Pacific Northwest
NATIONAL LABORATORY

PNNL- 36677

An ML-based terrestrial data fusion and augmentation framework to enable advanced understanding of the terrestrial carbon and water interactions

September 2024

Mingjie Shi
Lingcheng Li
Xinming Lin
Yilin Fang
Z. Jason Hou

U.S. DEPARTMENT OF ENERGY

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

DISCLAIMER

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from
the Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062
www.osti.gov
ph: (865) 576-8401
fox: (865) 576-5728
email: reports@osti.gov

Available to the public from the National Technical Information Service
5301 Shawnee Rd., Alexandria, VA 22312
ph: (800) 553-NTIS (6847)
or (703) 605-6000
email: info@ntis.gov
Online ordering: http://www.ntis.gov

PNNL- 36677

An ML-based terrestrial data fusion and augmentation framework to enable advanced understanding of the terrestrial carbon and water interactions

September 2024

Mingjie Shi
Lingcheng Li
Xinming Lin
Yilin Fang
Z. Jason Hou

Prepared for
the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99354

Abstract

Soil moisture is essential to the terrestrial carbon and water cycles and land–atmosphere interactions. There are various types of soil moisture data, and each type has the distinct spatiotemporal strengths and limitations, depending on the diverse applications and retrieval methodologies of different data types (Li et al., in review; The PNNL-82151 FY23 Report). However, the limitations of different soil moisture data in terms of accuracy and spatiotemporal coverage hinder our ability to further understand the soil moisture dynamics across scales. To have a gap free soil moisture data product with a fine spatiotemporal coverage and vertical profiles, we train extreme gradient boosting (XGBoost) models by using (1) in-situ soil moisture measurements from the International Soil Moisture Network (ISMN), (2) soil moisture from the ECMWF reanalysis (ERA) at the 9 km and sub-daily spatiotemporal resolution, (3) the Daymet meteorological fields, and (4) data products that characterize surface conditions, including soil texture, organic content, topography, vegetation type, and rooting depth. We use the trained XGBoost models that have consistent performance across seven soil layers, i.e., 0–5 cm, 5–10 cm, 10–20 cm, 20–40 cm, 40–60 cm, 60–100 cm, and 100–200 cm, and the gridded model predictors to generate a soil moisture data at the 1 km and daily spatiotemporal resolution for the Continental United States (CONUS) from 2001–2020. This dataset can be broadly used for Earth system model benchmark, monitoring extreme weathers, making informed decisions regarding agriculture, water resource management, climate change mitigation, and ecosystem preservation.

Summary

This study develops a gap free soil moisture data product at the 1 km and daily spatiotemporal resolution with seven soil layers over the Continental United States (CONUS) during 2001–2020. To develop this data, we train machine learning (i.e., XGBoost) models by using existing datasets, including in-situ soil moisture measurements, a reanalysis soil moisture dataset at the 9 km and sub-daily spatiotemporal resolution, meteorological fields from Daymet, and data products that characterize surface conditions, such as soil texture, organic content, topography, vegetation type, and rooting depth. The XGBoost models have a high performance and are used to generate the daily, 1km soil moisture data over CONUS. This CONUS based soil moisture data product can enhance the fundamental understanding of the carbon and water interactions and land–atmosphere feedback under the varying climate and extreme weather conditions at a fine spatiotemporal resolution. The data product can be adequately used to benchmark and parameterize Earth system models, such as the Energy Exascale Earth System Model (E3SM) developed by Department of Energy (DOE).

Acknowledgments

This research was supported by the Earth and Biological Sciences Investment, under the Laboratory Directed Research and Development (LDRD) Program at Pacific Northwest National Laboratory (PNNL). PNNL is a multi-program national laboratory operated for the U.S. Department of Energy (DOE) by Battelle Memorial Institute under Contract No. DE-AC05-76RL01830.

Acronyms and Abbreviations

CONUS: Continental United States
DOE: Department of Energy
E3SM: Energy Exascale Earth System Model
ESM: Earth system models
ML: Machine learning
ERA: The ECMWF reanalysis
ISMN: International Soil Moisture Network
SHAP: SHapley Additive exPlanations
XGBoost: Extreme Gradient Boosting

Contents

Figures

1.0 Introduction

Soil moisture is a key to the terrestrial carbon and water interactions, and determines the water and energy fluxes, which are essential to the land–atmosphere feedback (Humphrey et al., 2021; Liu et al., 2022; Ochsner et al., 2013; Seneviratne et al., 2010). To develop soil moisture products with high data quality and accuracy, both the observational and modelling communities have been working on the enhancement of soil moisture monitoring and simulation (e.g., Chan et al., 2016; Dorigo et al., 2013; Rodell et al., 2004). Therefore, there are a variety of types of soil moisture data sources, including in-situ soil moisture measurements, remote sensing products, land surface model output, reanalysis data, and machine learning data products (Li et al., in review). These data types show distinct strengths and limitations, reflecting their diverse applications and methodologies.

Soil moisture observational methods include in-situ soil moisture sensors and remote-sensing. In-situ soil moisture has high temporal but limited spatial coverages (Dorigo et al., 2021), thus it is insufficient to use in-situ data to represent soil moisture changes at the kilometer (e.g., the gridcell sizes of Earth system models [ESMs]), regional, and global scales. Remote sensing can monitor surface soil moisture from the regional to global scales. However, remote-sensing data samplings are limited by the spatial resolution, revisiting time frequency, and penetration depth of instrumental design. For example, due to the failure of the radar of the Soil Moisture Active Passive [SMAP]), SMAP can only monitor surface (i.e., 0–5 cm) soil moisture at the 9 km spatial resolution with a global survey per 2–3 days. Therefore, there is still a lack of data coverage and consistent resolution across both time and space of soil moisture measurements.

By using meteorological conditions and surface features, land surface models can solve water and energy balance equations and simulate soil moisture in different soil layers (e.g., Schaake et al. 2004). Reanalysis systems, on the other hand, can use existing soil moisture measurements and the land surface model frameworks to generate soil moisture data products with continued spatiotemporal coverages (e.g., Balsamo et al., 2013; Lievens et al., 2017). However, these datasets also have relatively coarse spatial resolutions (e.g., from 9 km to 0.25 deg; Li et al., in review) and large uncertainties determined by the input data and the physics and parameterizations of the host models of the reanalysis systems. All these limitations hinder our ability to accurately evaluate ESMs and make informed decisions regarding agriculture, water resource management, climate change mitigation, and ecosystem preservation.

To develop high quality soil moisture data with high spatiotemporal coverage, this study uses both in-situ and gridded data from a variety of sources to train machine learning (ML) models (Section 1.1), and develops a 1 kilometer and daily soil moisture dataset with seven layers, ranging from the surface to a depth of 2 meters for the contiguous United States (CONUS) over the period 2001–2020. This report summarizes methods used for generating of this dataset, discusses the broader impacts and future directions of this research.

1.1 Methods and Deliverables

To develop a gap free soil moisture data product with multiple vertical layers at the 1 km and daily spatiotemporal resolution, we train extreme gradient boosting (XGBoost) models by using data from several sources: (1) in-situ soil moisture measurements from the International Soil Moisture Network (ISMN), (2) hydrological features (e.g., soil moisture, runoff, snow water equivalent, groundwater table depth) obtained from ECMWF reanalysis (ERA) at the 9 km and sub-daily spatiotemporal resolution, (3) the Daymet meteorological fields, and (4) data products that characterize surface conditions, such as soil texture, organic content, topography, vegetation type, and rooting depth. The predictors used for the ML model include meteorological conditions (e.g., precipitation, air temperature, shortwave radiation), soil properties (i.e., sand percentage, organic content), hydrological features (e.g., multi-layer soil moisture, snow water equivalent, water table depth, evapotranspiration), vegetation related features (e.g., LAI, rooting depth, land cover), and topography (i.e., elevation, slope, aspect). The target variable is the in-situ soil moisture data collected from ISMN at seven different depths, i.e., 0–5 cm, 5–10 cm, 10–20 cm, 20–40 cm, 40–60 cm, 60–100 cm, and 100–200 cm. Separate models are trained for each soli layer. We use the trained XGBoost models that have consistent performance in seven soil layers, and the grided model predictors, and generate our soil moisture data over CONUS for the period 2001–2020 (Figure 1).

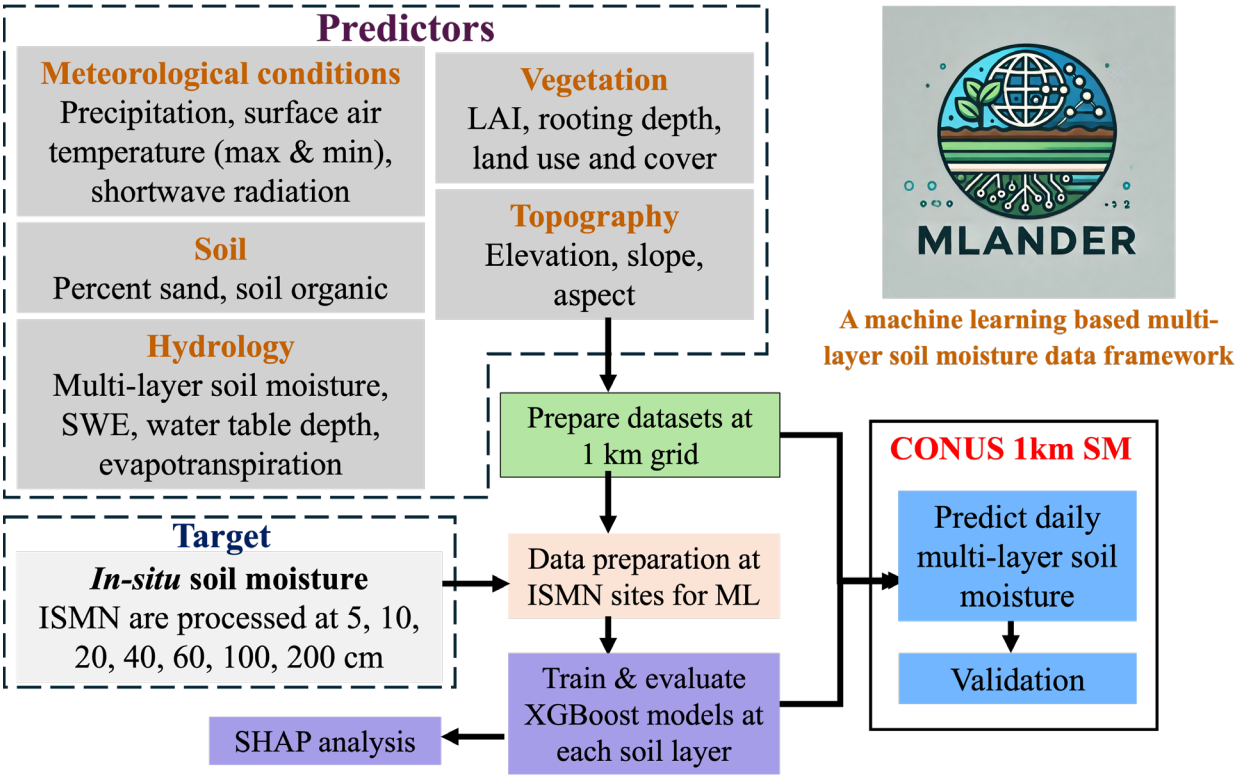

Figure 1. The workflow of this study.

Due to the large data volume (~11 terabyte), the process of running the ML code for data generation is still ongoing. We show the soil moisture training and testing results and a one-day (August 1st, 2010) soil moisture record in Figures 2(a)–(d). Figure 2(a) shows the robustness of the testing statistics. In Figures 2(b)–(c), the dots represent the locations of all the ISMN sites with

0–5 cm soil moisture measurements, and the coefficient of determination (R^2) values and root mean square errors (RMSE) further show that the ML model trained for the 0–5 cm layer performs well, spatially. In addition, the performance of the ML models varies across different soil layers because of the differences in the number of available ISMN soil moisture sites that can be used for ML model training (Figure now shown). Here, we use a one-day 0–5 cm soil moisture spatial map to indicate that our research framework (Figure 1) is well established for soil moisture data development, and we will further perform data validation spatially and temporally using the in-situ (e.g., AmeriFlux) soil moisture datasets.

We also use SHapley Additive exPlanations (SHAP) to gain an in-depth interpretation of the contributions of different XGBoost model predictors to soil moisture (Figure 1). The SHAP analysis is based on the five predictor types, and the results show that the hydrological features and soil characteristics are the top two most important factors influencing the 0–5 cm soil moisture (Figure 2(e)). The importance of different factors varies across soil layers. Overall, soil characteristics (i.e., percentage of sand and organic content) are the most important predictors in determining soil moisture profiles (i.e., soil moisture in the 5–200 cm layers; figure not shown). This research implies the substantial needs of enhancing the spatial and vertical coverage of both in-situ and remote sensing soil moisture measurements.

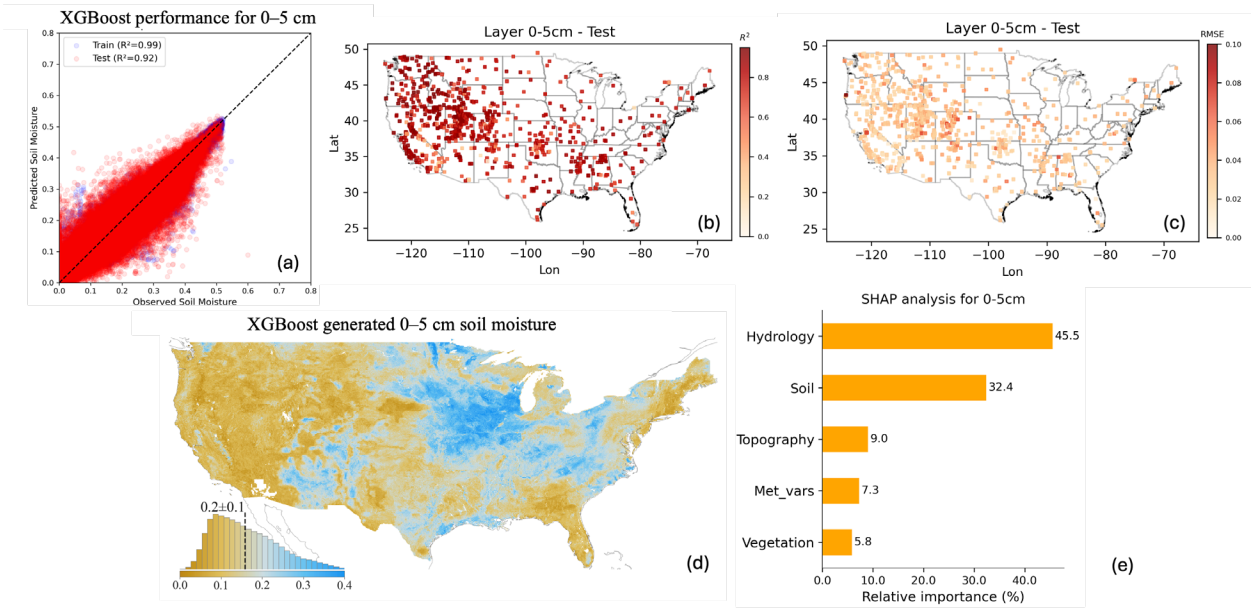

Figure 2. The results of ML model training and testing (a)–(c), (d) the developed soil moisture product (the day of August 1st, 2010 is shown here as a demonstration), and (e) the SHAP analysis based on the five types of predictors (Figure 1) in the 0–5 cm layer over CONUS.

2.0 Boarder Impacts and Future Directions

Most of the existing ML based soil moisture products focus primarily on the surface soil moisture (Li et al., 2024, in review). However, soil moisture at the rooting zone determines the carbon–water–energy exchanges between the land and atmosphere through the root systems with varying complexity across ecosystems (Fan et al., 2017). The existing large-scale soil moisture measurements are limited to the surface due to the remote-sensing instrument penetration capacity and the soil moisture retrieval methods (Wang et al., 2024). Thus, observational based studies that quantify the relationships between soil moisture and gross primary production (GPP) at the regional scale mostly rely on surface soil moisture due to the lack of products that can represent root zone soil moisture (e.g., He et al., 2017). By using a set of algorithms dedicated to the estimation of terrestrial evaporation and root-zone soil moisture from satellite data, Marten et al. (2017) developed the Global Land Evaporation Amsterdam Model (GLEAM) soil moisture, which has root zone soil moisture. However, the GLEAM soil moisture is at a relatively coarse spatiotemporal resolution, i.e., 0.25 degree and monthly. Therefore, the existing soil moisture products limit the comprehensive understanding of soil moisture dynamics at high spatiotemporal resolutions across different soil layers.

The soil moisture product developed by this study provides 20 years (2001–2020) of soil moisture data records at the 1 km and daily spatiotemporal resolution to the depth of 2 meters over CONUS. This invaluable dataset can be used to enhance the understanding of the carbon–water interactions (e.g., soil moisture induced GPP changes) across different biomes of CONUS. This dataset allows for advanced investigation of soil moisture dynamics under water stressed conditions, including drought, flash drought, heatwave, and fire weathers. Thus, we can use this data to further study the soil moisture dynamics under pre- and post-water stresses with various severity, which facilities a comprehensive understanding of the carbon–water–energy interactions and land–atmosphere feedback.

This data product can significantly benefit the modelling community. By solving water and energy balance equations, ESMs or their land components simulate soil moisture in different soil layers, and can develop spatially- and temporally- continuous records with vertical profiles of soil moisture (e.g., Schaake et al., 2004). As ESMs are progressively advancing towards the kilometer scale (Li et al., 2024), this data can be broadly used by the modelling community for model evaluation and uncertainty quantification. Due to the high spatiotemporal resolution of this data, which have high computational demands for running the ML models and data archiving, we use CONUS as the testbed for data development of this study. The ISMN has soil moisture sensors installed over the globe (Dorigo et al., 2017). Since our ML models have demonstrated feasibility and success over CONUS, in our next step, we will use this framework to develop a kilometer scale global soil moisture dataset with vertical soil layers. The development of a global dataset will fundamentally enhance the soil moisture research for both the observational and modelling communities.

3.0 References

Balsamo, G., Albergel, C., Beljaars, A., Boussetta, S., Cloke, H., Dee, D., ... & Vitart, F. (2013). ERA-Interim/Land: a global land water resources dataset. *Hydrology & Earth System Sciences Discussions*, *10*(12).

Chan, S. K., Bindlish, R., O'Neill, P. E., Njoku, E., Jackson, T., Colliander, A., ... & Kerr, Y. (2016). Assessment of the SMAP passive soil moisture product. IEEE Transactions on Geoscience and Remote Sensing, 54(8), 4994-5007.

Dorigo, W. A., Xaver, A., Vreugdenhil, M., Gruber, A., Hegyiová, 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 J., 12, https://doi.org/10.2136/vzj2012.0097, 2013.

Fan, Y., Miguez-Macho, G., Jobbágy, E. G., Jackson, R. B., & Otero-Casal, C. (2017). Hydrologic regulation of plant rooting depth. *Proceedings of the National Academy of Sciences*, *114*(40), 10572-10577.

He, L., Chen, J. M., Liu, J., Bélair, S., & Luo, X. (2017). Assessment of SMAP soil moisture for global simulation of gross primary production. *Journal of Geophysical Research: Biogeosciences*, *122*(7), 1549-1563.

Humphrey, V., Berg, A., Ciais, P., Gentine, P., Jung, M., Reichstein, M., ... & Frankenberg, C. (2021). Soil moisture–atmosphere feedback dominates land carbon uptake variability. *Nature*, *592*(7852), 65-69.

Li, L., Bisht, G., Hao, D., & Leung, L. R. (2024). Global 1 km land surface parameters for kilometer-scale Earth system modeling. Earth System Science Data, 16(4), 2007-2032.

Li, L., Lin, X., Fang, Y., et al, A Unified Ensemble Soil Moisture Dataset Across the Continental United States, Scientific Data, in review.

Lievens, H., Reichle, R. H., Liu, Q., De Lannoy, G. J., Dunbar, R. S., Kim, S. B., ... & Wagner, W. (2017). Joint Sentinel-1 and SMAP data assimilation to improve soil moisture estimates. Geophysical research letters, 44(12), 6145-6153.

Liu, W., Zhang, Q., Li, C., Xu, L., & Xiao, W. (2022). The influence of soil moisture on convective activity: a review. Theoretical and Applied Climatology, 149(1-2), 221-232.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., ... & Lee, S. I. (2020). From local explanations to global understanding with explainable AI for trees. Nature machine intelligence, 2(1), 56-67.

Martens, B., Miralles, D. G., Lievens, H., et al., 2017. GLEAM v3: Satellite-based land evaporation and root-zone soil moisture. Geoscientific Model Development, 10(5), 1903-1925.

Ochsner, T. E., Cosh, M. H., Cuenca, R. H., Dorigo, W. A., Draper, C. S., Hagimoto, Y., ... & Zreda, M. (2013). State of the art in large-scale soil moisture monitoring. *Soil Science Society of America Journal*, *77*(6), 1888-1919.

Rodell, M., et al. (2004), The global land data assimilation system, *Bull. Am. Meteorol. Soc.*, 85(3), 381–394.

Thornton, P.E., Shrestha, R., Thornton, M. *et al.* (2021). Gridded daily weather data for North America with comprehensive uncertainty quantification. *Sci Data* 8, 190.

Seneviratne, S. I., Corti, T., Davin, E. L., Hirschi, M., Jaeger, E. B., Lehner, I., ... & Teuling, A. J. (2010). Investigating soil moisture–climate interactions in a changing climate: A review. *Earth-Science Reviews*, *99*(3-4), 125-161.

Schaake, J. C., Duan, Q., Koren, V., Mitchell, K. E., Houser, P. R., Wood, E. F., ... & Tarpley, J. D. (2004). An intercomparison of soil moisture fields in the North American Land Data Assimilation System (NLDAS). *Journal of Geophysical Research: Atmospheres*, *109*(D1).

Wang, C., Wu, Q., Weimer, M. and Zhu, E., 2021. FLAML: A fast and lightweight automl library. Proceedings of Machine Learning and Systems, 3, pp.434-447.

Appendix A – Milestones

Manuscripts:

1. Li, L., Lin, X., Fang, Y., et al, 2024, Assessment of Multi-Source Soil Moisture Products Across the Continental United States, *Scientific Data*, (in revision).
2. Huang, J., Sehgal, V., Fisher, J.B., et al, **inc**. Fang, Y., Li, Y., and Shi, M., 2024, High-Resolution Soil Moisture and Evapotranspiration: Bridging the Gap between Science and Society, *Water Resources Research*, (in review).
3. Hao, Y., Mao, J., Bachmann, C. M., **inc.** Fang, Y., Li, Y., and Shi, M., 2024, Soil Moisture Controls over Greenhouse Gas Emissions and Carbon Sequestration: A Review, *npj Climate and Atmospheric Sciences*, (in revision).
4. Li, C., Batbeniz F., Koren, G., et al., **inc.** Fang, Y., Li, Y., and Shi, M., 2024, The role of soil moisture in climate-driven compound hazards, *manuscript proposal to Nature Geoscience*, (in review).

Presentations:

5. Shi, M., Li, L., Lin, X., et al, Unified Ensemble Soil Moisture Datasets Across the Continental United States**,** September 11, 2024, RUBISCO SMWG Mini Workshop (Talk).
6. Li, L., Lin, X., Fang, Y., et al, 2024, Developing High-Resolution Root Zone Soil Moisture Using Machine Learning Across the Contiguous United States, AGU 2024 Meeting Abstract.
7. Shi., Li, L., Lin, X., et al., 2024, A Unified Ensemble Soil Moisture Dataset Across the Continental United States, AGU 2024 Meeting Abstract.
8. Li, L., 2024, Advancing kilometer Scale Land Surface Modeling using E3SM Land Model: Developments and Case Studies, Digital Earths Webinar Series.

Datasets:

9. Li, L., Lin, X., Fang, Y., et al, 2023, Evaluation of Multi-Fidelity Soil Moisture Products Across the Continental United States**,** PNNL DataHub, https://doi.org/10.25584/2001040.

Pacific Northwest National Laboratory

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99354

1-888-375-PNNL (7665)

www.pnnl.gov