

High-resolution soil moisture mapping in northern boreal forests using SMAP data and downscaling techniques

Emmihenna Jääskeläinen¹, Miska Luoto², Pauli Putkiranta³, Mika Aurela¹, and Tarmo Virtanen³

¹Finnish Meteorological Institute, Erik Palmenin aukio 1, 00560 Helsinki, Finland

²Department of Geosciences and Geography, University of Helsinki, P.O. Box 64, Helsinki 00014, Finland

³Ecosystems and Environment Research Programme, Faculty of Biological and Environmental Sciences, University of Helsinki, P.O. Box 65, Helsinki 00014, Finland

Correspondence: Emmihenna Jääskeläinen (emmihenna.jaaskelainen@fmi.fi)

Abstract.

Soil moisture plays an important part in predicting different forest-related phenomena, such as tree growth or forest fire risk. As ~~these phenomena~~ they influence the carbon storage capacity of boreal forest ecosystems, it is crucial to provide soil moisture information at high ~~temporal and spatial~~ spatio-temporal scales. Current satellite-based soil moisture products often have high temporal resolution at the expense of spatial resolution. Therefore, we developed a machine-learning-based model to estimate soil moisture at high ~~temporal and high~~ spatial resolution over boreal forested areas for the annual time period from May to October, while retaining the high temporal resolution. The basis data of the model is the ~~enhanced 9–36~~ km spatial resolution soil moisture data from the Soil Moisture Active Passive (SMAP) mission. Additionally, ~~soil and~~ vegetation properties, ~~reanalysis-based~~ weather-related parameters, and measured in situ soil moisture data are used to guide the model construction process. The analysis of the developed model shows that ~~the model~~ it retains the temporal and large-scale spatial variability of SMAP soil moisture. Furthermore, comparisons with the independent in situ soil moisture data ~~show that the soil moisture values predicted by the developed model have a better agreement~~ indicate that the model's predictions align more closely with in situ values than SMAP soil moisture, as RMSE decreases from ~~0.097–0.103~~ 0.065–0.092 m³/m³ to ~~0.065–0.092~~ 0.046 to 0.55 m³/m³, and correlation increases from ~~0.30 to 0.52~~ 0.46 to 0.55 over forest sites. Therefore, this machine-learning-based model can be used to predict high-resolution soil moisture over boreal forested areas.

1 Introduction

Boreal forest ecosystems are important carbon sinks and stocks (Pan et al. (2011), Pan et al. (2024)). Trees, mineral soil, and organic layer account for about 70% of the carbon pool in boreal forests (Merilä et al. (2023)). Trees remove carbon dioxide from the atmosphere through photosynthesis, turn it into organic carbon compounds, and use them for growing. Carbon is stored in all parts of the tree, i.e. in branches, stems, leaves, bark, and roots (e.g. Clemmensen et al. (2013), Thurner et al. (2014)). This carbon stored in the boreal ecosystems is released back into the atmosphere ~~for example~~ due to forest fires (Walker et al. (2019)) and the decomposition of trees, turning forests from carbon sinks to sources. As soil moisture plays a significant role in

predicting tree growth (Larson et al. (2024)), forest fire risks (Walker et al. (2019)), and carbon stock partitioning (Larson et al. (2023)), it is essential to provide soil moisture data at a large spatial scale and high temporal frequency across boreal forests.

Due to the considerable local variation in soil moisture and the sparsity of the in-situ measurement network, the only viable way to extensively observe soil moisture over boreal forests is to use satellite-based soil moisture data sets. However, persistent cloud cover, other weather-related phenomena, and high solar zenith angles hinder the use of optical satellite-based soil moisture data, making microwave-based soil moisture the most feasible option. For example, Sentinel-1 C-band Synthetic Aperture data (SAR) has been used to retrieve soil moisture with high resolution (e.g. Bauer-Marschallinger et al. (2019), Balenzano et al. (2021), Manninen et al. (2022)), but the dense vegetation prevents the radar signal from reaching the soil surface (Flores et al. (2019)), and thus causes uncertainty in the results (Bauer-Marschallinger et al. (2019), Flores et al. (2019)). A longer wavelength band, like L-band, can penetrate the vegetation ~~all the way to to reach~~ the soil surface (Flores et al. (2019)), and possibly ~~also even~~ deeper than the documented -5 cm depth in the boreal forest (Ambadan et al. (2022)). The well-known L-band-based soil moisture mission Soil Moisture Active Passive (SMAP, <https://smap.jpl.nasa.gov/>) has been measuring soil moisture globally from 2015 onwards, and has been reported to be sensitive to soil moisture changes under the forest canopy (Colliander et al. (2020), Ayres et al. (2021)). The disadvantage of soil moisture data from SMAP is that the ~~original~~ spatial resolution is ~~very coarse, with grid size being coarse~~, 36 km (Entekhabi et al. (2014)). SMAP soil moisture data has been ~~enhanced~~ regridded to 9 km, but as soil moisture ~~can change a lot even in a short distance is known to be spatially heterogeneous~~ (Mälicke et al. (2020)), there is a need for soil moisture data in a finer spatial resolution. SMAP L-band radiometer data has been combined with the Sentinel-1 C-band radar data to obtain higher resolution (1 km and 3 km) soil moisture data (O'Neill et al. (2021), Das et al. (2019)). The unbiased root-mean-square-error (RMSE) of this combined data is around 0.05 m³/m³ (Das et al. (2019)). The limitations of this data set include its temporal frequency, which is around 6 days over Europe, and only around 12 days elsewhere, and also its limited global coverage, as the product covers only the area between -60°S and 60°S, excluding the most of the boreal forest zone.

~~Instead~~ In addition to of directly using satellite-based measurements to retrieve soil moisture, another approach is downscaling. This involves enhancing coarse-resolution soil moisture data to a finer spatial scale using regression or more advanced machine learning methods. Downscaling has been used widely (e.g. Peng et al. (2017), Sabaghy et al. (2018), and references therein) with promising results. A few examples of downscaled soil moisture in 1 km spatial resolution include GLASS SM (Zhang et al. (2023)), which is based on ERA5-Land soil moisture; an over 20-year gap-free global and daily soil moisture data set (Zheng et al. (2023)) based on ESA-CCI soil moisture; and downscaled SMAP (Fang et al. (2022)). Since the original data sets have very coarse spatial resolutions, the downscaled data sets typically aim for a spatial resolution of 1 km or coarser. ~~However, a disadvantage of these downsealed data sets is that they are not specifically designed for boreal forests.~~

~~In this study,~~

Our main goal is to develop a model for estimating high-resolution soil moisture data over boreal forests of Northern Finland. There we can use two elements to our advantage. First is the SMAP soil moisture retrieval process, which is based on the dominant land use classification in each pixel. In the studied area, the SMAP soil moisture is most accurate for forested areas with a shrub and herb dominated understory, because, by definition, the dominant land cover classification is woody

savanna (i.e. a herbaceous understory and forest canopy cover between 30-60%; different class definitions can be found in Strahler et al. (1999)). The second element is that most available in situ sites are located in forested sites. By combining these two details we provide a model optimized to calculate a high resolution (1 km and 250 m pixel-sized) soil moisture for boreal forests. We use SMAP soil moisture data in 936 km spatial resolution as the basis, and we combine SMAP soil moisture with in situ soil moisture observations. Other high-resolution and soil moisture linked parameters (like vegetation and soil properties) are added to our downscaling machine learning model to guide the process. ~~The coarse resolution of SMAP, the small amount of in situ measuring sites over boreal forest (i.e. low variability in static input parameters), and the mainly forested locations of the in situ sites, hinder the model construction process, and therefore our model is limited to forested areas, excluding other areas like peatlands.~~

This paper is constructed as follows. First, all the used data sets are introduced in Section 2, followed by preprocessing steps and model construction in Section 3. In Section 4, the results of the model analysis and model validation are shown. We conclude with a discussion and conclusions in Sections 5 and 6, respectively.

2 Data

In this study, we used in situ data from two ~~larger boreal forests~~ large boreal forest in situ observation areas with easily accessible data, one in Northern Finland (NF), operated by the Finnish Meteorological Institute (FMI), and the other in Alaska, operated by multiple networks (see section 2.9 below). For model construction, we decided to use NF sites, leaving the Alaska sites for validation. The locations of the sites are shown in Figure 1. Based on the observations between the years 2019 and 2023 from the weather stations located in the NF (<https://www.ilmatiiteenlaitos.fi/havaintojen-lataus>), there is snow cover typically from mid-October to May, depending on the site and location. Therefore, ~~for the study time period,~~ we chose the annual ~~period interval~~ spanning from the first of May to 15 of October, ~~covering the years 2019–2023~~ as the study period. From here onwards, the use of the soil moisture term indicates volumetric water content (%).

2.1 Study area

The main study area is located in Northern Finland, between latitudes 65.5°N and 70.0°N, placing it in the boreal forest biome. Based on the land cover classification data in 100 m resolution from CORINE land cover (see Section 2.8) for NF (area shown in Figure 1), around 61% of the area is covered by tree cover (42% coniferous trees, 10% broadleaved trees, and 9% mixed trees). One fifth of the area (almost 20%) is covered in peat bogs, and 4.5% is covered by water bodies (lakes). Water bodies include three large lakes (Lake Inari, Lokka Reservoir, and Porttipahta Reservoir), but also many smaller lakes. The rest of the land use (around 15%) is different urban areas, heatlands, bare areas, and agriculture.

2.2 SMAP soil moisture

The SMAP mission was meant to combine radiometer (passive) and radar (active) observations. However, since the radar broke down just months after the launch, the radiometer is currently the only instrument observing the surface. This SMAP L-band

(1.41 GHz) radiometer has a native spatial footprint of 36 km ~~, but the footprint is enhanced to 9 km resolution by using the~~
90 ~~Backus-Gilbert optimal interpolation algorithm (Chan et al. (2018)). The enhanced products are~~ and the data is provided on
the global cylindrical EASE-Grid 2.0 (Brodzik et al. (2012)).

SMAP soil moisture is based on retrieved brightness temperature data in horizontal and vertical polarizations (O'Neill et al.
(2021)). Water body correction is applied to the brightness temperature data first to remove water bodies, as they lower the
brightness temperature values and hence cause overestimated soil moisture values. Then tau-omega-model (tau, vegetation
95 optical depth τ and, omega, vegetation single scattering albedo, ω) is applied to the single channel (horizontal and vertical)
brightness temperature data to separate soil and vegetation contributions from the total brightness temperature. After that,
soil moisture is retrieved by inversion from the tau-omega corrected brightness temperature. Land use classification data is
used to determine the τ and ω values for different areas. For dual-channel retrieved soil moisture, the tau-omega corrected
single-channel brightness temperature data is used.

100 In this study, we use the ~~enhanced SMAP L3_SM_P_E product (O'Neill et al. (2023b))~~ SMAP SPL3SMP V009 product
(O'Neill et al. (2023a)), in which the global surface soil moisture (0–5 cm) in m^3/m^3 is provided twice a day, at 6:00 am
(descending) and at 6:00 pm (ascending). Three different soil moisture products are available, one calculated from each single
channel and one dual-channel product. As the latter one is currently the baseline product (Chan and Dunbar (2021)), we chose
that for this study. Further, we ~~chose the north polar maps as they cover the boreal zone, and we~~ focus on the soil moisture data
105 at 6 am (descending overpasses). This SMAP data in 936 km resolution is used as an input for the soil moisture model.

2.3 MODIS

The Moderate Resolution Imaging Spectroradiometer (MODIS) instruments are aboard the Terra and Aqua satellites, which
were launched in 1999 and 2002, respectively. As Sun-synchronous satellites, they provide almost global coverage every 1 to 2
days. Terra is set to a descending orbit (measurements at 10.30 am) and Aqua to an ascending orbit (measurements at 1.30 pm).

110 The MODIS instrument measures multiple wavelength bands, resulting in a wide range of obtained parameters. In this study,
we use vegetation indices from both MODIS instruments. We use MYD13Q1 (from Aqua, Didan (2021a)) and MOD13Q1
(from Terra, Didan (2021b)) products (version 6.1) which are global 16-day-mean data sets with 250 m spatial resolution. The
~~used data, enhanced vegetation index data used, the Enhanced Vegetation Index~~ (EVI) and the Normalized Vegetation Index
(NDVI), are provided in the Sinusoidal tile grid. EVI and NDVI contribute to the vegetation effects of the soil moisture model.

115 2.4 SMAP-based 1 km soil moisture data

SMAP, enhanced to 9 km spatial resolution, was further downscaled to 1 km (Lakshmi and Fang (2023)) by using thermal
inertia theory (Fang et al. (2022)). Based on that theory, the land surface temperature (LST) difference between night and day
is negatively correlated to the soil moisture. For downscaling SMAP, the MODIS LST data in 1 km spatial resolution from
Terra (night) and Aqua (day) were used, combined with the NDVI, also from MODIS. The NDVI, divided into 10 groups by
120 using an interval of 0.1, is used for grouping soil moisture and LST differences. The assumption behind this is that changes in
NDVI affect the relationship between soil moisture and LST difference. Based on the validation, the downscaled SMAP data

performs better in low latitudes and warm months, compared to high latitudes and cold months (Fang et al. (2022)). The SMAP in 1 km resolution is used in this study as an example of downscaled data based on SMAP ~~in 9 km resolution. data.~~

2.5 Interpolated daily weather observations

125 Finnish Meteorological Institute provides different weather-related parameters in spatial resolution of 1 km (<https://en.ilmatieteenlaitos.fi/gridded-observations-on-aws-s3>), covering the time period starting from 1961 through the present day. Daily weather station observations are interpolated into a 1 km × 1 km grid by using kriging with external drift. In that method, external predictors are used as covariates. Elevation, relative altitude, the effect of the seas, and the effect of the lakes are the chosen external predictors for these weather station observation based maps (Aalto et al. (2016)). Daily precipitation sum and daily mean temperature are
130 used as inputs for the soil moisture model.

2.6 GPM

The Global Precipitation Measurement mission (GPM, <https://gpm.nasa.gov/missions/GPM>) is a network of satellites, aiming to provide precipitation observations every 2-3 hours. This is achieved by using active radar observations and passive microwave radiometer measurements. Precipitation data is provided in multiple levels and processing steps, of which we use the level 3
135 Integrated Multi-satellitE Retrievals for GPM (IMERG) Final Run data. This data is based on intercalibrated data from all microwave precipitation estimates, and microwave-calibrated infrared satellite estimates, as well as bias corrected by using precipitation gauge analyses (Huffman et al. (2023)). The Final Run product is provided in either 30-minute intervals or daily and monthly means. The spatial resolution is 0.1° (around 10 km). For this study, we use daily means of precipitation.

2.7 ERA5-Land

140 ERA5-Land, the land component of the fifth generation of European Centre for Medium-Range Weather Forecasts (ECMWF) atmospheric reanalysis of the global climate (ERA5), is produced by the Copernicus Climate Change Service. Similarly to ERA5, the land component also covers the period from 1940 to the present day (Hersbach et al. (2020)), but with enhanced spatial resolution (from 31 km to 9 km). ERA5-Land provides hourly data of various surface parameters, of which we used ~~total precipitation (m), air temperature at 2 m above the surface (K), and surface net solar radiation (J/m²) as input's for the~~
145 ~~soil moisture model. They provided the weather contribution to the model.~~

2.8 Soil properties

~~SoilGrids version 2.0 provides multiple soil properties globally in 250 m spatial resolution. These data sets are constructed by using quantile regression forests to about 240 000 in-situ soil observations and numerous environmental covariates (Poggio et al. (2021)). From the resulting data sets, we chose bulk density (cg/cm³) and silt content (g/kg) at the depth interval 0–5 cm to be used as~~
150 ~~input in models. As static inputs, they provides the invariable soil contribution to the model. SoilGrids also provide uncertainty information for each soil property, and therefore they were acquired as well.~~

2.8 CORINE land cover

The Coordination of Information on the Environment (CORINE) program was launched in the 1980s, as there was a need for detailed and harmonized land cover data set over the European continent (Büttner et al. (2017)). The current land cover data covers the pan-European area with 100 m spatial resolution. The data set consists of 44 classes, and it is updated every six years. In this study, we use CORINE land cover data from 2018 to determine the land use classifications of the study area, the land cover classes of the used in situ sites in the NF area, and we also used land cover data to create a mask to exclude water bodies and all the other land covers except forested areas.

2.9 In situ data

In situ soil moisture data for model training and testing are from the Arctic Space Centre of the Finnish Meteorological Institute (FMI-ARC, <https://fmiarc.fmi.fi/>). FMI-ARC hosts a measurement infrastructure, which is used to monitor, for example, the atmosphere, soil properties, snow properties, precipitation, and carbon and water cycles. All collected observations can be found at <https://litdb.fmi.fi/> (last accessed: 30 September 2024). For in situ soil moisture observations, the measurement sites are located around Sodankylä and Saariselkä, and they cover mostly boreal forested sites. The chosen in situ sites with additional information can be found in Table A1, and their locations are shown in Figure 1. The in situ soil moisture is measured at different depths, and for this study, we chose a depth of -5 cm.

The in situ data for validation of the constructed model are located in the boreal zone in Alaska. In situ soil moisture data has been collected in the International Soil Moisture Network (ISMN) database starting from 1952 (Dorigo et al. (2013), Dorigo et al. (2021)). In situ soil moisture data is provided to the ISMN by multiple organizations for free use. ~~We~~ From ISMN we chose 16 stations located in the boreal zone (Figure 1, right side), and information about those sites can be found in Table A2. ~~We focused on the year 2019, as all 16 stations had observations during that year.~~ Similarly to NF sites, the in situ soil moisture is also measured ~~in at~~ different depths, and for this study, we chose a depth of -5 cm.

Additionally, we included one in situ site from the U.S National Science Foundation's National Ecological Observatory Network (NEON). NEON has multiple measurement sites around the United States, of which 5 are located in Alaska. From these 5 sites, we chose Delta Junction as its dominant land use class is evergreen forest. The used data product is DP1.00094.001 (National Ecological Observatory Network (2025)), in which soil volumetric water content is included in various depths. For this study, we chose a depth of -6 cm, which is the shallowest one.

3 Methods

3.1 Preprocessing

All gridded data used (SMAP, EVI, NDVI, ~~ERA5-Land data, and soil properties~~ and interpolated weather station observations) are reprojected to the global EASE2-grid if needed and resampled to achieve a spatial resolution of 1 km. This means that the projection matches that of SMAP, but the spatial resolution is finer than that of SMAP. If the original resolution is coarser than

the resampled one, the resampling is done by using the nearest neighbor. On the other hand, if the original resolution is finer than the resampled one, then the resampling is done by taking the average of all values within the coarser pixel. The average is taken even if there is only one value within the coarser pixel. This ensures that the model inputs have a minimal number of missing values.

After resampling and reprojecting, some of the data are further preprocessed. As EVI and NDVI data from both MODIS instruments are originally provided every 16 days, we obtain daily maps of EVI and NDVI by linear interpolation over time using the closest available observations. The linear interpolation was chosen because it is easy to implement and does not cause any major discrepancies in the interpolated data for vegetation types with weak seasonal changes, such as evergreen needle-leaved forests (Li et al. (2021)).

After interpolation, we calculate the mean value of Terra and Aqua -based vegetation maps to obtain only one EVI and NDVI map per day. Precipitation, ~~air temperature, and surface net solar radiation from ERA5-Land are provided for every hour. Instead of using instantaneous precipitation and air temperature from interpolated weather station data are provided as~~ daily means. Based on preliminary testing, we decided to use a precipitation sum of 2 and 7 days preceding each SMAP observation, and a temperature sum of 8 and ~~air temperature reanalysis values, we calculated a sum over 72 hours before 10 days preceding~~ each SMAP observation instead of using just the daily means of one previous day. This approach takes into account the cumulative effects of temperature and precipitation. ~~From surface net solar radiation, we used instantaneous values only.~~

In situ data for training and testing was cleaned by removing those stations and those years where soil moisture values were abnormally low (below 0.05 continuously, or decreased to zero regularly), as including those values might lead to the model underestimating soil moisture. Also, there are two in situ sites located in or close to the peatlands, where soil moisture values of those sites are extremely high ($> 0.75 \text{ m}^3/\text{m}^3$). Including those locations in the training set caused the model to predict erroneous soil moisture values. Therefore, those two sites were excluded from the study data set.

After preprocessing and data cleaning, all the gridded data are matched with NF in situ locations. If there are multiple in situ values within the same 1 km pixel, we take a mean value of those soil moisture values and use that instead to represent the soil moisture in that location. By doing this, we end up with only ~~14~~ 10 individual locations, as most of the in situ sites are located near each other. ~~During the individual location acquiring process, one of the two peatland sites (SUO0004) ended up being the only site in one 1 km pixel with extremely high in situ soil moisture values ($> 0.8 \text{ m}^3/\text{m}^3$). Including that location in the training set caused the model to predict erroneous soil moisture values. Therefore, that site was excluded from the study data set.~~

3.2 Model for soil moisture

The data set for model construction consists of only ~~13 individual locations, meaning that there are only 13 different bulk density or silt content values~~ 10 individual locations. We aimed ~~for high variability in soil properties in the training set and, conversely, low variability in the test set to produce a model that is as generalized as possible using the available data~~ to have similar distributions of soil moisture values in both training and test sets. Therefore, we chose ~~9 of those 13~~ 7 of those 10 sites

for the training data set ~~based on their soil properties. The other 4, and the other 3~~ were left for the test set. ~~In the training set, the bulk density values vary from 0.57 g/cm³ to 0.66 g/cm³, and silt content from 27.7 g/kg to 34.6 g/kg. In the test set, the bulk density varies from 0.55 g/cm³ to 0.63 g/cm³, and silt content from 27.8 g/kg to 30.5 g/kg.~~ The placing of the individual
220 in situ sites to training or test set is shown in Figure 1 and Table A1.

We used all the available data from the chosen annual periods covering the years 2019–2023, and hence we had ~~4480~~
~~3415~~ values for training and ~~2221–1775~~ for testing. Tree-based algorithms are commonly used in soil moisture predictions
(e.g. Wei et al. (2019), Trambly and Quintana Seguí (2022), Ning et al. (2023), Shokati et al. (2024)), and it has been reported
that tree-based methods can outperform deep-learning methods (Li and Yan (2024)). The Gradient Boosting (GB) method
225 (Breiman (1997), Friedman (2001), Friedman (2002)), in which the weak learners (decision trees) are trained sequentially
by correcting the residuals of the previous model, was therefore chosen for model construction. We used a framework for
tree-based algorithms called Light Gradient-Boosting Machine (lightGBM), as it is faster to use (Ke et al. (2017)).

We hypertuned the model parameters by using the GridSearchCV method from scikit-learn (Pedregosa et al. (2011)). It is
a method where all possible combinations of given model parameters and their grids are tested and evaluated by using cross-
230 validation. In our model building, we used CV=3. The chosen parameters with their test ranges are shown in Table 1. The
learning rate was chosen to be 0.05, ~~and we set the number of trees to 200. We also used a feature fraction of 0.4 to limit the~~
~~overfitting and prevent one static input (bulk density or silt content) from having too much importance over that of the other~~
~~inputs. We also limited the maximum bins to 200, and a minimum number of data values in one leaf to 40 at maximum to limit~~
~~overfitting.~~

235 4 Results

4.1 Analysis of the model

The SHapley Additive exPlanations (SHAP, Lundberg et al. (2020)) values (which specify the effect of different individual
inputs on the output) indicate that ~~static vegetation~~ inputs dominate the results, as can be seen from Figure 2. ~~Bulk density and~~
~~silt content have a higher influence on the predicted soil moisture compared to other inputs. Also, their effect is nonlinear.~~
240 ~~On the other hand, SMAP soil moisture has a clear linear effect. All inputs have clear linear effects~~ on the results, ~~although~~
~~smaller (compared to static inputs). The rest of the inputs (EVI, NDVI, and ERA5-Land based) have much lower impacts but~~
~~are significant enough to be included as inputs. Precipitation-related inputs have the smallest effect on the model.~~

The RMSE, R, and R² values between the training and test set indicate ~~at least small overfitting when using all test data no~~
~~overfitting~~ (Table 2). ~~Due to the small size of the RMSE and uRMSE values between in situ values and~~ training and test sets
245 ~~, along with low variability in static inputs, preventing overfitting is challenging. These limitations also make the statistical~~
~~values sensitive to changes in the data sets. One site (DIS0004 spots 1 and 2, in Test site D) had a clear deviation from the~~
~~SMAP values in the test data set, and if we remove those values from the test set, overfitting decreases remarkably, which~~
~~can be seen in Figure 3 are almost identical (0.058 and 0.062, and 0.058 and 0.061, respectively). On the other hand, R and R²~~
~~values are higher between in situ values and predicted soil moisture values from the training set compared to values between~~

250 in situ values and test set predicted soil moisture values. Based on results in Figure 3, there is a possibility of the model underestimating higher soil moisture values ($> 0.3 \text{ m}^3/\text{m}^3$). Also, as there are no higher than $0.4 \text{ m}^3/\text{m}^3$ soil moisture values in the training set, the model will have difficulties predicting soil moisture values above $0.4 \text{ m}^3/\text{m}^3$.

As the original highest spatial resolution of some inputs is 250 m (NDVI, ~~EVI, and soil properties~~ and EVI), we also re-sampled SMAP soil moisture and ~~ERA5-Land-based-weather-related~~ inputs to that same 250 m spatial resolution using nearest neighbor resampling. We then calculated soil moisture maps from those 250 m resolution data maps using the constructed GB model to study how sensitive the developed model is to small changes in vegetation ~~and soil property~~-values (i.e. as those are the only parameter values changing within one time step). Exemplary time series for NF test sites for the year ~~2023-2020~~ are shown in Figure 4. The individual in situ sites are located close to each other and therefore the Test sites ~~A-C~~ A-B have the same in situ sites in both resolutions. Only in Test site ~~D-C~~ one site (~~DIS0004-spot2~~MET0010) locates in a different pixel. ~~Due to that, in situ soil moisture values for 250 m resolution in site D would be systematically about $0.04 \text{ m}^3/\text{m}^3$ lower compared to the in situ soil moisture values in 1 km resolution. Because of only a small difference between in situ values in different resolutions, we omitted~~ Therefore, for Figure 4, we have added an extra Test site D, which includes the in situ soil moisture in 250 m resolution from the test site D in Figure 4 site MET0010. Overall, as all sites (A-D) are boreal forest sites, SMAP soil moisture is temporally well in line with in situ soil moisture values. ~~Furthermore, predicted, but due to the coarse resolution, there are systematic differences, especially in Test site C (and D). Predicted~~ values calculated for both 1 km and 250 m resolution data are better in line with in situ values ~~and SMAP soil moisture, even though predicted data in 250 m resolution can have systematic differences (Test site B).~~ Based on these results for NF sites, the developed model is not overly sensitive to small changes in ~~soil-weather-related~~ and vegetation properties data. Also, based on these time series results, the developed model detects temporal changes well. In hindsight, as the model is constructed using SMAP soil moisture, and SMAP soil moisture data is noisy, some of the same noisy features can be found in predicted values. Also, due to the SMAP being the basis for the developed model, the predicted values have the same temporal resolution as SMAP, meaning that data can be predicted almost daily if SMAP soil moisture data are available. Mean relative differences (Table 3) between in situ values and GB model-based predicted values indicate varying under- and overestimations. In 1 km resolution, the underestimation for the whole test set is just $< 1\%$, which is to be expected. For 250 m resolution, the underestimation is higher, almost 20% for the whole test set.

275 We also calculated the soil moisture values for the whole NF area using the constructed model to analyze how well the model captures the spatial variations and also to show the impact of missing pixels on the predicted maps. We calculated soil moisture maps using 1 km and 250 m resolution data. Examples of these predicted soil moisture maps are shown in Figures 5 and 6. Predicted soil moisture values are lower than SMAP soil moisture values, and for 250 m resolution maps the number of missing pixels increases. Nevertheless, spatial changes are well detected by the predicted values when compared to SMAP soil moisture. The missing values in predicted maps are due to the missing data in the inputs. SMAP data have missing data because of water bodies or otherwise failed soil moisture retrievals, ~~whereas soil property data are missing only due to water bodies.~~ Similarly, vegetation properties are not retrieved over water bodies, but vegetation data are also missing because of missing measurements, caused typically by cloud cover (as vegetation properties are based on optical data). Furthermore, as the model is developed mainly for forested areas, a land cover mask was applied to the results (shown only in Figure 6, and

285 omitted in Figure 5 for clarity). We used CORINE land cover data in 100 m spatial resolution as the basis of the mask. Land cover data was resampled to the 1 km and 250 m spatial resolutions and those pixels where forest classes covered under 50% of the coarser pixel were masked.

4.2 Model uncertainty

We used the sensitivity of the most important inputs and the standard deviation of the difference between predicted soil moisture values and in situ values from test data as the uncertainty of the model. First, we approximated the uncertainty each **important** input causes to the results. Predicted soil moisture from the training data was used as the reference data. Then we added errors to the important inputs separately from their error distributions $\epsilon \sim \mathcal{N}(0, \sigma^2)$. For **soil properties, the σ was taken from their uncertainty estimations.** For vegetation indexes, we used the reported uncertainties, 0.015 for EVI and 0.025 for NDVI (<https://modis-land.gsfc.nasa.gov/ValStatus.php?ProductID=MOD13>). For SMAP soil moisture, we used the standard deviation from the difference between SMAP soil moisture and in situ soil moisture from the whole in situ data set from NF. The obtained standard deviation was **0.12. For ERA5-Land based 0.097.** For weather-related inputs, we **decided to omit their uncertainty, as their impact on the model is small.** used reported RMSE values (Aalto et al. (2016)), 1.4 mm for precipitation and 0.58°C for temperature. As we use cumulative sums, we used error propagation of sum to estimate the uncertainty of them. The uncertainties have therefore a form of $\sqrt{x} \cdot RMSE_i$, where x is the number of days the cumulative sum is obtained, and i is either precipitation or temperature. This way we obtained 1.98 mm uncertainty for precipitation sum over 2 days, and 3.7 mm uncertainty for precipitation sum over 7 days, and 1.64°C and 1.83°C uncertainties for temperature sums over 8 and 10 days, respectively. We calculated the difference between the error-added values and the reference data 100 times. The sensitivity of each varied input, the test std, and the total uncertainty for the constructed model are shown in Table 4. The total uncertainty is calculated as a squared sum between the individual sensitivities and test std, that is:

$$305 \quad \epsilon = \sum_i^N (u_i^2). \quad (1)$$

The sensitivities of each input are in line with their importance to the model. Bulk density and silt content have the

SMAP soil moisture has the highest impact on the model uncertainty for individual inputs. On the other hand, vegetation properties and weather-related data have the lowest impact. In total, the model uncertainty is around **0.119** 0.080 m³/m³.

4.3 Validation with Alaska sites

310 **Altogether 16** The weather station network over Alaska is sparse, and thus kriging-based interpolation to obtain precipitation and temperature in high resolution (as done over Northern Finland) is not possible. Therefore, we decided to use satellite-based data for precipitation (GPM data) and for temperature, we used ERA5-Land temperature data. GPM data was calculated to required cumulative sums without any modifications, but as ERA5-Land data is provided hourly, we preprocessed it in daily mean temperatures and then further processed it to required cumulative sums.

315 Altogether 17 stations from Alaska were used as an independent model validation set. One site located in Alaska, Tokositna Valley, was excluded from the validation set, because its soil moisture values varied abnormally. In addition, the predicted soil moisture values below $0.05 \text{ m}^3/\text{m}^3$ were also excluded. We calculated statistical values (RMSE, ~~uRMSE~~RMSE, and R) for each site between in situ soil moisture and SMAP in ~~9-36~~ km resolution, SMAP enhanced to 1 km resolution, and GB-model-based predicted values, both 1 km and 250 m. ~~SMAP in 1 km resolution has a slightly lower mean RMSE compared to others, but predicted values in 250 m have the lowest mean uRMSE value.~~ The median statistical values are similar (Figure 7), only R values are slightly higher with SMAP in 36 km resolution compared to others. ~~GB-model-based predicted values in both 1 km and 250 m resolution have clearly higher R values compared to SMAP soil moisture data sets.~~

Of the 16 stations, only ~~4-5~~ were reported to be located in forested sites (information is based on ESA CCI Land Cover (ESA (2017)) and NLCD (<https://www.mrlc.gov/>)). Soil moisture data comparisons from those ~~four-five~~ sites are shown in Figure 8. 325 SMAP soil moisture in both resolutions has a lot of variability compared to predicted estimates. ~~Predicted values are closer to the in situ values.~~ It is also evident that the GB-model cannot predict high soil moisture values ($> 0.4 \text{ m}^3/\text{m}^3$), as was expected. Overall, there are clear correlations between satellite-based estimates and in situ soil moisture values when taking into account all data, but correlations are less clear when focusing on individual sites.

The Exemplary time series for summer 2019 (May to October) for sites Aniak sites Nenana and Gulkana River (tree-covered 330 sites) are shown in Figure ??. ~~For Aniak (left side), SMAP in 9 km resolution and predicted values in 1 km and 250 m resolution are most of the time well in line with in situ values. SMAP in 1 km has lower 9, and a close-up focusing on the year 2019 in Figure 10. The high~~ soil moisture values compared to the in situ values. For the Gulkana River site (right side), all data sets have difficulties at the beginning of the summer .The high soil moisture values are probably (due to the snow melting increasing the soil moisture, which is melt) are not detected by SMAP .Predicted values catch the high values at the beginning 335 of the summer slightly better than SMAP soil moisture data sets. Predicted values in 1 km have too much temporal variability, which is missing from the 250 m data. On the other hand, the GB-model-based estimates do catch them better. Otherwise, SMAP data in both resolutions detect the soil moisture values well. The GB-model-based soil moisture estimates have more temporal variation compared to SMAP data. The close-up of the year 2019 shows that the model can detect the U-shape of the in situ soil moisture better than SMAP data.

340 Sixteen in situ sites in Alaska were grouped into coarser land use classification classes (forest, mosaic, shrub, and sparse), and RMSE, ~~uRMSE~~RMSE, and R values were calculated between in situ values and each satellite-based data, the values are shown in Table 5. For forested sites, predicted values in 250 m have the lowest RMSE and ~~uRMSE~~RMSE, and highest R values compared to other data sets. Predicted values in 1 km resolution have the second-highest model validation statistics. For mosaic sites, SMAP in ~~9-36~~ km has the lowest RMSE ~~, but the predicted values have the highest uRMSE and correlation~~ 345 ~~values and uRMSE, and the highest R value.~~ All data sets struggle to predict soil moisture values in shrub and sparse sites. In shrub sites, predicted values in 1 km resolution are more in line with in situ values compared to SMAP soil moisture values in both resolutions. Based on these validation results, the developed model predicts temporal changes relatively well.

5 Discussion

Spatio-temporal data on the variation in soil moisture for boreal regions is crucial for predicting forest-related phenomena, such as tree growth and forest fire risk, both of which influence the carbon storage capacity of these ecosystems. However, existing satellite-based soil moisture products for vegetated areas often have coarse spatial resolution. To address this issue, higher-resolution data is necessary to capture the finer spatial variations in soil moisture. Consequently, we developed a model utilizing satellite data to estimate soil moisture at high resolution (1 km and 250 m) over boreal forested regions. We used a tree-based machine learning method called gradient boosting with SMAP soil moisture in 936 km spatial resolution as a basis. Produced data maps have the same temporal resolution as SMAP (typically daily, but are missing if SMAP soil moisture retrieval has failed). The developed model is shown to retain the temporal and spatial variability of SMAP soil moisture, but validated against independent data, the predicted values show better agreement compared to the SMAP soil moisture (RMSE decreasing from $0.097\text{--}0.103\text{ m}^3/\text{m}^3$ to $0.065\text{--}0.092\text{ m}^3/\text{m}^3$, and correlation increasing from $0.30\text{--}0.52$ to $0.46\text{--}0.55$ over forest sites).

Even though SMAP soil moisture has been reported to be sensitive to soil moisture changes under canopy cover (Colliander et al. (2020); Ayres et al. (2021)), the SMAP soil moisture enhancement from 36 km to 9 km spatial resolution has been reported to cause uncertainty to the soil moisture values (Ambadan et al. (2022)). However, based on comparisons (mean error and standard deviation) between enhanced SMAP soil moisture and in situ values from NF sites in There exist numerous other soil moisture products at a 1 km spatial resolution, which differ on the underlying data they use, the methods they implement, and also whether they are global or regional (Table 6). Overall, our constructed model has higher uRMSE values than many other 1 km resolution, the SMAP mission requirement target for soil moisture is achieved, as the mean error ($0.03\text{ m}^3/\text{m}^3$) is below the target ($0.04\text{ m}^3/\text{m}^3$, O'Neill et al. (2021)), even though the standard deviation is high ($0.12\text{ m}^3/\text{m}^3$). Therefore, the enhanced SMAP soil moisture is a suitable basis data for our model. The disadvantage of SMAP soil moisture is that tau-omega-model parameters for each 9 km pixel spatial resolution data sets, but most of them do not cover boreal forest areas or are not validated against boreal forest soil moisture. Of those that do cover boreal forests, the uRMSE and R^2 values are in line with the results we obtained from validation against forested sites in Alaska. Those data products which are based on the dominant land use classification in each pixel. This assumption of homogeneous area ignores the heterogeneous nature of the boreal forest areas (typically a mosaic of forests, peatlands, and water bodies). Sentinel-1 SAR and cover boreal forest zone (Bauer-Marschallinger et al. (2018), Fan et al. (2025), Meyer et al. (2022)) have difficulties with dense vegetation, which is to be expected due to the C-band being sensitive to vegetation. On the other hand, good results are obtained when using ERA5 soil moisture as the basis data (Zhang et al. (2023)). Used downscaling methods and algorithms vary. Change detection method (used in Bauer-Marschallinger et al. (2018)) and forward model (Fan et al. (2025)) are used for Sentinel-1 SAR data, whereas for Sentinel-1 and SMAP combination uses SMAP active-passive algorithm (used in Das et al. (2019) and Meyer et al. (2022)), and is based on work by Das et al. (2014), Entekhabi et al. (2014), and Das et al. (2018)). In the studied area, the SMAP soil moisture is most accurate for forested areas with a herbaceous understory, because, by definition, the dominant land use classification is woody savanna (i.e. a herbaceous understory and forest canopy cover between 30-60%; different class

definitions can be found in Strahler et al. (1999)). Therefore, we also have to restrict our model to forested areas. Based on the land use classification data in 100 m resolution from CORINE land cover for NF (area shown in Figure 1), around 61% of the area is covered by tree cover (42% coniferous trees, 10% broadleaved trees, and 9% mixed trees). One-fifth of the area (almost 20%) is covered in peat bogs, and 4.5% is covered by water bodies (lakes). Water bodies include three large lakes (Lake Inari, Lokka Reservoir, and Porttipahta Reservoir), but also many smaller lakes. The rest of the land use (around 15%) is for different urban activities, grasslands, and agriculture. As our model provides high-resolution soil moisture for forested areas, it covers approximately 60% of the NF area being covered with this model. Additionally, Machine-learning methods are also implemented (Kovačević et al. (2020), Rao et al. (2022), Zhang et al. (2023), Zheng et al. (2023), Zhang et al. (2024)), mostly when using ESA-CCI or ERA5 soil moisture data. When SMAP soil moisture has a lot of noise, and those features are also transferred to our model predictions. Smoothing would have been one option to decrease the effect of noise, but choosing the method that would have retained the actual temporal variations, was not a straightforward task. Therefore, we decided to leave the noise in the end results. is used as a data source, typical algorithms are based either on the thermal inertia theory (used in Fang et al. (2022) and Dandridge et al. (2020)) or the Temperature-Vegetation (T-V) method (used in Yin et al. (2020) and Mohseni et al. (2024), based on Sandholt et al. (2002)). Once again, the validation results from our constructed ML-method-based model are consistent with other ML-based data sets.

~~For modeled soil moisture data, most spatial features are from used soil properties, bulk density, and silt content. Soil properties are commonly used inputs for soil moisture models (e.g. Ranney et al. (2015), O et al. (2022), Ma et al. (2023), Zhang et al. (2023)), although typically also clay content and sand content are used. As we have an exceedingly small set of individual static inputs in the training set (only 9), we had to exclude clay content and sand content, to prevent overfitting a small number of individual sites in training and test sets, we excluded soil properties data from this study.~~ Additionally, other commonly used inputs include topography and geography data (i.e. elevation, slope, aspect, latitude, and longitude). As Again, as we have a very-relatively small amount of model construction data, adding geographical information would have caused major overfitting. We also excluded topography data, as it has been found that models using topography data as inputs may not be useful in other locations (Kemppinen et al. (2023)). Weather-related data, i.e. precipitation, temperature, and solar radiation from ERA5-Land, and temperature, are included as inputs because they are related to the soil moisture. Precipitation is positively correlated with soil moisture (Sehler et al. (2019)), but air temperature has the opposite effect (Feng and Liu (2015)). Based on feature importances (Figure 2), air temperature has in both cumulative sums (sums over 8 and 10 days) have a negative impact on the results as expected, but precipitation does not have has a varying effect. Precipitation sum over 7 days has the expected positive effect. This, but the precipitation sum over 2 days has the opposite effect. The latter might be due to the coarse resolution of ERA5-Land data (9 km) and the dense vegetation in the area causing soil moisture to be less sensitive to precipitation canopy interception and no-rain values. The canopy interception of precipitation can be up to 50% in the dense boreal forest (Molina and del Campo (2012), Zabret et al. (2017), Hassan et al. (2017)), leading to only a small amount of rain attributed to the soil moisture, but also it is possible that there are no rain events happening, leading the sum of rain over 2 days to cause negative effect to the soil moisture values. It has also been studied that air temperature has a higher impact on soil moisture than precipitation, even over forest areas (Feng and Liu (2015)). This effect can be seen in feature importances (Figure

2), as temperature has a clearly higher impact on the soil moisture estimates compared to precipitation. ~~Solar radiation impacts soil moisture values indirectly through evapotranspiration (Liu et al. (2012)).~~ The cumulated precipitation and temperature values ~~(a sum over 72 hours)~~ increased the model accuracy compared to the instantaneous values, and therefore they were
420 chosen. ~~The solar radiation had the opposite effect and hence we chose the instantaneous values.~~ An additional useful data source would have been the land surface temperature (LST), as the LST difference between night and day correlates with soil moisture. LST data has been widely used for estimating soil moisture (e.g. Matsushima et al. (2012), Hao et al. (2022), Han et al. (2023)). The disadvantage of LST is that it is obtained from optical measurements. Due to the difficulties caused by cloud cover, obtaining even moderately gap-free LST data regularly over the whole NF area was an impossible task, and therefore
425 we did not include LST as an input.

To choose the best model, we tested three ~~different~~ tree-based methods: random forest, level-wise gradient boosting, and leaf-wise growth-based gradient boosting. The leaf-wise growth GB (lightGBM) produced the results with the highest accuracy and was therefore chosen. However, because it is a tree-based method, it cannot extrapolate well when inputs differ from training data, as the decision boundaries are determined during training. Therefore, our predicted estimates are more or less bounded,
430 and unexpectedly high or low soil moisture values are not predicted correctly (i.e. soil moisture values below $0.05 \text{ m}^3/\text{m}^3$ or above $0.4 \text{ m}^3/\text{m}^3$). To overcome this disadvantage, there needs to be much more data from diverse in situ locations. The available data from the in situ networks is limited and thus hinders model's predictive ability. ~~During the model construction process, we discovered that model construction is highly sensitive to changes in static inputs due to the small study data set. Therefore, the data division into training and test sets was not random but carefully considered. Almost all in situ sites are~~
435 ~~located in forested areas, a fact that further strengthened the model's usefulness for only forested areas.~~

We included SMAP in 1 km resolution to be compared to our model predictions. ~~There exist other~~ Other downscaled data sets were also considered, but as the soil moisture network is sparse, they, unfortunately, use some of the same in situ sites from NF and Alaska for training, and therefore we couldn't use them as independent data sets. The comparisons with SMAP in 1 km resolution indicate that downscaled SMAP lacks some of the variability found in our model. For some sites in Alaska,
440 SMAP in 1 km even performed weaker compared to SMAP in 936 km resolution. Based on those results, it could be possible that thermal inertia theory is not ideal for downscaling soil moisture data over forested areas.

As our model provides high-resolution soil moisture for forested areas, it covers approximately 60% of the NF area (see Section 2.1). Additionally, SMAP soil moisture has a lot of noise, and some of those features are also transferred to our model predictions. Smoothing would have been one option to decrease the effect of noise, but choosing the method that would have
445 retained the actual temporal variations was not a straightforward task. Also, we tried to implement precipitation and temperature data to smooth some of the noise, but due to the dense nature of the boreal forest, there was no clear relationship between soil moisture changes and weather. Therefore, we decided to leave the noise in the end results.

In the future, L-band-based missions, like the NASA-ISRO SAR mission (NISAR, <https://nisar.jpl.nasa.gov/>, Lal et al. (2023), Lal et al. (2024)) with a planned launch ~~at the end of 2024~~ around April 2025, and Radar Observing System for Europe in L-
450 band (ROSE-L, <https://sentiwiki.copernicus.eu/web/rose-l>) with a planned launch in 2028, are aiming to provide soil moisture data with higher spatial resolution (around ~~10~~ 200 m for NISAR and around 25 m for ROSE-L). With those resolutions, even

peatlands can be taken into account. As they cover over 20% of ~~the NF area is covered by peats and bogs, and considering their importance as NF and are important~~ carbon sinks, peatlands need to be included in soil moisture studies. If there were more in situ observation sites located in varying kinds of peatlands, one could construct a model based on them, and then combine
455 models focused on forested areas and peatlands to better account for all the variability in soil moisture over boreal forest areas. As it is, the constructed GB-model does provide an alternative to downscale SMAP soil moisture in 36 km resolution to finer spatial scales over boreal forests.

6 Conclusions

We developed a model to predict high-resolution soil moisture ~~values~~ in boreal forests. This model specifically targets ~~forested~~
460 ~~regions~~forests, as peatlands are not represented in SMAP soil moisture data, and most in situ soil moisture observation sites are located ~~in forested areas~~within forests. The model ~~is~~was developed by using SMAP soil moisture at ~~9-36~~ km spatial resolution as the basis data, and additional vegetation ~~and soil properties are~~properties and weather-related data were used to guide the machine learning model together with in situ soil moisture values. The model produces predictions at a resolution of 1 km, which aligns well with SMAP measurements. However, it can also generate soil moisture estimates at a finer resolution of
465 250 m, offering improved accuracy in certain applications, for example hydrological modelling and carbon exchange studies. Consequently, ~~this~~the model provides a valuable tool for predicting soil moisture in high resolution across boreal forested landscapes.

Data availability. The data sets associated with this paper are available in the Finnish Meteorological Institute Research Data repository METIS (<http://hdl.handle.net/11304/a0c2b377-2e2f-4054-8d82-37abc28beec9>).

Table 1. Gradient Boosting model parameters, their ranges and chosen values to be used for model building. Parameter ranges are constrained to prevent overfitting. The chosen values are determined by using GridSeachCV method with CV=3.

Parameter name	Range	Chosen value
Number of leaves	[3 ,4,5,6]	5
Maximum depth	[2 , 3 ,4,5, 6]	4
Minimal amount of data in one leaf	[10 ,20,30,40]	20 <u>40</u>
Maximum number of bins	[30 , 40 , 50 , 60 , 70 <u>100,150,200</u>]	60 <u>200</u>
<u>Number of trees</u>	[<u>200,300,400</u>]	<u>300</u>

Table 2. Statistical values between predicted values and in situ soil moisture for training and test sets.

	RMSE	URMSE <u>uRMSE</u>	R	R ²	N
Training	0.064 <u>0.058</u>	0.064 <u>0.058</u>	0.86 <u>0.81</u>	0.74 <u>0.66</u>	4480 <u>3415</u>
Test all	0.078 <u>0.062</u>	0.063 <u>0.061</u>	0.40 <u>0.71</u>	0.16 <u>0.51</u>	2221 <u>1725</u>
Test (DIS0004 excluded) 0.049 0.045 0.66 0.44 2221 heightTest site A	0.046 <u>0.069</u>	0.043 <u>0.059</u>	0.64 <u>0.43</u>	0.41 <u>0.19</u>	534 <u>585</u>
Test site B	0.070 <u>0.044</u>	0.049 <u>0.039</u>	0.72 <u>0.71</u>	0.52 <u>0.50</u>	579 <u>524</u>
Test site C	0.040 <u>0.067</u>	0.034 <u>0.061</u>	0.71 <u>0.51</u>	0.50 <u>0.26</u>	509 <u>616</u>
Test site height					

Table 3. Mean relative differences [%] between in situ soil moisture values and predicted soil moisture estimates from both GB model. Values are from test set, and they cover the time period 2019–2023.

Resolution	Test site A	Test site B	Test site C	Test site D	All
1 km	0.023–17.37	0.77–7.11	0.59–24.25	599–	–0.88
Test site D (DIS0004 excluded) 250 m	0.034–0.12	0.025–19.14	0.72–58.36	0.52–0.83	599–19.55

Table 4. Sensitivities for chosen inputs, standard deviation between test set in situ soil moisture and predicted soil moisture, and calculated total uncertainty of the model. All results have the unit m^3/m^3 .

source <u>Source</u>	uncertainty <u>Uncertainty</u>
Bulk density 0.057 Silt content 0.075 SMAP soil moisture	0.035 <u>0.029</u>
NDVI	0.003 <u>0.020</u>
EVI	<u>0.019</u>
<u>Prec. 2d sum</u>	<u>0.003</u>
<u>Prec. 7d sum</u>	0.004
<u>T2m 8d sum</u>	<u>0.002</u>
<u>T2m 10d sum</u>	<u>0.003</u>
Test standard deviation	0.065 <u>0.069</u>
total	0.119 <u>0.080</u>

Table 5. Model validation statistics between observed in situ soil moisture from Alaska sites and predicted soil moisture values.

RMSE	forest	mosaic	shrub	sparse	all
SMAP	0.097 <u>0.103</u>	0.122 0.149	0.193 <u>0.174</u>	0.112 <u>0.156</u>	0.137 <u>0.144</u>
SMAP 1 km	0.112 <u>0.095</u>	0.136 <u>0.162</u>	0.170 <u>0.197</u>	0.123 <u>0.152</u>	0.133 <u>0.150</u>
Predicted 1 km	0.067 <u>0.093</u>	0.152 <u>0.175</u>	0.182 <u>0.147</u>	0.110 0.131	0.137 0.138
Predicted 250 m	0.065 0.092	0.138 <u>0.174</u>	0.185 0.126	0.110 <u>0.136</u>	0.134 <u>0.145</u>
URMSE <u>uRMSE</u>					
SMAP	0.096 <u>0.101</u>	0.121 0.147	0.134 <u>0.138</u>	0.104 <u>0.142</u>	0.133 <u>0.144</u>
SMAP 1 km	0.097 <u>0.094</u>	0.131 <u>0.156</u>	0.164 <u>0.161</u>	0.108 <u>0.134</u>	0.128 <u>0.149</u>
Predicted 1 km	0.067 <u>0.093</u>	0.119 <u>0.163</u>	0.114 0.091	0.106 0.128	0.118 0.131
Predicted 250 m	0.065 0.092	0.117 <u>0.163</u>	0.106 <u>0.104</u>	0.108 <u>0.133</u>	0.113 <u>0.137</u>
R					
SMAP	0.30 <u>0.46</u>	0.37 0.61	-0.17 <u>0.41</u>	0.06 <u>0.15</u>	0.18
SMAP 1 km	0.02 <u>0.43</u>	0.13 <u>0.40</u>	-0.65 <u>0.43</u>	0.00 <u>0.03</u>	0.12 <u>0.09</u>
Predicted 1 km	0.47 <u>0.50</u>	0.54 <u>0.21</u>	-0.14 0.42	-0.01 0.14	0.22 0.25
Predicted 250 m	0.52 0.55	0.58 <u>0.14</u>	-0.01	-0.02 <u>0.03</u>	0.20 <u>0.16</u>
N					
SMAP	418 <u>2680</u>	402 <u>2556</u>	439 <u>1550</u>	367 <u>2530</u>	4626 <u>9316</u>
SMAP 1 km	434 <u>2680</u>	407 <u>2556</u>	240 <u>1550</u>	374 <u>2530</u>	4455 <u>9316</u>
Predicted 1 km	497 <u>2680</u>	484 <u>2556</u>	548 <u>1550</u>	421 <u>2530</u>	4950 <u>9316</u>
Predicted 250 m	405 <u>2788</u>	288 <u>2497</u>	409 <u>1554</u>	308 <u>2460</u>	4432 <u>9299</u>

Table 6. A collection of downscaled soil moisture data sets in 1 km spatial resolution, with reported accuracies. The uRMSE values have the unit m^3/m^3 .

Area	Source data	uRMSE	R	R ²	Boreal forest	Reference
global	S1-SAR	0.080 ^a	0.39	—	RMSE = 0.035 m^3/m^3 ; R = 0.28	Bauer-Marschallinger et al. (2018)
global	S1-SAR	0.077	—	—	uRMSE = 0.097 m^3/m^3	Fan et al. (2025)
global	S1-SAR and SMAP	0.050	—	—	does not cover	Das et al. (2019)
global	S1-SAR and SMAP	0.056 ^a	—	0.70	RMSE $\in$ [0.043, 0.162]; R $\in$ [0.33, 0.61]	Meyer et al. (2022)
global	ESA-CCI SM	0.058	—	0.73	does not cover	Kovačević et al. (2020)
global	ESA-CCI SM	0.052	0.82	—	not validated	Zhang et al. (2024)
global	ESA-CCI SM	0.045	0.89	—	not validated	Zheng et al. (2023)
global	ERA5 SM	0.048	0.89	—	uRMSE = 0.061 m^3/m^3 ; R = 0.73	Zhang et al. (2023)
global	SMAP	0.063	—	—	not optimal over high latitudes	Fang et al. (2022)
global	SMAP	0.054	0.64	—	includes dense vegetation	Yin et al. (2020)
regional ^b	SMAP	—	—	0.70	not for boreal forest	Dandridge et al. (2020)
regional ^c	SMAP	0.069	0.64	—	not for boreal forest	Mohseni et al. (2024)
regional ^d	SMAP	0.057	—	—	not for boreal forest	Rao et al. (2022)

^aRMSE instead of uRMSE

^bLower Mekong River Basin

^cAfrica

^dChina

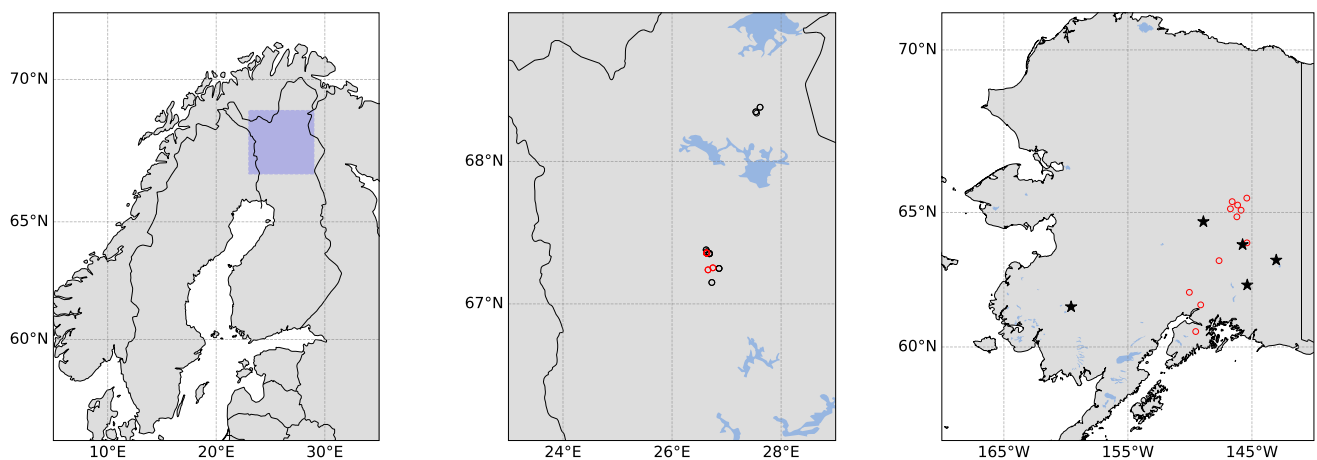


Figure 1. Locations of the chosen training, test, and validation data sites. Left: Northern Finland study area in a broader context (blue squared area). Middle: Location of the chosen model training (black circles) and test (red circles) in situ sites. Right: Location of the chosen model validation in situ sites. Black stars indicate forest sites and red circles indicate other sites (mosaic, shrub, sparse).

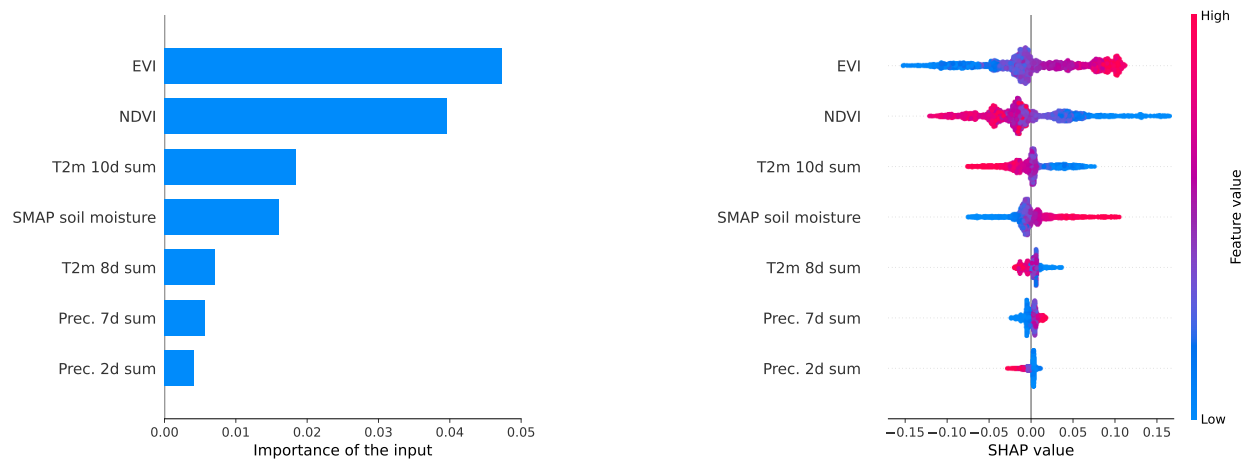


Figure 2. The SHapley Additive exPlanations (SHAP) values for the constructed gradient boosting model. Left: the mean SHAP values for each predictor. Right: More detailed view of the effect of different feature values on predictions.

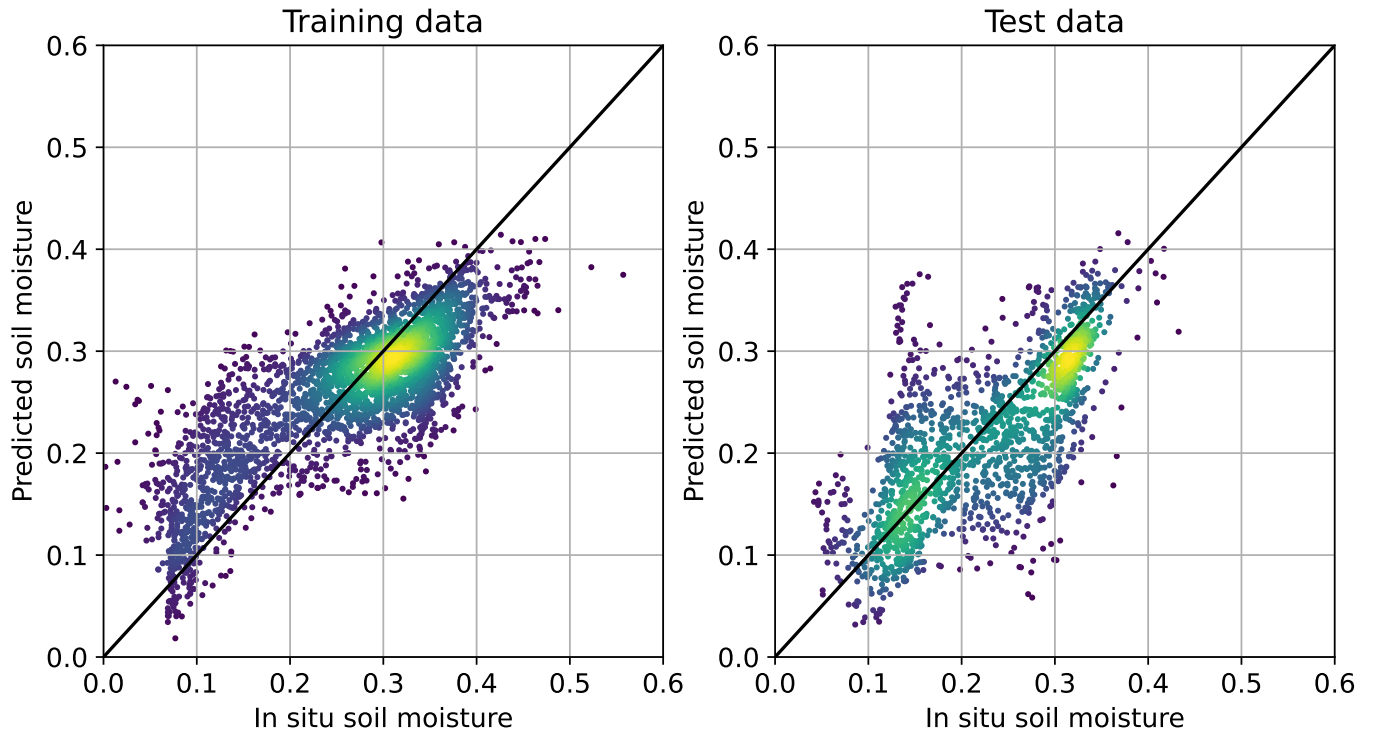


Figure 3. Scatter plots of predicted training and test set soil moisture values from years 2019–2023. Left: scatter plot of training data set. ~~Middle: scatter plot of test set.~~ Right: scatter plot of test set, ~~where one site (DIS0004) is removed from the set as site's soil moisture values deviated clearly from the SMAP soil moisture values.~~

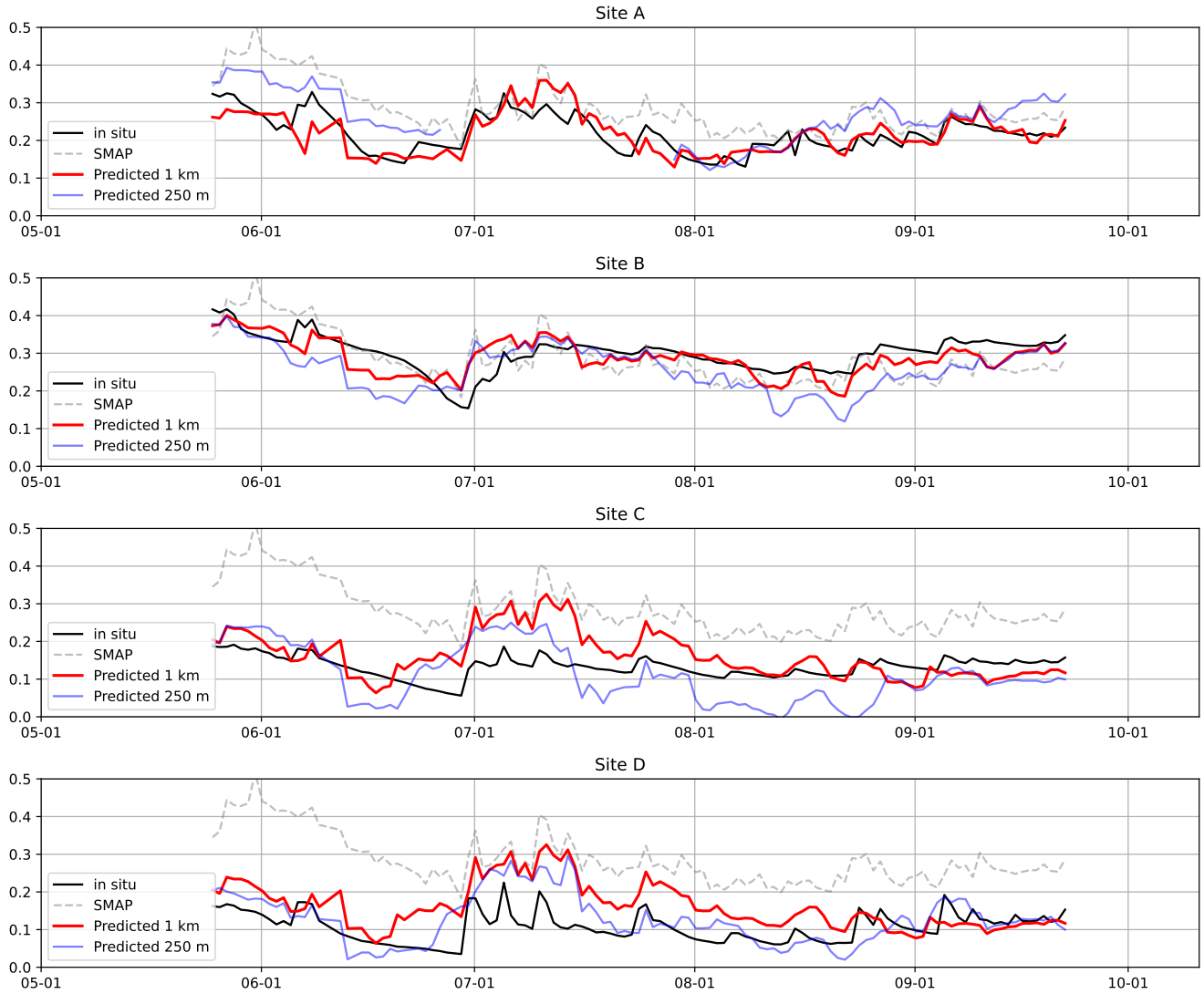


Figure 4. Exemplary time series of test sites for the year ~~2023~~.2020. Predicted soil moisture values in 1 km and 250 m resolutions are from a developed gradient boosting model. In test site C, one in situ site (MET0010) locates in different pixel in 250 m resolution. Therefore, we added an extra Test site D, which includes the in situ site MET0010.

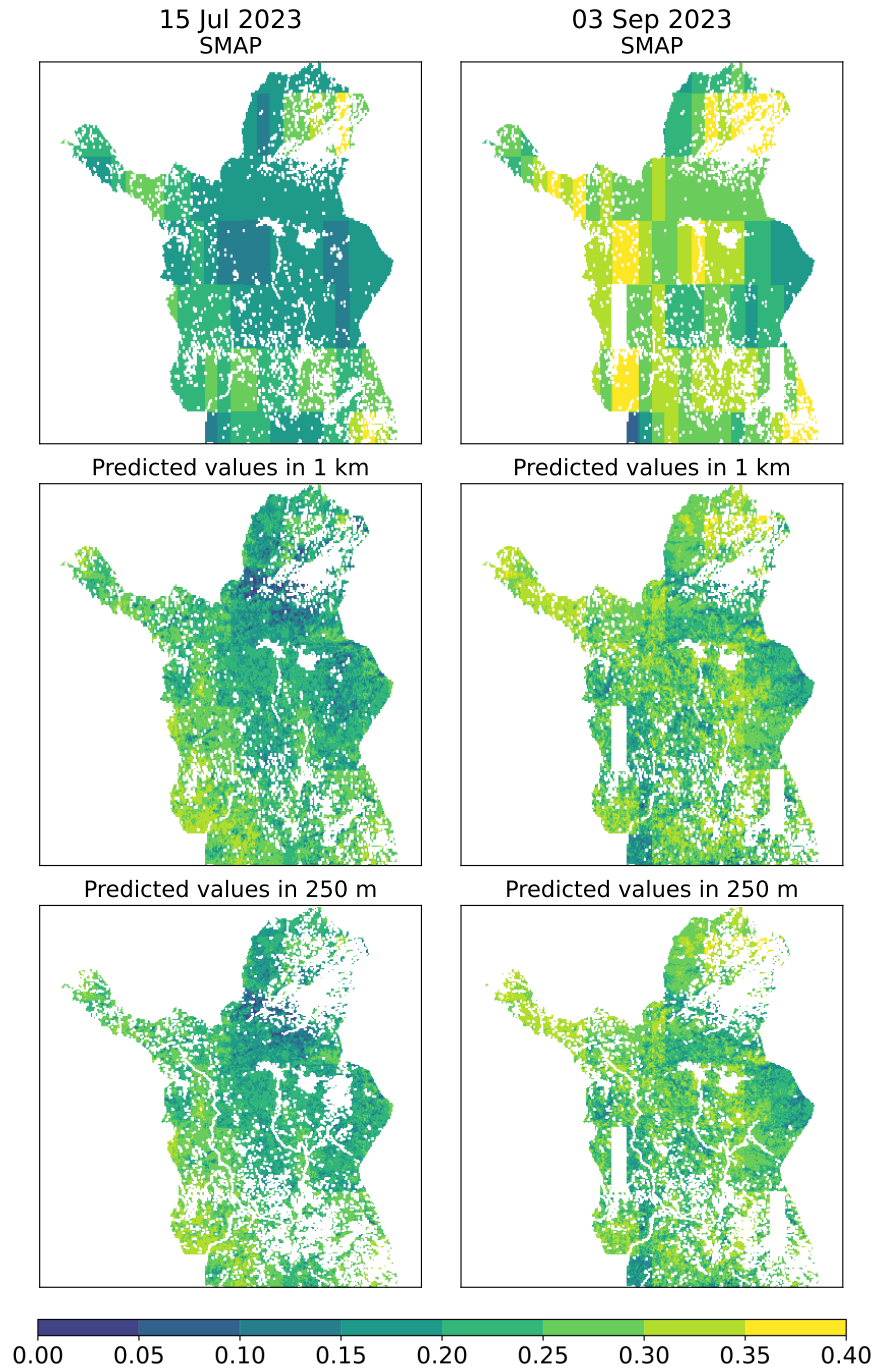


Figure 5. Exemplary maps show SMAP soil moisture for two dates, along with predicted soil moisture at spatial resolutions of 1 km and 250 m. Missing values due to the missing values in inputs and water bodies are indicated with ~~in~~ white. Even though developed model is just for forested areas, all pixels with data in these maps are shown for clarity.

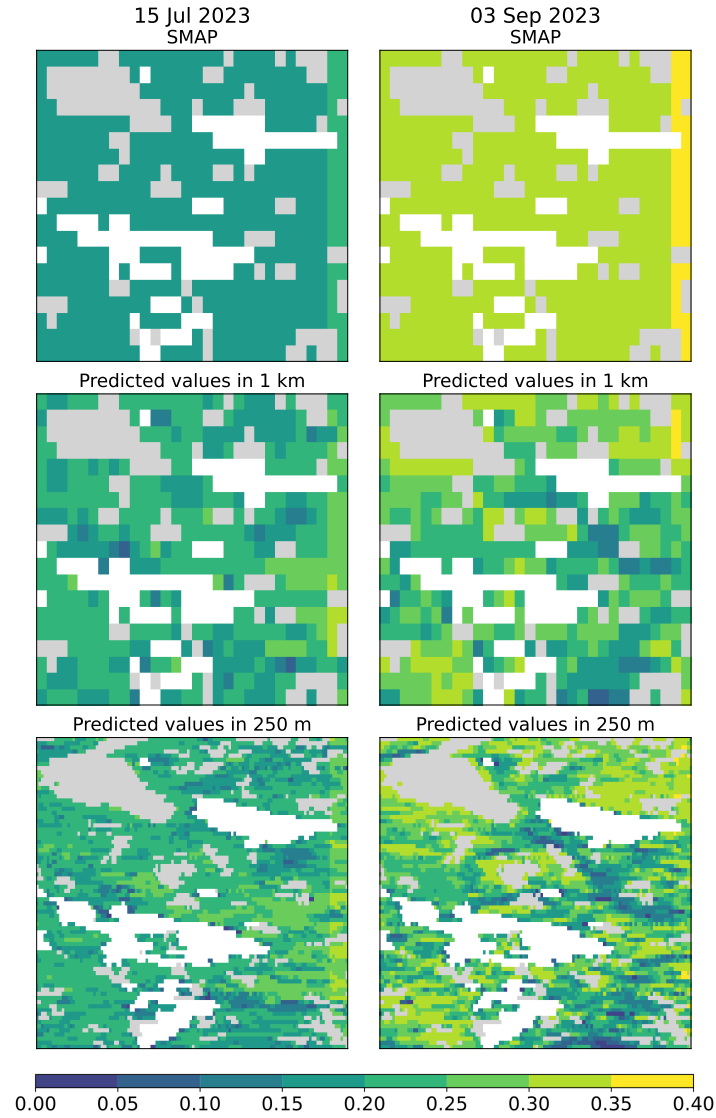


Figure 6. Exemplary maps show SMAP soil moisture for two dates, along with predicted soil moisture at spatial resolutions of 1 km and 250 for a smaller area located around Lake Pallas (N68.033°, E24.197°). Missing values due to the water bodies are indicated with-in white and other land uses than forest are indicated with-in grey. The land use mask is based on CORINE land use classification in 100 m resolution. Pixel is assumed to be forest if the forest class fraction is above 50 %.

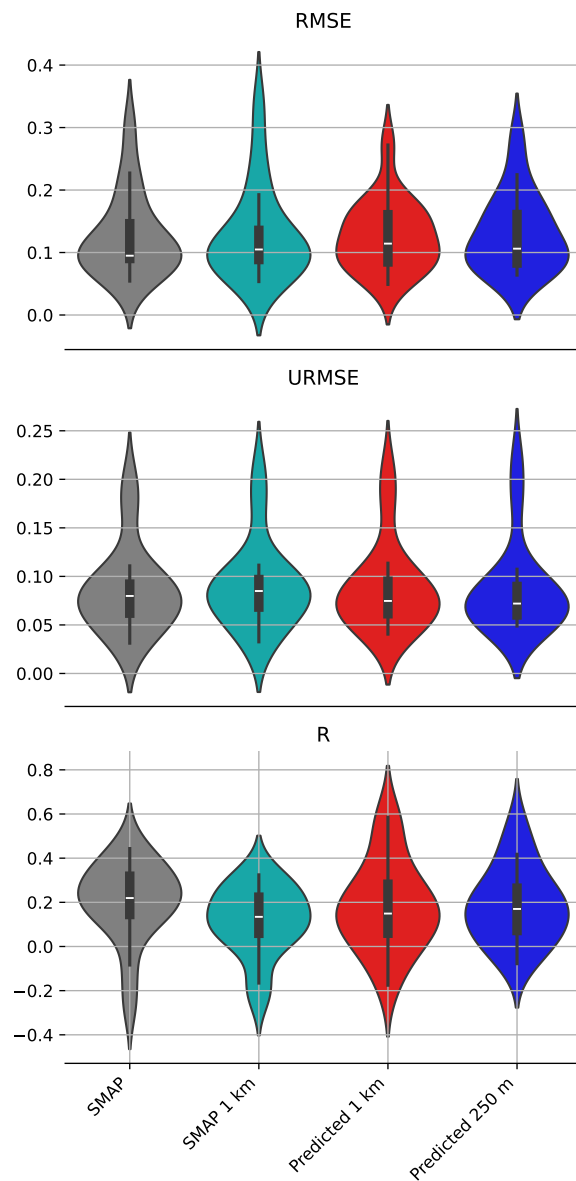


Figure 7. RMSE, URMSE, and R values between in situ values from Alaska validation data set and different soil moisture data sets (SMAP in 936 km resolution, SMAP in 1 km resolution, predicted values using 1 km resolution data, and 250 m resolution data) shown as violin plots. The data is from the annual time period between 1 May 2019 and 15 Oct 2019, covering varying number of years depending on the in situ site.

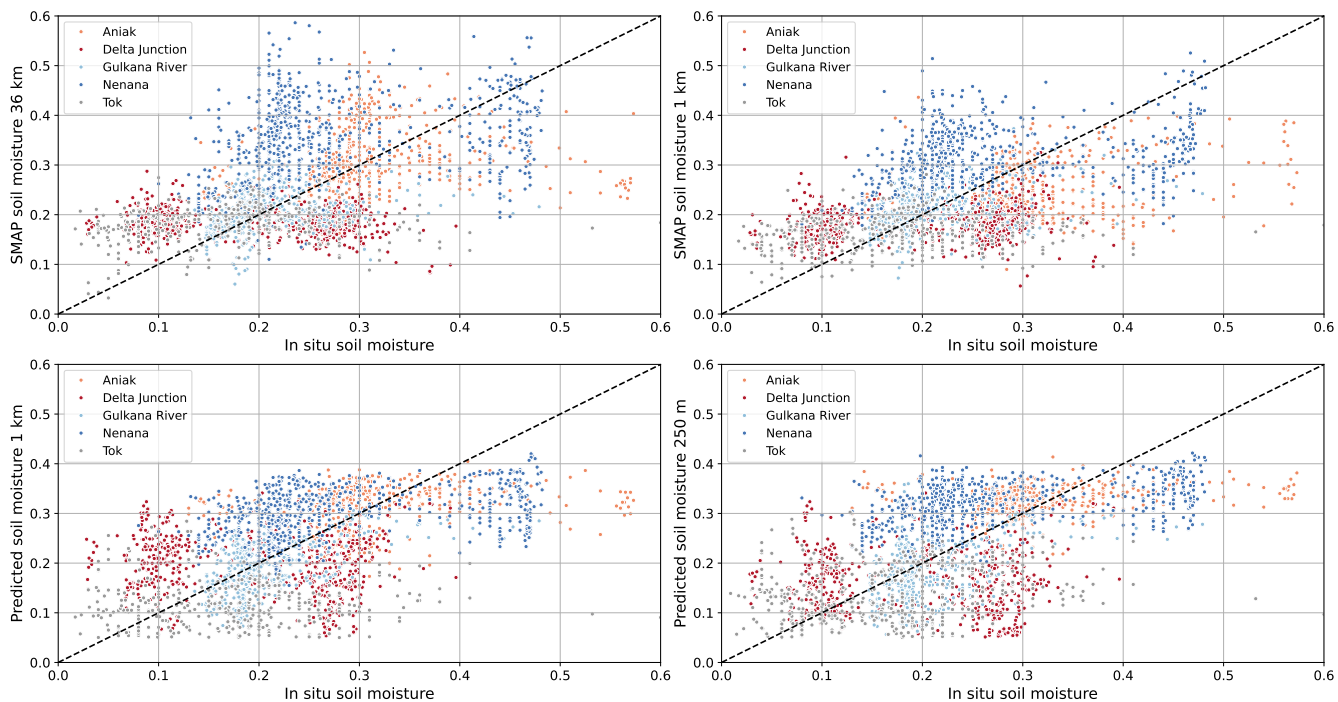


Figure 8. Comparisons between in situ values from Alaska sites (~~four~~ five forested sites) and different soil moisture data sets (SMAP in ~~936~~ 936 km resolution, SMAP in 1 km resolution, predicted values using 1 km resolution data, and 250 m resolution data). The data is from the annual time period between ~~May-1, 2019, May and October-15 Oct, 2019~~ May-1, 2019, May and October-15 Oct, 2019 covering varying number of years depending on the in situ site.

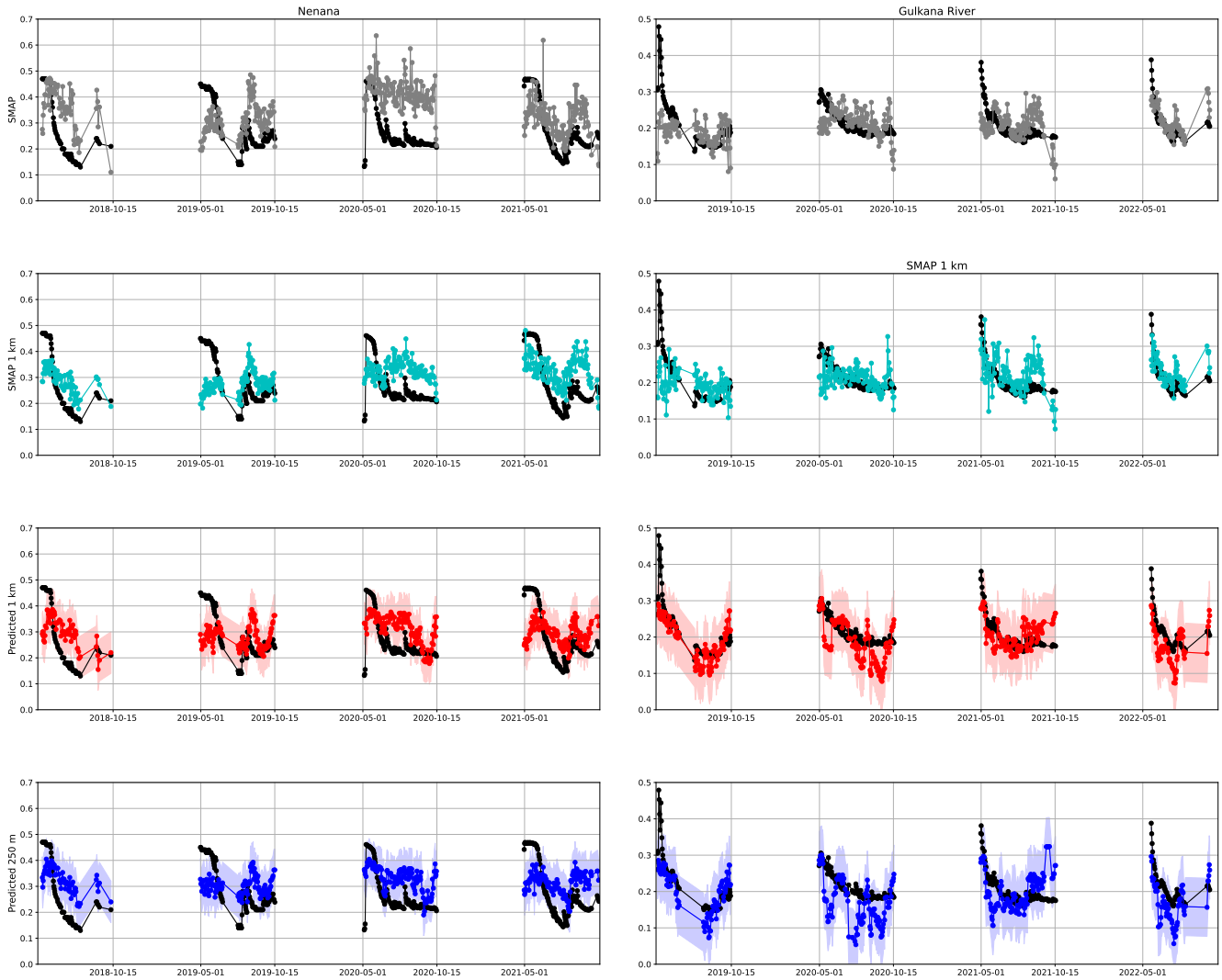


Figure 9. Exemplary time series of soil moisture for two Alaska sites located in forested areas. Left: ~~Aniak~~Nenana, right: Gulkana River. Black indicates in situ soil moisture and grey, turquoise, red, and blue satellite-based soil moisture data. The curtain in the two bottom rows indicates the model uncertainty (uncertainty 0.119-0.080 m³/m³ added and ~~subtracted~~subtracted from the predicted values). ~~All data are~~ Data for Nenana is from the year 2019, years 2018–2021, and for Gulkana River from years 2019–2022.

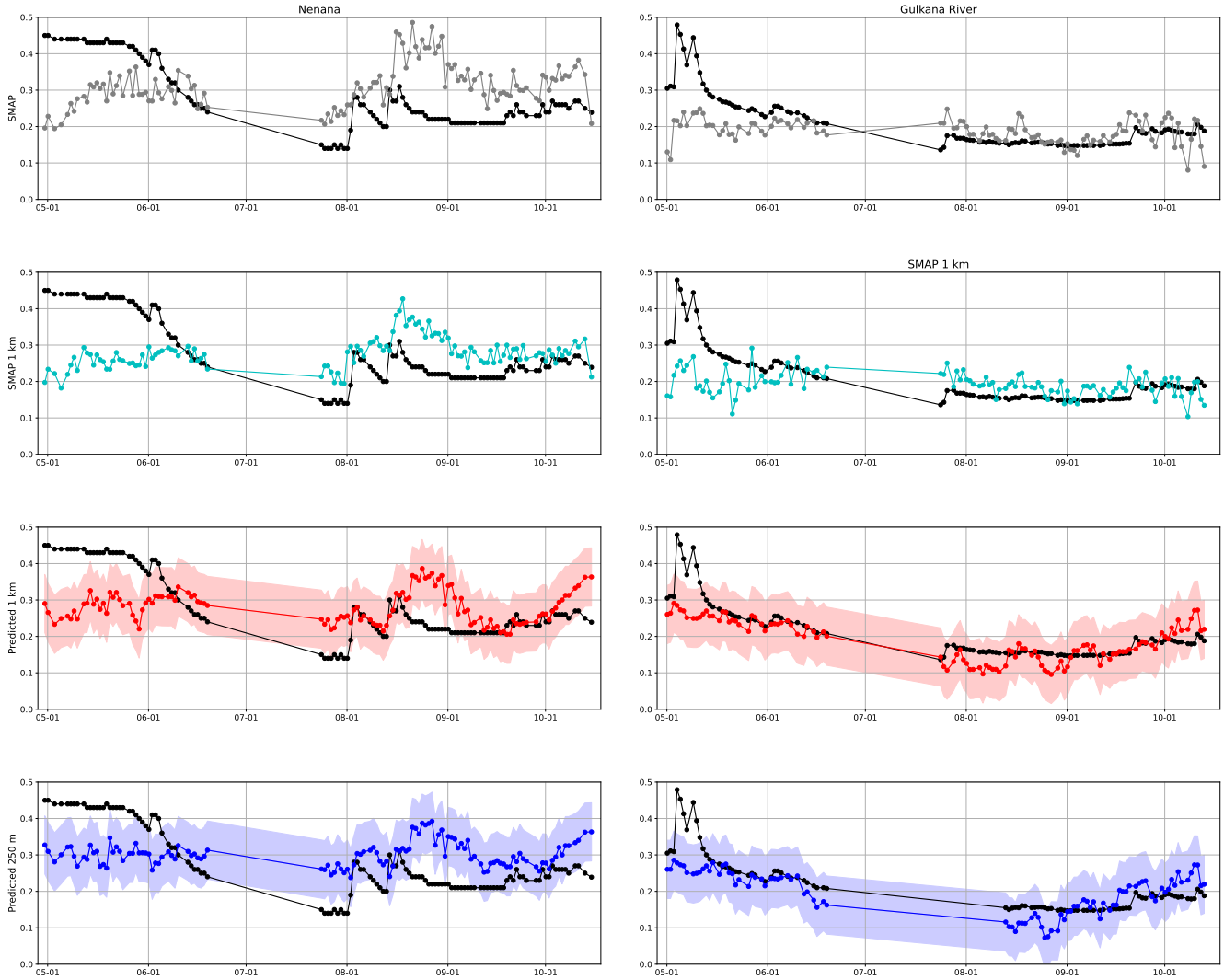


Figure 10. Close-up from the exemplary time series for two Alaska sites located in forested areas. Left: Nenana right: Gulkana River. Black indicates in situ soil moisture and grey, turquoise, red, and blue satellite-based soil moisture data. The curtain in the two bottom rows indicates the model uncertainty (uncertainty $0.080 \text{ m}^3/\text{m}^3$ added and subtracted from the predicted values). Data are from the year 2019.

Table A1. In situ sites for training and testing the soil moisture model, located in Northern Finland. Land cover information is from CORINE land cover

Name	Number of spots
Training set	
IOA0002 <u>DIS0002</u>	1 <u>3</u>
IOA0003 <u>DIS0004</u>	7 <u>2</u>
IOA0007 <u>DIS0005</u>	2 <u>1</u>
KAI0001	3
KAI0002 ^a	2
LEN0001	2
MET0003 1 N67.362 , E26.637 Coniferous forest 0.584 27.70 MET0010 1 N67.362 , E26.638 Coniferous forest 0.584 27.70 SAA0001	3
SAA0002	2
SAA0003	2
SUO0004 2 N67.367 , E26.651 Peats and bogs 0.616 28.31 SUO0005 2 N67.370 , E26.640 Coniferous forest 0.585 29.32 height	Test set
DIS0001 ¹	3
DIS0002 <u>IOA0002</u> ²	3 <u>1</u>
DIS0003 <u>IOA0003</u> ^{b,2}	2 <u>8</u>
DIS0004 <u>IOA0007</u> ²	2
DIS0005 <u>MET0010</u> ^{2,4}	1
<u>DIS0003</u> ³	0.55 <u>1</u> 8

^a ~~OA0003 spot 4 locates in a different pixel (bulk density 0.585 cg/cm³, silt content 29.32g/kg)~~KAI0002 has 3 spots, but one of them had abnormally low

^b ~~KAI0002 has 3~~IOA0003 has 8 spots, but ~~one two~~ of them had abnormally low soil moisture values and ~~was were~~ therefore removed.).

^{e1} Test site A.

^{e2} Test site B.

^{e3} Test site C.

^{e4} Test site D in 250 m resolution.

Table A2. In situ sites for model validation, sites located in Alaska. Data are from ~~three~~four different networks, SCAN (Schaefer et al. (2007)), SNOTEL (Leavesley et al. (2008), Leavesley (2010)), USCRN (Bell et al. (2013)), and NEON (National Ecological Observatory Network (2025)). Land cover information for SCAN, SNOTEL, and USRN sites is from ESA CCI Land Cover (ESA (2017)), and for NEON, the land cover information is from NLCD (<https://www.mrlc.gov/>).

Name	Location
Aniak	N61.58, W159.58
<u>Delta Junction</u>	<u>N63.88, W145.75</u>
Eagle Summit	N65.49, W145.41
Granite Creek	N63.94, W145.40
Gulkana River	N62.41, W145.38
Kenai-Moose Pens N60.73, W150.48 SCAN Shrub or herbaceous cover flooded fresh/saline/brakish water Little Chena Ridge	N65.12, W146.73
Monahan Flat	N63.31, W147.65
Monument Creek	N65.08, W145.87
Mt. Ryan	N65.25, W146.15
Munson Ridge	N64.85, W146.21
Nenana	N64.68, W148.92
<u>Spring Creek</u>	<u>N61.66, W149.13</u>
Summit Creek	N60.62, W149.53
Susitna Valley High	N62.13, W150.04
Tokositna Valley-Tok	N62.63, W150.78 <u>N6</u>
Upper None Creek	N65.37, W146.59

Author contributions. TV and MA are responsible for acquiring the funding for the research presented in this paper. TV and EJ are behind the conceptualization of research questions presented in this paper. EJ collected and processed all the data used in this paper. EJ developed methodology for analyses, carried them out and investigated the obtained results. EJ also led the preparation of manuscript, with contributions
475 from all co-authors. All authors reviewed and edited the manuscript.

Competing interests. The contact author has declared that none of the authors has any competing interests.

Acknowledgements. This work was financially supported by the Research Council of Finland in the project CWI (347662). We acknowledge support from the Ministry of Transport and Communications through the Integrated Carbon Observing System (ICOS) Finland. The authors would like to thank the National Snow and Ice Data Center, Copernicus Data Space Ecosystem, and NASA for providing the satellite data
480 used in this study, ~~ISRIC for providing the soil properties,~~ and the Finnish Meteorological Institute ~~and~~, the International Soil Moisture Network, and the U.S National Science Foundation's National Ecological Observatory Network for providing the in situ measurements used in this study. Furthermore, this publication has been prepared using European Union's Copernicus Land Monitoring Service information (<https://doi.org/10.2909/960998c1-1870-4e82-8051-6485205ebbac>). The authors would like to thank Aku Riihelä, Kimmo Rautiainen, Niilo Kalakoski, and Jaakko Ikonen (Finnish Meteorological Institute) for helpful discussions.

- Aalto, J., Pirinen, P., and Jylhä, K.: New gridded daily climatology of Finland: Permutation-based uncertainty estimates and temporal trends in climate, *Journal of Geophysical Research: Atmospheres*, 121, 3807–3823, <https://doi.org/https://doi.org/10.1002/2015JD024651>, 2016.
- Ambadan, J. T., MacRae, H. C., Colliander, A., Tetlock, E., Helgason, W., Gedalof, Z., and Berg, A. A.: Evaluation of SMAP Soil Moisture Retrieval Accuracy Over a Boreal Forest Region, *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–11, <https://doi.org/10.1109/TGRS.2022.3212934>, 2022.
- Ayres, E., Colliander, A., Cosh, M. H., Roberti, J. A., Simkin, S., and Genazzio, M. A.: Validation of SMAP Soil Moisture at Terrestrial National Ecological Observatory Network (NEON) Sites Show Potential for Soil Moisture Retrieval in Forested Areas, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14, 10 903–10 918, <https://doi.org/10.1109/JSTARS.2021.3121206>, 2021.
- 495 Balenzano, A., Mattia, F., Satalino, G., Lovergine, F. P., Palmisano, D., Peng, J., Marzahn, P., Wegmüller, U., Cartus, O., Dąbrowska-Zielińska, K., Musial, J. P., Davidson, M. W., Pauwels, V. R., Cosh, M. H., McNairn, H., Johnson, J. T., Walker, J. P., Yueh, S. H., Entekhabi, D., Kerr, Y. H., and Jackson, T. J.: Sentinel-1 soil moisture at 1 km resolution: a validation study, *Remote Sensing of Environment*, 263, 112 554, <https://doi.org/https://doi.org/10.1016/j.rse.2021.112554>, 2021.
- Bauer-Marschallinger, B., Schaufler, S., and Navacchi, C.: Copernicus Global Land Operations, Vegetation and Energy, CGLOPS-1, Validation report, available at: <https://land.copernicus.eu/en/technical-library/validation-report-surface-soil-moisture-version-1/@@download/file>, 2018.
- 500 Bauer-Marschallinger, B., Freeman, V., Cao, S., Paulik, C., Schaufler, S., Stachl, T., Modanesi, S., Massari, C., Ciabatta, L., Brocca, L., and Wagner, W.: Toward Global Soil Moisture Monitoring With Sentinel-1: Harnessing Assets and Overcoming Obstacles, *IEEE Transactions on Geoscience and Remote Sensing*, 57, 520–539, <https://doi.org/10.1109/TGRS.2018.2858004>, 2019.
- 505 Bell, J., Palecki, M., Baker, B., Collins, W., Lawrimore, J., Leeper, R., Hall, M., Kochendorfer, J., Meyers, T., Wilson, T., and Diamond, H.: U.S. Climate Reference Network Soil Moisture and Temperature Observations, *Journal of Hydrometeorology*, 14, 977–988, <https://doi.org/10.1175/JHM-D-12-0146.1>, 2013.
- Breiman, L.: Arcing the edge, Tech. rep., Statistics Department University of California, Berkeley CA. 94720, 1997.
- Brodzik, M. J., Billingsley, B., Haran, T., Raup, B., and Savoie, M. H.: EASE-Grid 2.0: Incremental but significant improvements for Earth-gridded data sets, *ISPRS International Journal of Geo-Information*, 1, 32–45, 2012.
- 510 Büttner, G., Kostztra, B., Soukup, T., Sousa, A., and Langanke, T.: CLC2018 Technical Guidelines, available at: <https://land.copernicus.eu/en/technical-library/clc-2018-technical-guidelines/>, 2017.
- Chan, S. and Dunbar, R. S.: Soil Moisture Active Passive (SMAP) Mission: Enhanced Level 3 Passive Soil Moisture Product Specification Document, https://nsidc.org/sites/default/files/d5629220smap2013_sm_p_e20psd_version2050_final.pdf, 2021.
- 515 Chan, S., Bindlish, R., O’Neill, P., Jackson, T., Njoku, E., Dunbar, S., Chaubell, J., Piepmeier, J., Yueh, S., Entekhabi, D., Colliander, A., Chen, F., Cosh, M., Caldwell, T., Walker, J., Berg, A., McNairn, H., Thibeault, M., Martínez-Fernández, J., Uldall, F., Seyfried, M., Bosch, D., Starks, P., Holifield Collins, C., Prueger, J., van der Velde, R., Asanuma, J., Palecki, M., Small, E., Zreda, M., Calvet, J., Crow, W., and Kerr, Y.: Development and assessment of the SMAP enhanced passive soil moisture product, *Remote Sensing of Environment*, 204, 931–941, <https://doi.org/https://doi.org/10.1016/j.rse.2017.08.025>, 2018.

- 520 Clemmensen, K. E., Bahr, A., Ovaskainen, O., Dahlberg, A., Ekblad, A., Wallander, H., Stenlid, J., Finlay, R. D., Wardle, D. A., and Lindahl, B. D.: Roots and Associated Fungi Drive Long-Term Carbon Sequestration in Boreal Forest, *Science*, 339, 1615–1618, <https://doi.org/10.1126/science.1231923>, 2013.
- Colliander, A., Cosh, M. H., Kelly, V. R., Kraatz, S., Bourgeau-Chavez, L., Siqueira, P., Roy, A., Konings, A. G., Holtzman, N., Misra, S., Entekhabi, D., O'Neill, P., and Yueh, S. H.: SMAP Detects Soil Moisture Under Temperate Forest Canopies, *Geophysical Research Letters*, 47, e2020GL089697, <https://doi.org/https://doi.org/10.1029/2020GL089697>, 2020.
- 525 Dandridge, C., Fang, B., and Lakshmi, V.: Downscaling of SMAP Soil Moisture in the Lower Mekong River Basin, *Water*, 12, <https://doi.org/10.3390/w12010056>, 2020.
- Das, N. N., Entekhabi, D., Njoku, E. G., Shi, J. J. C., Johnson, J. T., and Colliander, A.: Tests of the SMAP Combined Radar and Radiometer Algorithm Using Airborne Field Campaign Observations and Simulated Data, *IEEE Transactions on Geoscience and Remote Sensing*, 52, 2018–2028, <https://doi.org/10.1109/TGRS.2013.2257605>, 2014.
- 530 Das, N. N., Entekhabi, D., Dunbar, R. S., Colliander, A., Chen, F., Crow, W., Jackson, T. J., Berg, A., Bosch, D. D., Caldwell, T., Cosh, M. H., Collins, C. H., Lopez-Baeza, E., Moghaddam, M., Rowlandson, T., Starks, P. J., Thibeault, M., Walker, J. P., Wu, X., O'Neill, P. E., Yueh, S., and Njoku, E. G.: The SMAP mission combined active-passive soil moisture product at 9km and 3km spatial resolutions, *Remote Sensing of Environment*, 211, 204–217, <https://doi.org/https://doi.org/10.1016/j.rse.2018.04.011>, 2018.
- 535 Das, N. N., Entekhabi, D., Dunbar, R. S., Chaubell, M. J., Colliander, A., Yueh, S., Jagdhuber, T., Chen, F., Crow, W., O'Neill, P. E., Walker, J. P., Berg, A., Bosch, D. D., Caldwell, T., Cosh, M. H., Collins, C. H., Lopez-Baeza, E., and Thibeault, M.: The SMAP and Copernicus Sentinel 1A/B microwave active-passive high resolution surface soil moisture product, *Remote Sensing of Environment*, 233, 111380, <https://doi.org/https://doi.org/10.1016/j.rse.2019.111380>, 2019.
- Didan, K.: MODIS/Aqua Vegetation Indices 16-Day L3 Global 250m SIN Grid V061, <https://doi.org/10.5067/MODIS/MYD13Q1.061>, accessed: 2024-10-02, 2021a.
- 540 Didan, K.: MODIS/Terra Vegetation Indices 16-Day L3 Global 250m SIN Grid V061, <https://doi.org/10.5067/MODIS/MOD13Q1.061>, accessed: 2024-10-02, 2021b.
- Dorigo, W., Xaver, A., Vreugdenhil, M., Gruber, A., Dostálová, A., Sanchis-Dufau, A. D., Zamojski, D., Cordes, C., Wagner, W., and Drusch, M.: Global Automated Quality Control of In Situ Soil Moisture Data from the International Soil Moisture Network, *Vadose Zone Journal*, 12, vzj2012.0097, <https://doi.org/10.2136/vzj2012.0097>, 2013.
- 545 Dorigo, W., Himmelbauer, I., Aberer, D., Schremmer, L., Petrakovic, I., Zappa, L., Preimesberger, W., Xaver, A., Annor, F., Ardö, J., Baldocchi, D., Bitelli, M., Blöschl, G., Boga, H., Brocca, L., Calvet, J.-C., Camarero, J. J., Capello, G., Choi, M., Cosh, M. C., van de Giesen, N., Hajdu, I., Ikonen, J., Jensen, K. H., Kanniah, K. D., de Kat, I., Kirchengast, G., Kumar Rai, P., Kyrouac, J., Larson, K., Liu, S., Loew, A., Moghaddam, M., Martínez Fernández, J., Mattar Bader, C., Morbidelli, R., Musial, J. P., Osenga, E., Palecki, M. A., Pellarin, T., Petropoulos, G. P., Pfeil, I., Powers, J., Robock, A., Rüdiger, C., Rummel, U., Strobel, M., Su, Z., Sullivan, R., Tagesson, T., Varlagin, A., Vreugdenhil, M., Walker, J., Wen, J., Wenger, F., Wigneron, J. P., Woods, M., Yang, K., Zeng, Y., Zhang, X., Zreda, M., Dietrich, S., Gruber, A., van Oevelen, P., Wagner, W., Scipal, K., Drusch, M., and Sabia, R.: The International Soil Moisture Network: serving Earth system science for over a decade, *Hydrology and Earth System Sciences*, 25, 5749–5804, <https://doi.org/10.5194/hess-25-5749-2021>, 2021.
- 550 Entekhabi, D., Yueh, S., O'Neill, P. E., Kellogg, K. H., Allen, A., Bindlish, R., Brown, M., Chan, S., Colliander, A., Crow, W. T., et al.: SMAP handbook—soil moisture active passive: Mapping soil moisture and freeze/thaw from space, Jet Propulsion Lab., California Inst. Technol., Pasadena, Calif, 2014.

- ESA: Land Cover CCI Product User Guide Version 2., available at: maps.elie.ucl.ac.be/CCI/viewer/download/ESACCI-LC-Ph2-PUGv2.0.pdf, 2017.
- Fan, D., Zhao, T., Jiang, X., García-García, A., Schmidt, T., Samaniego, L., Attinger, S., Wu, H., Jiang, Y., Shi, J., Fan, L., Tang, B.-H., Wagner, W., Dorigo, W., Gruber, A., Mattia, F., Balenzano, A., Brocca, L., Jagdhuber, T., Wigneron, J.-P., Montzka, C., and Peng, J.: A Sentinel-1 SAR-based global 1-km resolution soil moisture data product: Algorithm and preliminary assessment, *Remote Sensing of Environment*, 318, 114 579, <https://doi.org/https://doi.org/10.1016/j.rse.2024.114579>, 2025.
- Fang, B., Lakshmi, V., Cosh, M., Liu, P.-W., Bindlish, R., and Jackson, T. J.: A global 1-km downscaled SMAP soil moisture product based on thermal inertia theory, *Vadose Zone Journal*, 21, e20 182, <https://doi.org/https://doi.org/10.1002/vzj2.20182>, 2022.
- Feng, H. and Liu, Y.: Combined effects of precipitation and air temperature on soil moisture in different land covers in a humid basin, *Journal of Hydrology*, 531, 1129–1140, <https://doi.org/https://doi.org/10.1016/j.jhydrol.2015.11.016>, 2015.
- Flores, A., Herndon, K., Thapa, R., and Cherrington, E.: The SAR Handbook: Comprehensive Methodologies for Forest Monitoring and Biomass Estimation, <https://doi.org/10.25966/nr2c-s697>, 2019.
- Friedman, J. H.: Greedy Function Approximation: A Gradient Boosting Machine, *The Annals of Statistics*, 29, 1189–1232, <https://doi.org/10.1214/aos/1013203451>, 2001.
- Friedman, J. H.: Stochastic gradient boosting, *Computational Statistics & Data Analysis*, 38, 367–378, [https://doi.org/10.1016/S0167-9473\(01\)00065-2](https://doi.org/10.1016/S0167-9473(01)00065-2), 2002.
- Han, Q., Zeng, Y., Zhang, L., Wang, C., Prikaziuk, E., Niu, Z., and Su, B.: Global long term daily 1 km surface soil moisture dataset with physics informed machine learning, *Scientific Data*, 10, 101, 2023.
- Hao, G., Su, H., Zhang, R., Tian, J., and Chen, S.: A Two-Source Normalized Soil Thermal Inertia Model for Estimating Field-Scale Soil Moisture from MODIS and ASTER Data, *Remote Sensing*, 14, <https://doi.org/10.3390/rs14051215>, 2022.
- Hassan, S. T., Ghimire, C. P., and Lubczynski, M. W.: Remote sensing upscaling of interception loss from isolated oaks: Sardon catchment case study, Spain, *Journal of Hydrology*, 555, 489–505, <https://doi.org/https://doi.org/10.1016/j.jhydrol.2017.08.016>, 2017.
- Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N.: The ERA5 global reanalysis, *Quarterly Journal of the Royal Meteorological Society*, 146, 1999–2049, <https://doi.org/https://doi.org/10.1002/qj.3803>, 2020.
- Huffman, G. J., Bolvin, D. T., Joyce, R., Nelkin, E. J., Tan, J., Braithwaite, D., Hsu, K., Kelley, O. A., Nguyen, P., Sorooshian, S., Watters, D. C., West, B. J., and Xie, P.: Algorithm Theoretical Basis Document (ATBD) NASA Global Precipitation Measurement (GPM) Integrated Multi-satellite Retrievals for GPM (IMERG) Version 07, available online: https://gpm.nasa.gov/sites/default/files/2023-07/IMERG_V07_ATBD_final_230712.pdf, 2023.
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y.: LightGBM: A Highly Efficient Gradient Boosting Decision Tree, in: *Neural Information Processing Systems*, <https://api.semanticscholar.org/CorpusID:3815895>, 2017.
- Kemppinen, J., Niittynen, P., Rissanen, T., Tyystjärvi, V., Aalto, J., and Luoto, M.: Soil Moisture Variations From Boreal Forests to the Tundra, *Water Resources Research*, 59, e2022WR032 719, <https://doi.org/https://doi.org/10.1029/2022WR032719>, 2023.
- Kovačević, J., Cvijetinović, , Stančić, N., Brodić, N., and Mihajlović, D.: New Downscaling Approach Using ESA CCI SM Products for Obtaining High Resolution Surface Soil Moisture, *Remote Sensing*, 12, <https://doi.org/10.3390/rs12071119>, 2020.

Lakshmi, V. and Fang, B.: SMAP-Derived 1-km Downscaled Surface Soil Moisture Product, Version 1.,
595 <https://doi.org/https://doi.org/10.5067/U8QZ2AXE5V7B>, accessed: 30 September 2024, 2023.

Lal, P., Singh, G., Das, N. N., Entekhabi, D., Lohman, R., Colliander, A., Pandey, D. K., and Setia, R.: A multi-scale
algorithm for the NISAR mission high-resolution soil moisture product, *Remote Sensing of Environment*, 295, 113 667,
<https://doi.org/https://doi.org/10.1016/j.rse.2023.113667>, 2023.

Lal, P., Singh, G., Das, N. N., Entekhabi, D., Lohman, R. B., and Colliander, A.: Uncertainty estimates in the NISAR
600 high-resolution soil moisture retrievals from multi-scale algorithm, *Remote Sensing of Environment*, 311, 114 288,
<https://doi.org/https://doi.org/10.1016/j.rse.2024.114288>, 2024.

Larson, J., Wallerman, J., Peichl, M., and Laudon, H.: Soil moisture controls the partitioning of carbon stocks across a managed boreal forest
landscape, *Sci Rep*, 13, <https://doi.org/10.1038/s41598-023-42091-4>, 2023.

Larson, J., Vigren, C., Wallerman, J., Ågren, A. M., Mensah, A. A., and Laudon, H.: Tree growth potential and its relationship with soil moisture
605 conditions across a heterogeneous boreal forest landscape, *Sci Rep*, 14, <https://doi.org/10.1038/s41598-024-61098-z>, 2024.

Leavesley: A Modelling Framework for Improved Agricultural Water-Supply Forecasting, 2010.

Leavesley, G. H., David, O., Garen, D. C., Lea, J., Marron, J. K., Pagano, T. C., Perkins, T. R., and Strobel, M. L.: A Modeling Framework for
Improved Agricultural Water Supply Forecasting, 2008.

Li, M. and Yan, Y.: Comparative Analysis of Machine-Learning Models for Soil Moisture Estimation Using High-Resolution Remote-Sensing
610 Data, *Land*, 13, <https://doi.org/10.3390/land13081331>, 2024.

Li, X., Zhu, W., Xie, Z., Zhan, P., Huang, X., Sun, L., and Duan, Z.: Assessing the Effects of Time Interpolation of NDVI Composites on
Phenology Trend Estimation, *Remote Sensing*, 13, <https://doi.org/10.3390/rs13245018>, 2021.

Liu, M., Bárdossy, A., Li, J., and Jiang, Y.: Physically-based modeling of topographic effects on spatial evapotranspiration and soil moisture
patterns through radiation and wind, *Hydrology and Earth System Sciences*, 16, 357–373, <https://doi.org/10.5194/hess-16-357-2012>, 2012.

615 Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I.: From local
explanations to global understanding with explainable AI for trees, *Nature Machine Intelligence*, 2, 2522–5839, 2020.

Ma, Y., Hou, P., Zhang, L., Cao, G., Sun, L., Pang, S., and Bai, J.: High-Resolution Quantitative Retrieval of Soil Moisture Based
on Multisource Data Fusion with Random Forests: A Case Study in the Zoige Region of the Tibetan Plateau, *Remote Sensing*, 15,
<https://doi.org/10.3390/rs15061531>, 2023.

620 Mälikke, M., Hassler, S. K., Blume, T., Weiler, M., and Zehe, E.: Soil moisture: variable in space but redundant in time, *Hydrology and Earth
System Sciences*, 24, 2633–2653, <https://doi.org/10.5194/hess-24-2633-2020>, 2020.

Manninen, T., Jääskeläinen, E., Lohila, A., Korkiakoski, M., Räsänen, A., Virtanen, T., Muhić, F., Marttila, H., Ala-Aho, P., Markovaara-
Koivisto, M., Liwata-Kenttälä, P., Sutinen, R., and Hänninen, P.: Very High Spatial Resolution Soil Moisture Observation of Heterogeneous
Subarctic Catchment Using Nonlocal Averaging and Multitemporal SAR Data, *IEEE Transactions on Geoscience and Remote Sensing*, 60,
625 1–17, <https://doi.org/10.1109/TGRS.2021.3109695>, 2022.

Matsushima, D., Kimura, R., and Shinoda, M.: Soil Moisture Estimation Using Thermal Inertia: Potential and Sensitivity to Data Conditions,
Journal of Hydrometeorology, 13, 638 – 648, <https://doi.org/10.1175/JHM-D-10-05024.1>, 2012.

Merilä, P., Lindroos, A.-J., Helmisaari, H.-S., Hilli, S., Nieminen, T. M., Nöjd, P., Rautio, P., Salemaa, M., Ľupek, B., and Ukonmaanaho, L.:
Carbon Stocks and Transfers in Coniferous Boreal Forests Along a Latitudinal Gradient, *Ecosystems*, 27, <https://doi.org/10.1007/s10021->
630 024-00921-0, 2023.

- Meyer, R., Zhang, W., Kragh, S. J., Andreasen, M., Jensen, K. H., Fensholt, R., Stisen, S., and Looms, M. C.: Exploring the combined use of SMAP and Sentinel-1 data for downscaling soil moisture beyond the 1 km scale, *Hydrology and Earth System Sciences*, 26, 3337–3357, <https://doi.org/10.5194/hess-26-3337-2022>, 2022.
- Mohseni, F., Ahrari, A., Haunert, J.-H., and and, C. M.: The synergies of SMAP enhanced and MODIS products in a ran-
635 dom forest regression for estimating 1 km soil moisture over Africa using Google Earth Engine, *Big Earth Data*, 8, 33–57, <https://doi.org/10.1080/20964471.2023.2257905>, 2024.
- Molina, A. J. and del Campo, A. D.: The effects of experimental thinning on throughfall and stemflow: A contribu-
tion towards hydrology-oriented silviculture in Aleppo pine plantations, *Forest Ecology and Management*, 269, 206–213, <https://doi.org/https://doi.org/10.1016/j.foreco.2011.12.037>, 2012.
- 640 National Ecological Observatory Network: Soil water content and water salinity (DP1.00094.001), <https://doi.org/10.48443/QHMT-HH62>, 2025.
- Ning, J., Yao, Y., Tang, Q., Li, Y., Fisher, J. B., Zhang, X., Jia, K., Xu, J., Shang, K., Yang, J., Yu, R., Liu, L., Zhang, X., Xie, Z., and Fan, J.: Soil moisture at 30 m from multiple satellite datasets fused by random forest, *Journal of Hydrology*, 625, 130010, <https://doi.org/https://doi.org/10.1016/j.jhydrol.2023.130010>, 2023.
- 645 O, S., Orth, R., Weber, U., and Park, S. K.: High-resolution European daily soil moisture derived with machine learning (2003–2020), *Sci Data*, 9, <https://doi.org/10.1038/s41597-022-01785-6>, 2022.
- O’Neill, P., Chan, S., Njoku, E., Jackson, T., Bindlish, R., and Chaubell, J.: SMAP L3 Radiometer Global Daily 36 km EASE-Grid Soil Moisture, Version 9, <https://doi.org/10.5067/4XXOGX0OOW1S>, 2023a.
- O’Neill, P. E., Chan, S., Njoku, E. G., Jackson, T., Bindlish, R., Chaubell, J., and Colliander, A.: SMAP Enhanced L3 Radiometer Global and Polar Grid Daily 9 km EASE-Grid Soil Moisture. (SPL3SMP_E, *Version 6*), <https://doi.org/10.5067/M20OXIZHY3RJ>, *accessed* : 2024 – 10 – 07, 2023b.
- O’Neill, P., Bindlish, R., Chan, S., Chaubell, J., Colliander, A., Njoku, E., and Jackson, T.: Algorithm Theoretical Basis Document Level
650 2 3 Soil Moisture (Passive) Data Products, available online: https://nsidc.org/sites/default/files/l2_sm_p_atbd_rev_g_final_oct2021_0.pdf, Accessed: 2024-10-08, 2021.
- Pan, Y., Birdsey, R. A., Fang, J., Houghton, R., Kauppi, P. E., Kurz, W. A., Phillips, O. L., Shvidenko, A., Lewis, S. L., Canadell, J. G., Ciais, P., Jackson, R. B., Pacala, S. W., McGuire, A. D., Piao, S., Rautiainen, A., Sitch, S., and Hayes, D.: A Large and Persistent Carbon Sink in the World’s Forests, *Science*, 333, 988–993, <https://doi.org/10.1126/science.1201609>, 2011.
- 655 Pan, Y., Birdsey, R. A., Phillips, O. L., Houghton, R. A., Fang, J., Kauppi, P. E., Keith, H., Kurz, W. A., Ito, A., Lewis, S. L., Nabuurs, G.-J., Shvidenko, A., Hashimoto, S., Lerink, B., Schepaschenko, D., Castanho, A., and Murdiyarso, D.: The enduring world forest carbon sink, *Nature*, 631, 563–569, <https://doi.org/10.1038/s41586-024-07602-x>, 2024.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.: Scikit-learn: Machine Learning in Python, *Journal of Machine
660 Learning Research*, 12, 2825–2830, 2011.
- Peng, J., Loew, A., Merlin, O., and Verhoest, N. E. C.: A review of spatial downscaling of satellite remotely sensed soil moisture, *Reviews of Geophysics*, 55, 341–366, <https://doi.org/https://doi.org/10.1002/2016RG000543>, 2017.
- Poggio, L., de Sousa, L. M., Batjes, N. H., Heuvelink, G. B. M., Kempen, B., Ribeiro, E., and Rossiter, D.: SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty, *SOIL*, 7, 217–240, <https://doi.org/10.5194/soil-7-217-2021>, 2021.

- 665Ranney, K. J., Niemann, J. D., Lehman, B. M., Green, T. R., and Jones, A. S.: A method to downscale soil moisture to fine resolutions using topographic, vegetation, and soil data, *Advances in Water Resources*, 76, 81–96, <https://doi.org/https://doi.org/10.1016/j.advwatres.2014.12.003>, 2015.
- Rao, P., Wang, Y., Wang, F., Liu, Y., Wang, X., and Wang, Z.: Daily soil moisture mapping at 1 km resolution based on SMAP data for desertification areas in northern China, *Earth System Science Data*, 14, 3053–3073, <https://doi.org/10.5194/essd-14-3053-2022>, 2022.
- 670Sabaghy, S., Walker, J. P., Renzullo, L. J., and Jackson, T. J.: Spatially enhanced passive microwave derived soil moisture: Capabilities and opportunities, *Remote Sensing of Environment*, 209, 551–580, <https://doi.org/https://doi.org/10.1016/j.rse.2018.02.065>, 2018.
- Sandholt, I., Rasmussen, K., and Andersen, J.: A simple interpretation of the surface temperature/vegetation index space for assessment of surface moisture status, *Remote Sensing of Environment*, 79, 213–224, [https://doi.org/https://doi.org/10.1016/S0034-4257\(01\)00274-7](https://doi.org/https://doi.org/10.1016/S0034-4257(01)00274-7), recent *Advances in Remote Sensing of Biophysical Variables*, 2002.
- 675Schaefer, G., Cosh, M., and Jackson, T.: The USDA natural resources conservation service soil climate analysis network (SCAN), *Journal of Atmospheric and Oceanic Technology - J ATMOS OCEAN TECHNOL*, 24, 2073 – 2077, <https://doi.org/10.1175/2007JTECHA930.1>, 2007.
- Sehler, R., Li, J., Reager, J., and Ye, H.: Investigating Relationship Between Soil Moisture and Precipitation Globally Using Remote Sensing Observations, *Journal of Contemporary Water Research Education*, 168, 106–118, <https://doi.org/10.1111/j.1936-704x.2019.03324.x>, 2019.
- Shokati, H., Mashal, M., Noroozi, A., Abkar, A. A., Mirzaei, S., Mohammadi-Doqozloo, Z., Taghizadeh-Mehrjardi, R., Khosravani, P., Nabi-680 ollahi, K., and Scholten, T.: Random Forest-Based Soil Moisture Estimation Using Sentinel-2, Landsat-8/9, and UAV-Based Hyperspectral Data, *Remote Sensing*, 16, <https://doi.org/10.3390/rs16111962>, 2024.
- Strahler, A., Muchoney, D., Borak, J., Friedl, M., Gopal, S., Lambin, E., and Moody, A.: MODIS Land Cover Product Algorithm Theoretical Basis Document (ATBD) Version 5.0, available at: https://modis.gsfc.nasa.gov/data/atbd/atbd_mod12.pdf, 1999.
- Turner, M., Beer, C., Santoro, M., Carvalhais, N., Wutzler, T., Schepaschenko, D., Shvidenko, A., Kompter, E., Ahrens, B., Levick, S. R., 685 and Schmulilius, C.: Carbon stock and density of northern boreal and temperate forests, *Global Ecology and Biogeography*, 23, 297–310, <https://doi.org/https://doi.org/10.1111/geb.12125>, 2014.
- Tramblay, Y. and Quintana Seguí, P.: Estimating soil moisture conditions for drought monitoring with random forests and a simple soil moisture accounting scheme, *Natural Hazards and Earth System Sciences*, 22, 1325–1334, <https://doi.org/10.5194/nhess-22-1325-2022>, 2022.
- Walker, X. J., Baltzer, J. L., Cumming, S. G., Day, N. J., Ebert, C., Goetz, S., Johnstone, J. F., Potter, S., Rogers, B. M., Schuur, E. 690 A. G., Turetsky, M. R., and Mack, M. C.: Increasing wildfires threaten historic carbon sink of boreal forest soils, *Nature*, pp. 520–523, <https://doi.org/10.1038/s41586-019-1474-y>, 2019.
- Wei, Z., Meng, Y., Zhang, W., Peng, J., and Meng, L.: Downscaling SMAP soil moisture estimation with gradient boosting decision tree regression over the Tibetan Plateau, *Remote Sensing of Environment*, 225, 30–44, <https://doi.org/https://doi.org/10.1016/j.rse.2019.02.022>, 2019.
- 695Yin, J., Zhan, X., Liu, J., Moradkhani, H., Fang, L., and Walker, J. P.: Near-real-time one-kilometre Soil Moisture Active Passive soil moisture data product, *Hydrological Processes*, 34, 4083–4096, <https://doi.org/https://doi.org/10.1002/hyp.13857>, 2020.
- Zabret, K., Rakovec, J., Mikoš, M., and Šraj, M.: Influence of Raindrop Size Distribution on Throughfall Dynamics under Pine and Birch Trees at the Rainfall Event Level, *Atmosphere*, 8, <https://doi.org/10.3390/atmos8120240>, 2017.
- Zhang, D., Lu, L., Li, X., Zhang, J., Zhang, S., and Yang, S.: Spatial Downscaling of ESA CCI Soil Moisture Data Based on Deep Learning 700 with an Attention Mechanism, *Remote Sensing*, 16, <https://doi.org/10.3390/rs16081394>, 2024.

Zhang, Y., Liang, S., Ma, H., He, T., Wang, Q., Li, B., Xu, J., Zhang, G., Liu, X., and Xiong, C.: Generation of global 1 km daily soil moisture product from 2000 to 2020 using ensemble learning, *Earth System Science Data*, 15, 2055–2079, <https://doi.org/10.5194/essd-15-2055-2023>, 2023.

Zheng, C., Jia, L., and Zhao, T.: A 21-year dataset (2000–2020) of gap-free global daily surface soil moisture at 1-km grid resolution, *Sci Data*, 10, vzj2012.0097, <https://doi.org/10.1038/s41597-023-01991-w>, 2023.