1038/s41558-023-01909-9

Schaefer, K., Lantuit, H., Romanovsky, V. E., Schuur, E. A. G., & Witt, R. (2014). The impact of the permafrost carbon feedback on global climate. *Environmental Research Letters*, *9*(8), 085003. https://doi.org/10.1088/1748-9326/9/8/085003

Schuur, E. A. G., Abbott, B. W., Commane, R., Ernakovich, J., Euskirchen, E., Hugelius, G., et al. (2022). Annual review of environment and resources permafrost and climate change: Carbon cycle feedbacks from the warming Arctic. https://doi.org/10.1146/annurev-environ-012220

Schuur, E. A. G., Bockheim, J., Canadell, J. G., Euskirchen, E., Field, C. B., Goryachkin, S. V., et al. (2008). Vulnerability of Permafrost carbon to climate change: Implications for the global carbon cycle. *BioScience*, *58*(8), 701–714. https://doi.org/10.1641/B580807

21698961, 2026, 5, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025JG009307 by Helmholtz-Zentrum Potsdam GFZ, Wiley Online Library on [29/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Schuur, E. A. G., Bracho, R., Celis, G., Belshe, E. F., Ebert, C., Ledman, J., et al. (2021). Tundra underlain by thawing Permafrost persistently emits carbon to the atmosphere over 15 years of measurements. *Journal of Geophysical Research: Biogeosciences*, *126*(6), e2020JG006044. https://doi.org/10.1029/2020JG006044

Schuur, E. A. G., Mcguire, A. D., Schädel, C., Grosse, G., Harden, J. W., Hayes, D. J., et al. (2015). Climate change and the permafrost carbon feedback. *Nature*, *520*(7546), 171–179. https://doi.org/10.1038/nature14338

Schwalm, C. R., Schaefer, K., Fisher, J. B., Huntzinger, D., Elshorbany, Y., Fang, Y., et al. (2019). Divergence in land surface modeling: Linking spread to structure. *Environmental Research Communications*, *1*(11), 111004. https://doi.org/10.1088/2515-7620/AB4A8A

Schwalm, C. R., Huntinzger, D. N., Michalak, A. M., Fisher, J. B., Kimball, J. S., Mueller, B., et al. (2013). Sensitivity of inferred climate model skill to evaluation decisions: A case study using CMIP5 evapotranspiration. *Environmental Research Letters*, *8*(2), 024028. https://doi.org/10.1088/1748-9326/8/2/024028

Sitch, S., Friedlingstein, P., Gruber, N., Jones, S. D., Murray-Tortarolo, G., Ahlström, A., et al. (2015). Recent trends and drivers of regional sources and sinks of carbon dioxide. *Biogeosciences*, *12*(3), 653–679. https://doi.org/10.5194/BG-12-653-2015

Sonnentag, O., & Quinton, W. L. (2021). AmeriFlux BASE CA-SCB scotty Creek bog, ver. 2-5, AmeriFlux AMP [Dataset]. https://doi.org/10.17190/AMF/1498754

Sonnentag, O., & Quinton, W. L. (2025). AmeriFlux BASE CA-SCC scotty Creek landscape [Dataset]. *Ver. 2-5, AmeriFlux AMP*. https://doi.org/10.17190/AMF/1480303

Stoy, P. C., Katul, G. G., Siqueira, M. B. S., Juang, J. Y., Novick, K. A., Uebelherr, J. M., & Oren, R. (2006). An evaluation of models for partitioning eddy covariance-measured net ecosystem exchange into photosynthesis and respiration. *Agricultural and Forest Meteorology*, *141*(1), 2–18. https://doi.org/10.1016/j.agrformet.2006.09.001

Sun, W., Luo, X., Fang, Y., Shiga, Y. P., Zhang, Y., Fisher, J. B., et al. (2023). Biome-scale temperature sensitivity of ecosystem respiration revealed by atmospheric CO2 observations. *Nature Ecology & Evolution*, *7*(8), 1199–1210. https://doi.org/10.1038/s41559-023-02093-x

Suseela, V., Conant, R. T., Wallenstein, M. D., & Dukes, J. S. (2012). Effects of soil moisture on the temperature sensitivity of heterotrophic respiration vary seasonally in an old-field climate change experiment. *Global Change Biology*, *18*(1), 336–348. https://doi.org/10.1111/j.1365-2486.2011.02516.x

Tang, J., & Riley, W. J. (2015). Weaker soil carbon-climate feedbacks resulting from microbial and abiotic interactions. *Nature Climate Change*, *5*(1), 56–60. https://doi.org/10.1038/nclimate2438

Tang, X., Fan, S., Zhang, W., Gao, S., Chen, G., & Shi, L. (2019). Global variability in belowground autotrophic respiration in terrestrial ecosystems. *Earth System Science Data*, *11*(4), 1839–1852. https://doi.org/10.5194/essd-11-1839-2019

Todd-Brown, K., Zheng, B., & Crowther, T. W. (2018). Field-warmed soil carbon changes imply high 21st-century modeling uncertainty. In *Biogeosciences* (Vol. 15(12), pp. 3659–3671). Copernicus GmbH. https://doi.org/10.5194/bg-15-3659-2018

Todd-Brown, K. E. O., Randerson, J. T., Post, W. M., Hoffman, F. M., Tarnocai, C., Schuur, E. A. G., & Allison, S. D. (2013). Causes of variation in soil carbon simulations from CMIP5 Earth system models and comparison with observations. *Biogeosciences*, *10*(3), 1717–1736. https://doi.org/10.5194/BG-10-1717-2013

Torn, M. S., Abramoff, R. Z., Vaughn, L. J. S., Chafe, O. E., Curtis, J. B., & Zhu, B. (2025). Large emissions of CO2 and CH4 due to active-layer warming in Arctic tundra. *Nature Communications*, *16*(1), 124. https://doi.org/10.1038/s41467-024-54990-9

Ueyama, M., Iwata, H., & Harazono, Y. (2023). AmeriFlux BASE US-Fcr Cascaden Ridge fire scar [Dataset]. *Ver. 3-5, AmeriFlux AMP*. https://doi.org/10.17190/AMF/1562388

Ueyama, M., Iwata, H., & Harazono, Y. (2025a). AmeriFlux BASE US-Rpf Poker Flat Research Range: Succession from fire scar to deciduous forest [Dataset]. *Ver. 11-5, AmeriFlux AMP*. https://doi.org/10.17190/AMF/1579540

Ueyama, M., Iwata, H., & Harazono, Y. (2025). AmeriFlux BASE US-Uaf university of Alaska, fairbanks, ver. 13-5, AmeriFlux AMP [Dataset]. https://doi.org/10.17190/AMF/1480322

Ueyama, M., Iwata, H., Nagano, H., Tahara, N., Iwama, C., & Harazono, Y. (2019). Carbon dioxide balance in early-successional forests after forest fires in interior Alaska. *Agricultural and Forest Meteorology*, *275*, 196–207. https://doi.org/10.1016/j.agrformet.2019.05.020

Varney, R. M., Chadburn, S. E., Burke, E. J., & Cox, P. M. (2022). Evaluation of soil carbon simulation in CMIP6 Earth system models. *Biogeosciences*, *19*(19), 4671–4704. https://doi.org/10.5194/bg-19-4671-2022

Virkkala, A.-M., Rogers, B. M., Watts, J. D., Arndt, K., Potter, S., Wargowsky, I., & Natali, S. (2024). Machine learning-based Arctic-boreal terrestrial ecosystem CO2 fluxes, 2001-2020 [Dataset]. *ORNL DAAC*. https://doi.org/10.3334/ORNLDAAC/2377

Virkkala, A.-M., Wargowsky, I., Vogt, J., Kuhn, M. A., Madaan, S., O'Keefe, R., et al. (2025). ABCFlux v2: Arctic–boreal CO 2 and CH 4 monthly flux observations and ancillary information across terrestrial and freshwater ecosystems. https://doi.org/10.5194/essd-2025-585

Wang, J., Wu, Q., Yuan, Z., & Kang, H. (2020). Soil respiration of alpine meadow is controlled by freeze-thaw processes of active layer in the permafrost region of the Qinghai-Tibet Plateau. *The Cryosphere*, *14*(9), 2835–2848. https://doi.org/10.5194/tc-14-2835-2020

Wang, Y., Liu, H., Chung, H., Yu, L., Mi, Z., Geng, Y., et al. (2014). Non-growing-season soil respiration is controlled by freezing and thawing processes in the summer monsoon-dominated Tibetan alpine grassland. *Global Biogeochemical Cycles*, *28*(10), 1081–1095. https://doi.org/10.1002/2013GB004760

Wells, J. M., Poe, J., Huntzinger, D., Riley, W. J., Schuur, E. A. G., Bouskill, N. J., et al. (2026). Warming permafrost model Intercomparison Project (WrPMIP): 2000-2021 pan-arctic perturbation warming simulations [Dataset]. *ESGF*. https://doi.org/10.25981/ESGF.WrPMIP.CMIP6/3007981

Wilkman, E., Zona, D., Tang, Y., Gioli, B., Lipson, D. A., & Oechel, W. (2018). Temperature response of respiration across the heterogeneous landscape of the Alaskan Arctic Tundra. *Journal of Geophysical Research: Biogeosciences*, *123*(7), 2287–2302. https://doi.org/10.1029/2017JG004227

Wu, D., Liu, S., Wu, X., Yang, X., Xu, T., Xu, Z., & Shi, H. (2021). Diagnosing the temperature sensitivity of ecosystem respiration in Northern high-latitude regions. *Journal of Geophysical Research: Biogeosciences*, *126*(4), e2020JG005998. https://doi.org/10.1029/2020JG005998

Xia, J., Wang, J., & Niu, S. (2020). Research challenges and opportunities for using big data in global change biology. In *Global Change Biology* (Vol. 26(11), pp. 6040–6061). Blackwell Publishing Ltd. https://doi.org/10.1111/gcb.15317

Xue, K., Yuan, M. M., Shi, Z. J., Qin, Y., Deng, Y., Cheng, L., et al. (2016). Tundra soil carbon is vulnerable to rapid microbial decomposition under climate warming. *Nature Climate Change*, *6*(6), 595–600. https://doi.org/10.1038/nclimate2940

Zhan, W., Yang, X., Ryu, Y., Dechant, B., Huang, Y., Goulas, Y., et al. (2022). Two for one: Partitioning CO2 fluxes and understanding the relationship between solar-induced chlorophyll fluorescence and gross primary productivity using machine learning. *Agricultural and Forest Meteorology*, *321*, 108980. https://doi.org/10.1016/j.agrformet.2022.108980

21698961, 2026, 5, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025JG009307 by Helmholtz-Zentrum Potsdam GFZ, Wiley Online Library on [29/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

References From the Supporting Information

Burke, E. J., Chadburn, S. E., & Ekici, A. (2017). A vertical representation of soil carbon in the JULES land surface scheme (Vn4.3-permafrost) with a focus on permafrost regions. *Geoscientific Model Development*, *10*(2), 959–975. https://doi.org/10.5194/gmd-10-959-2017

Burrows, S. M., Maltrud, M., Yang, X., Zhu, Q., Jeffery, N., Shi, X., et al. (2020). The DOE E3SM v1.1 Biogeochemistry configuration: Description and simulated ecosystem-climate responses to historical changes in forcing. *Journal of Advances in Modeling Earth Systems*, *12*(9), e2019MS001766. https://doi.org/10.1029/2019MS001766

Chadburn, S. E., Krinner, G., Porada, P., Bartsch, A., Beer, C., Belelli Marchesini, L., et al. (2017). Carbon stocks and fluxes in the high latitudes: Using site-level data to evaluate Earth system models. *Biogeosciences*, *14*(22), 5143–5169. https://doi.org/10.5194/bg-14-5143-2017

Grant, R. F., Humphreys, E. R., & LaFleur, P. M. (2015). Ecosystem CO2 and CH4 exchange in a mixed tundra and a fen within a hydrologically diverse Arctic landscape: 1. Modeling versus measurements. *Journal of Geophysical Research: Biogeosciences*, *120*(7), 1366–1387. https://doi.org/10.1002/2014JG002888

Grant, R. F., Mekonnen, Z. A., Riley, W. J., Wainwright, H. M., Graham, D., & Torn, M. S. (2017). Mathematical modelling of Arctic Polygonal tundra with ecosys: 1. Microtopography determines how active layer depths respond to changes in temperature and precipitation. *Journal of Geophysical Research: Biogeosciences*, *122*(12), 3161–3173. https://doi.org/10.1002/2017JG004035

Guimberteau, M., Zhu, D., Maignan, F., Huang, Y., Yue, C., Dantec-N d lec, S., et al. (2018). ORCHIDEE-MICT (v8.4.1), a land surface model for the high latitudes: Model description and validation. *Geoscientific Model Development*, *11*(1), 121–163. https://doi.org/10.5194/gmd-11-121-2018

Lee, H., Swenson, S. C., Slater, A. G., & Lawrence, D. M. (2014). Effects of excess ground ice on projections of permafrost in a warming climate. *Environmental Research Letters*, *9*(12), 124006. https://doi.org/10.1088/1748-9326/9/12/124006

Melton, J. R., Arora, V. K., Wisernig-Cojoc, E., Seiler, C., Fortier, M., Chan, E., & Teckentrup, L. (2020). CLASSIC v1.0: The open-source community successor to the Canadian Land Surface Scheme (CLASS) and the Canadian Terrestrial Ecosystem Model (CTEM)-Part 1: Model framework and site-level performance. *Geoscientific Model Development*, *13*(6), 2825–2850. https://doi.org/10.5194/gmd-13-2825-2020

Mengis, N., Keller, D. P., Macdougall, A. H., Eby, M., Wright, N., Meissner, K. J., et al. (2020). Evaluation of the University of Victoria Earth System Climate Model version 2.10 (UVic ESCM 2.10). *Geoscientific Model Development*, *13*(9), 4183–4204. https://doi.org/10.5194/gmd-13-4183-2020

Pongracz, A., Wårlind, D., Miller, P. A., & Parmentier, F. J. W. (2021). Model simulations of arctic biogeochemistry and permafrost extent are highly sensitive to the implemented snow scheme in LPJ-GUESS. *Biogeosciences*, *18*(20), 5767–5787. https://doi.org/10.5194/bg-18-5767-2021

Reick, C., Gayler, V., Goll, D., Hagemann, S., Heidkamp, M., Nabel, J., et al. (2021). *Jsbach 3 – The land component of the MPI Earth System Model: Documentation of version 3.2*. (Technical Report). MPI für Meteorologie.

Smith, B., Wärlind, D., Arneth, A., Hickler, T., Leadley, P., Siltberg, J., & Zaehle, S. (2014). Implications of incorporating N cycling and N limitations on primary production in an individual-based dynamic vegetation model. *Biogeosciences*, *11*(7), 2027–2054. https://doi.org/10.5194/bg-11-2027-2014

Wiltshire, A. J., Carolina Duran Rojas, M., Edwards, J. M., Gedney, N., Harper, A. B., Hartley, A. J., et al. (2020). JULES-GL7: The Global Land configuration of the Joint UK Land Environment Simulator version 7.0 and 7.2. *Geoscientific Model Development*, *13*(2), 483–505. https://doi.org/10.5194/gmd-13-483-2020

Xi, Y., Qiu, C., Zhang, Y., Zhu, D., Peng, S., Hugelius, G., et al. (2024). Assessment of a tiling energy budget approach in a land surface model, ORCHIDEE-MICT (r8205). *Geoscientific Model Development*, *17*(12), 4727–4754. https://doi.org/10.5194/gmd-17-4727-2024

Zhu, Q., Riley, W. J., Tang, J., Collier, N., Hoffman, F. M., Yang, X., & Bisht, G. (2019). Representing nitrogen, phosphorus, and carbon interactions in the E3SM land model: Development and global benchmarking. *Journal of Advances in Modeling Earth Systems*, *11*(7), 2238–2258. https://doi.org/10.1029/2018MS001571

21698961, 2026, 5, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025JG009307 by Helmholtz-Zentrum Potsdam GFZ, Wiley Online Library on [29/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License