SMAP-HydroBlocks, a 30-m satellite-based soil moisture dataset for the conterminous US

Noemi Vergopolan1✉, Nathaniel W. Chaney2, Ming Pan1,3, Justin Sheffield4, Hylke E. Beck5, Craig R. Ferguson6, Laura Torres-Rojas2, Sara Sadri7 & Eric F. Wood1

Soil moisture plays a key role in controlling land-atmosphere interactions, with implications for water resources, agriculture, climate, and ecosystem dynamics. Although soil moisture varies strongly across the landscape, current monitoring capabilities are limited to coarse-scale satellite retrievals and a few regional in-situ networks. Here, we introduce SMAP-HydroBlocks (SMAP-HB), a high-resolution satellite-based surface soil moisture dataset at an unprecedented 30-m resolution (2015–2019) across the conterminous United States. SMAP-HB was produced by using a scalable cluster-based merging scheme that combines high-resolution land surface modeling, radiative transfer modeling, machine learning, SMAP satellite microwave data, and in-situ observations. We evaluated the resulting dataset over 1,192 observational sites. SMAP-HB performed substantially better than the current state-of-the-art SMAP products, showing a median temporal correlation of 0.73 ± 0.13 and a median Kling-Gupta Efficiency of 0.52 ± 0.20. The largest benefit of SMAP-HB is, however, the high spatial detail and improved representation of the soil moisture spatial variability and spatial accuracy with respect to SMAP products. The SMAP-HB dataset is available via zenodo and at https://waterai.earth/smaphb.

Background & Summary
Detailed and accurate information on the spatiotemporal distribution of soil moisture is important for numerous applications, such as monitoring of drought1–4 and crop irrigation demands4–6; mapping antecedent conditions that trigger wildfires7–9, landslides8,10, and flooding11,12; and quantifying water, energy, and carbon fluxes between the land and atmosphere13–15. Depending on the landscape heterogeneity, such physical processes can occur at the 1–100 m spatial scale, at which in-situ sensors could provide detailed information. However, in-situ observations’ representativeness can be limited to only a few meters from the sensors, they are costly to deploy and maintain, and therefore are not widely available at continental extents.

With satellite observations increasingly available16, optical and near-infrared satellite sensors (e.g., MODIS, Landsat, and Sentinel-2) can provide proxies for estimating soil moisture at high spatial resolution (10–250 m)17–19. However, estimates from these sensors can suffer attenuation from the atmosphere, high cloud coverage, dense vegetation, and infrequent revisit time (~1–2 weeks). Alternatively, passive microwave sensors were designed to penetrate through clouds and dense vegetation to retrieve surface soil moisture with a 25–50-km spatial resolution and 2–3-days revisit time20–23. NASA’s Soil Moisture Active-Passive mission22 (SMAP), for example, has a 36-km spatial resolution (or 9 km via the resampled SMAP L3 Enhanced23,26 product). Combining such passive sensors with a active sensor (e.g., Sentinel-1) and/or assimilating them into physical models can provide estimates with a 1–5-km27,28 and 9–25-km29–32 spatial resolution, respectively. These capabilities critically contributed for aiding regional- to global-scale water resources applications33,34. However, they still lack the spatial detail and accuracy necessary for local-scale (1–100 m) applications35,36. Thus, despite the increased demand, obtaining high-resolution data at continental extents remains a challenge.

1Princeton University, Department of Civil and Environmental Engineering, Princeton, NJ, United States. 2Duke University, Department of Civil and Environmental Engineering, Durham, NC, United States. 3Center for Western Weather and Water Extremes, Scripps Institution of Oceanography, University of California, San Diego, CA, United States. 4Southampton University, School of Geography and Environmental Science, Southampton, United Kingdom. 5GloH2O, Almere, the Netherlands. 6University at Albany, State University of New York, Atmospheric Sciences Research Center, Albany, NY, United States. 7University of Saskatchewan, Global Institute for Water Security, Saskatoon, Canada. ✉E-mail: noemi@princeton.edu
To address the need for high-resolution satellite-based soil moisture estimates, Vergopolan et al.37 developed an approach that combines HydroBlocks, a cluster-based high-resolution land surface model, with a Tau-Omega Radiative Transfer Model (RTM). This approach fuses HydroBlocks-RTM outputs (30-m resolution) and SMAP L3 brightness temperature observations (36-km resolution) using a cluster-based merging scheme. The uniqueness of this approach resides in leveraging HydroBlocks’ complex tiling for merging satellite observations in the cluster space. In this way, satellite-based soil moisture at an effective 30-m resolution can be achieved in a computationally efficient manner that, otherwise, would be challenging to scale using traditional regular grid approaches. Here, we introduce a new parameterization for this cluster-based merging scheme, which uses

---

Fig. 1 Climatology of SMAP-HB surface soil moisture, representing the top 5-cm of the soil column at 30-m spatial resolution over the CONUS (2015–2019). Insets show the soil moisture spatial detail at locations with different hydroclimatic and topographic conditions. For each location, two boxes of 100-km and 20-km size are shown along with illustrative Landsat satellite imagery. Water bodies are shown in blue. Inset labels show scale bar. Interactive visualization of the 30-m data is available at https://waterai.earth/smaphb.
Machine learning to regionalize the relationships between landscape characteristics and data from satellites, models, and in-situ observations. We apply this new approach to fuse brightness temperature from HydroBlocks-RTM (30-m resolution) and the SMAP L3 Enhanced (9-km resolution) product, and we demonstrate its scalability by developing SMAP-HydroBlocks (SMAP-HB), the first hyper-resolution satellite-based surface soil moisture dataset at over a continental extent (Fig. 1). SMAP-HB is available at 6-30 m spatial resolution (2015–2019) over the conterminous United States (CONUS).

SMAP-HB revealed a substantial spatial variability (Fig. 1), reflecting the complex interactions between hydroclimate and topography across CONUS, but also the impact of soil properties and land use evident at the local scales (insets). SMAP-HB captures the imprint of river reaches and wet riparian corridors in both wet and dry hydroclimates, such as over the wetlands of the Okefenokee National Wildlife Refuge (inset 6) and the perennial tributaries replenished by snowmelt in the California’s Sierra Nevada (inset 1). We evaluate the accuracy of SMAP-HB using in-situ observations and compare its performance against the HydroBlocks and SMAP L3E (representing the baseline products), and the NASA’s SMAP L4 data assimilation product \(^{(25,39)}\) at their respective spatial resolution (Fig. 2). Overall, SMAP-HB has the best temporal statistics, with a Root Mean Square Error (RMSE) of 0.07 m\(^3\)/m\(^3\) for the training and testing sites (Table S1), and Kling-Gupta Efficiency (KGE) scores of 0.53 and 0.48 for the training and testing sites, respectively. SMAP-HB showed temporal correlations of 0.71 and 0.77 at the training and testing sites, respectively, compared to 0.73 and 0.74 for SMAP L4. SMAP-HB performed substantially better than the baseline products (Fig. 2b). The largest gains are in the KGE score, with a 0.12 improvement compared to SMAP L3E. SMAP-HB showed the highest spatial accuracy (Fig. 3), evaluated through the spatial correlation across the CONUS (0.66), the New York Mesonet (0.42), and the Oklahoma Mesonet (0.54). As such, we anticipate this dataset can transform efforts to monitor water resources and natural hazards by enabling better representation and understanding of water, energy, and carbon cycle processes at spatial scales that have so far been unresolved.

**Methods**

**Satellite brightness temperature and soil moisture retrievals.** We used data from NASA’s Soil Moisture Active Passive (SMAP) Mission, in particular version 3 of the L3 Enhanced Global 9-km product \(^{(26)}\) (SMAP L3E). Relative to other satellites, the SMAP L-Band microwave sensor tends to offer the best sensitivity to soil moisture retrieval at the top 5 cm of the soil \(^{(40,41)}\). The SMAP L3E provides morning and afternoon composites of brightness temperature, ancillary data for the Tau-Omega Radiative Transfer Model, retrieved soil moisture, time of measurement, and quality control flags. This product spans from 31 March 2015 to the present, with a 2–3 days revisit time. We used the vertically polarized brightness temperature corrected and flagged for the presence of frozen ground, snow cover, transient water, and active precipitation at the time of the satellite overpass. SMAP L3E soil moisture retrievals were only used for evaluation purposes.

To expand the soil moisture dataset evaluation, we included the SMAP L4 Global 3-h 9-km EASE-Grid Surface Soil Moisture Analysis Update version \(^{(50)}\). This product was computed via dynamic assimilation of SMAP brightness temperatures into the NASA Catchment land surface model \(^{(42)}\) using a customized version of the Goddard Earth Observing System (GEOS) land data assimilation system.

**HydroBlocks land surface model.** Satellite Earth observation and physiographic data are increasingly available at higher spatial resolutions. However, traditional land surface models struggle to harness the opportunities afforded by these data due to their complex representation of physical processes, and they are unable to computationally scale with the massive data volumes across large domains. To address this challenge, the HydroBlocks land surface model was designed to leverage the repeating spatial patterns that exist over the landscape by implementing a hierarchical clustering algorithm to define its computational mesh \(^{(34)}\). This approach groups the fine-scale drivers of the landscape spatial heterogeneity using, for example, 30-m land cover, soil properties, topography data, into complex tiles/clusters of similar hydrologic behavior, herein called Hydrologic Response Unit (HRU) \(^{(44,45)}\). In this way, HydroBlocks simulates hydrological processes within the HRUs instead of regular grids, yielding an effective 30-m spatial resolution. This allows HydroBlocks to leverage the complex physics of land surface models while efficiently reducing the system’s dimensionality and computational requirements. For example, a 9-km grid box containing 90,000 30-m grid cells can be represented with ~300–500 clusters (a 180–300 times reduction) depending on the landscape complexity.

Here, HydroBlocks was set up to simulate soil moisture and soil temperature with a 3-h 30-m resolution, between 2015–2019 (with model spin up between 2010–2014). We used the 1-h 3-km Princeton CONUS Forcing \(^{(47)}\) (PCF) dataset as meteorological input. PCF downscales the North American Land Data Assimilation System 2 (NLDAS-2) data with several higher resolution datasets. PCF precipitation combines the Stage IV and Stage II radar/gauge products with NLDAS-2, and the shortwave radiation combines GOES Surface and Inolation Product (GSIP) with NLDAS-2. PCF also uses an elevation-based downscaling/fusion procedure to ensure physical consistency and mass/energy balance. To parameterize the land surface model, we used a 30-m SRTM-based elevation dataset \(^{(46)}\) and post-processed it to remove pits and derived slope, aspect, topographic index, flow direction, flow accumulation values, and height above the nearest drainage. We used the 2016 30-m land cover classification from the National Land Cover Database \(^{(47)}\) (NLCD). The soil-water hydraulic parameters were from the 30-m Probabilistic Remapping of SSURGO \(^{(48)}\) (POLARIS) dataset. No model calibration was performed, to allow the in-situ soil moisture observations to be used for independent validation.

To obtain HRU-level 30-m brightness temperature estimates, HydroBlocks was combined with a Tau-Omega Radiative Transfer Model (HydroBlocks-RTM). Using the 30-m HydroBlocks soil moisture (of the top 5-cm of the soil column), soil temperature, 30-m POLARIS clay content, and the 9-km SMAP L3E ancillary data (albedo, vegetation optical depth, surface roughness), we computed the 30-m brightness temperature with HydroBlocks-RTM. Further details of the HydroBlocks-RTM implementation are presented in Vergopolan et al. \(^{(39)}\).
Fig. 2 Temporal evaluation of daily SMAP L3E, SMAP L4, HydroBlocks, and SMAP-HB soil moisture products against in-situ observations. Evaluation statistics are the Pearson correlation, RMSE, and the KGE score. Panel (a) shows the temporal evaluation analysis split between in-situ observations used in the merging scheme random forest model (training sites, Table S1) and independent observations (the SMAP core calibration/validation sites). The in-situ observations were compared with the respective soil moisture product when data was simultaneously available for all four products, where \( n \) is the number of observational sites evaluated. To remove the influence of frozen soils, observations are masked when the HydroBlocks soil temperature is below 4 °C. Panel (b) shows the temporal statistics of the SMAP-HB product distributed in space and their respective improvement over those for the base products. The first row shows the correlation, RMSE, and KGE for SMAP-HB for all the sites. The following rows show the difference in the evaluation statistics between SMAP-HB and the base products. Blue colors indicate higher SMAP-HB performance. Inset histograms show the median and median absolute deviation values.
We merged the 30-m resolution brightness temperature from HydroBlocks-RTM with the 9-km resolution SMAP L3E observed brightness temperature. To do this, we used a spatial cluster-based merging scheme, introduced in Vergopolan et al.37. This merging scheme is implemented such that, in a given time step, the fine-scale merged brightness temperature $T_{BH}^+$ can be derived according to the state update equation:

$$T_{BH}^+ = T_{BH}^- + K \left( T_{SMAP}^{anom} - HT_{BH}^{anom} \right) \ast w_{short} + bias \ast w_{long}$$

Where $T_{SMAP}$ is the SMAP brightness temperature observation resampled to 9-km (SMAP L3E product), $T_{BH}^-$ is the cluster-space HydroBlocks-RTM brightness temperature, and the $anom$ subscript refers to the anomalies of each product. $T_{BH}^+$, $T_{BH}^-$, and $T_{BH}^{anom}$ have dimensions $nc \times 1$, where $nc$ is the total number of clusters in the domain. $T_{SMAP}^{anom}$ dimensions $ns \times 1$, where $ns$ is the total number of SMAP grids in the domain. $H$ is the observation operator that maps HydroBlocks-RTM brightness temperature anomalies ($T_{BH}^{anom}$) from the cluster space to the SMAP grid space. $H$ has dimensions $ns \times nc$, and it uses a Gaussian-shaped weighted area to account for the relative contribution of each cluster to each SMAP grid (Fig. 4).

The difference between $T_{SMAP}^{anom}$ and $HT_{BH}^{anom}$ accounts for the short-term (instantaneous) SMAP increments. The $bias$ term accounts for the systematic seasonal differences between $T_{SMAP}$ and $HT_{BH}$ and it was calculated using a 4-month moving window average. $w_{short}$ and $w_{long}$ are static parameters ranging between 0–1, and they are applied to control the contribution of SMAP anomalies and bias depending on how SMAP adds value to the merging scheme. As described in the sequence, to identify the added value of SMAP brightness temperature, we used a machine learning data-driven approach to extract relationships from in-situ observations, landscape characteristics, and SMAP ancillary data. The contribution of SMAP anomalies are also weighted by $K$, which represents the relative magnitude of the model and observation uncertainties.
Also operates in the cluster space and it has dimensions \( nc \times ns \). \( R \) is the observation error covariance matrix, and \( P \) is the model error covariance matrix. \( R \) has dimensions \( ns \times ns \), with the diagonal elements set to the SMAP radiometer uncertainty of 1.32 K^2 \( 49 \), and the off-diagonal set to zero—assuming the SMAP observation errors are uncorrelated with each other. For the model error covariance, we assume cluster pairs belonging to the same SMAP grid have correlated errors; otherwise, the errors are assumed to be uncorrelated. Thus, \( P \) has dimensions \( nc \times nc \), with the entries of correlated cluster pairs set to the HydroBlocks brightness temperature uncertainty of 52 K^2 \( 37 \) and the entries of uncorrelated cluster pairs set to zero.

With the merged brightness temperature estimates, we deployed the inverse HydroBlocks-RTM model to retrieve the 30-m (merged) satellite-based soil moisture estimates at each time-step independently. This spatial cluster-based merging scheme allows for efficiently combining regular-grid observational data into the cluster-space using matrices with \( nc \) dimension of ~300–500 elements instead of a fully distributed setup that would require ~90,000 elements (of 30-m grid cells) for merging data over the same 9-km grid.

Quantifying the added value of SMAP. Models and satellites have variable accuracy across the landscape, and these differences are reflected in the accuracy of merged products. Thus, identifying where the satellite data adds value and by how much is critical to improving the estimates. Here, we map the added value of SMAP brightness temperatures that would result in merged soil moisture with the highest accuracy. To this aim, we quantified the added value of SMAP based on how SMAP seasonal mean (4-month moving window) and anomalies (instantaneous differences with respect to the seasonal mean) improve soil moisture estimates with respect to the HydroBlocks model.

This approach relies on identifying the \( w_{short} \) and \( w_{long} \) parameters in Eq. 1 that result in merged soil moisture with the highest KGE score (defined in the Technical Validation section). Since these parameters control the contribution of SMAP to the merged brightness temperature, the higher the parameter values, the more SMAP contributes. When the parameters are close to zero, SMAP adds limited value with respect to the model. To quantify

\[
K = PH^T (HPH^T + R)^{-1}
\]
the $w_{\text{short}}$ and $w_{\text{long}}$ parameters, we used 958 in-situ soil moisture observations distributed across the CONUS (training sites in Table S1). We identified the added value at each site by testing all possible combinations of $w_{\text{short}}$ and $w_{\text{long}}$ (each parameter varied a 0.01 increment between 0–1), and we selected the pair that resulted in merged soil moisture with the highest KGE score. In this way, we compiled an observation-based training sample of $w_{\text{short}}$ and $w_{\text{long}}$, shown in Fig. 5a. Subsequently, we used this sample to train a random forest model (RF) to predict the added value of SMAP based on the relationship learned from physiographic and SMAP ancillary data predictors, listed in Table S2. For model training, the value of each RF predictor was defined at the colocated location of each observation with respect to the predictor grid cell. All the predictors were normalized based on their maximum and minimum values. For model prediction, the value of each RF predictor was defined as the predictor spatial mean at each cluster, with each predictor normalized based on the training set maximum and minimum values. In this way, after the RF is trained, it enables the prediction of the added value of SMAP seasonal mean and anomalies at each cluster, instead of every 30-m grid cell, while still yielding an effective 30-m spatial resolution.

This approach was applied to predict the added value of SMAP seasonal mean and anomalies across the CONUS (Fig. 5b). The seasonal mean represents the overall wet and dry biases of soil moisture, and the anomalies represent instantaneous contributions, such as from rainfall, irrigation, and flooding. Fig. 5b show how the SMAP seasonal mean adds more value in the Northern Great Plains, in the dry and heavily irrigated Southwest and California Central Valley, and in the wet and sandy soils of the Mississippi floodplains, correcting for the model bias. Short-term contributions (anomalies) tend to be more relevant across the irrigated Great Plains and in the sandier soil conditions of the West Coast and the Atlantic Coastal Plain, where SMAP can capture the timing of wetting events better than model-only estimates. This implies that at these locations SMAP contributes significantly to improve deficiencies in the precipitation data or in the way the model translates precipitation into soil moisture. However, SMAP anomalies provide a limited contribution in the northeast US, the Rocky Mountains, and the Appalachian Mountains, which could be attributable to the confounding effects of complex terrain and dense vegetation on the satellite retrievals, but also due to SMAP’s limited quality control in snow-dominated regions. The added value of SMAP seasonal mean and anomalies was applied to parameterize the SMAP-HB merging scheme in Eq. 1 via the $w_{\text{short}}$ and $w_{\text{long}}$ Parameters. This observation-driven parameterization enabled the merging scheme to benefit from the information contained in in-situ observations and physical landscape characteristics without solely relying on covariance errors.

**Data Records**

The SMAP-HydroBlocks surface soil moisture dataset at 30-m 6-h resolution (2015–2019) comprises a 22 TB dataset (with maximum compression). Due to the storage limitation of online repositories, we provide the raw data at the HRU level (time, hru) compressed to 33.8 GB. A python code and instructions for post-processing the data into geographic coordinates (time, latitude, longitude) is provided at GitHub (https://github.com/NoemiVergopolan/SMAP-HydroBlocks_postprocessing). An aggregated version at 1-km 6-h resolution already post-processed into geographic coordinates (time, latitude, longitude) is also made available comprising in 31.5 GB of data. Data are available for download from the Zenodo repository (https://doi.org/10.5281/zenodo.5206725). Different subsets of the data can also be made available upon request from the primary author. Please provide details on the intended and desired spatial and temporal resolution, domain, and period of interest in your request. Data will be provided via Google Drive shared link. The data are provided in self-describing netCDF-4 format (https://www.unidata.ucar.edu/software/netcdf/), and referenced to the World Geodetic Reference System 1984 (WGS 84) ellipsoid. The netCDF-4 files can be viewed, edited, and analyzed using most Geographic Information Systems (GIS) software packages, including ArcGIS, QGIS, and GRASS. As an illustration example, a 30-m map of the SMAP-HB annual and long-term climatology can be viewed through an interactive web interface at https://waternai.eai/smaphb.

**Technical Validation**

We quantified the spatial and temporal accuracy of the SMAP-HB 30-m soil moisture using observations from in-situ sensors at 1,191 sites. We compared it with the performance of the HydroBlocks and the SMAP L3E products (representing the baseline products), and the state-of-the-art SMAP L4 data assimilation product. Our evaluation used mean daily in-situ observations at the soil moisture products’ colocated grid cell only at the time steps in which all soil moisture products were simultaneously available. To remove the impact of frozen soils in the evaluation, we masked the soil moisture estimates when the HydroBlocks soil temperature was below 4 degrees Celsius.

In this way, after the RF is trained, it enables the prediction of the added value of SMAP seasonal mean and anomalies at each cluster, instead of every 30-m grid cell, while still yielding an effective 30-m spatial resolution.

This approach was applied to predict the added value of SMAP seasonal mean and anomalies across the CONUS (Fig. 5b). The seasonal mean represents the overall wet and dry biases of soil moisture, and the anomalies represent instantaneous contributions, such as from rainfall, irrigation, and flooding. Fig. 5b show how the SMAP seasonal mean adds more value in the Northern Great Plains, in the dry and heavily irrigated Southwest and California Central Valley, and in the wet and sandy soils of the Mississippi floodplains, correcting for the model bias. Short-term contributions (anomalies) tend to be more relevant across the irrigated Great Plains and in the sandier soil conditions of the West Coast and the Atlantic Coastal Plain, where SMAP can capture the timing of wetting events better than model-only estimates. This implies that at these locations SMAP contributes significantly to improve deficiencies in the precipitation data or in the way the model translates precipitation into soil moisture. However, SMAP anomalies provide a limited contribution in the northeast US, the Rocky Mountains, and the Appalachian Mountains, which could be attributable to the confounding effects of complex terrain and dense vegetation on the satellite retrievals, but also due to SMAP’s limited quality control in snow-dominated regions. The added value of SMAP seasonal mean and anomalies was applied to parameterize the SMAP-HB merging scheme in Eq. 1 via the $w_{\text{short}}$ and $w_{\text{long}}$ Parameters. This observation-driven parameterization enabled the merging scheme to benefit from the information contained in in-situ observations and physical landscape characteristics without solely relying on covariance errors.

**Data Records**

The SMAP-HydroBlocks surface soil moisture dataset at 30-m 6-h resolution (2015–2019) comprises a 22 TB dataset (with maximum compression). Due to the storage limitation of online repositories, we provide the raw data at the HRU level (time, hru) compressed to 33.8 GB. A python code and instructions for post-processing the data into geographic coordinates (time, latitude, longitude) is provided at GitHub (https://github.com/NoemiVergopolan/SMAP-HydroBlocks_postprocessing). An aggregated version at 1-km 6-h resolution already post-processed into geographic coordinates (time, latitude, longitude) is also made available comprising in 31.5 GB of data. Data are available for download from the Zenodo repository (https://doi.org/10.5281/zenodo.5206725). Different subsets of the data can also be made available upon request from the primary author. Please provide details on the intended and desired spatial and temporal resolution, domain, and period of interest in your request. Data will be provided via Google Drive shared link. The data are provided in self-describing netCDF-4 format (https://www.unidata.ucar.edu/software/netcdf/), and referenced to the World Geodetic Reference System 1984 (WGS 84) ellipsoid. The netCDF-4 files can be viewed, edited, and analyzed using most Geographic Information Systems (GIS) software packages, including ArcGIS, QGIS, and GRASS. As an illustration example, a 30-m map of the SMAP-HB annual and long-term climatology can be viewed through an interactive web interface at https://waternai.eai/smaphb.

**Technical Validation**

We quantified the spatial and temporal accuracy of the SMAP-HB 30-m soil moisture using observations from in-situ sensors at 1,191 sites. We compared it with the performance of the HydroBlocks and the SMAP L3E products (representing the baseline products), and the state-of-the-art SMAP L4 data assimilation product. Our evaluation used mean daily in-situ observations at the soil moisture products’ colocated grid cell only at the time steps in which all soil moisture products were simultaneously available. To remove the impact of frozen soils in the evaluation, we masked the soil moisture estimates when the HydroBlocks soil temperature was below 4 degrees Celsius.

The temporal evaluation was split between 958 training sites (used to parameterize our merging scheme via machine learning) and 233 independent testing sites (SMAP core calibration/validation sites; see Table S1). Training sites were selected such that no validation sites were within a 25 km radius from testing sites. We evaluate the soil moisture performance in terms of the temporal Pearson correlation, the Root Mean Squared Error (RMSE), and the Kling-Gupta Efficiency (KGE) score. The KGE score combines the linear Pearson correlation ($\rho$), the bias ratio ($\beta$), and the variability ratio ($\gamma$):

$$KGE = 1 - \sqrt{(\rho - 1)^2 + (\beta - 1)^2 + (\gamma - 1)^2}$$

where $\mu$ and $\sigma$ are the temporal mean and standard deviation of the soil moisture products (prod) and the observations (obs).

Fig. 2a presents the temporal evaluation results. Overall, SMAP-HB has the best temporal statistics, with RMSE values of 0.07 m$^3$/m$^3$ for both the training and testing sites, and Kling-Gupta Efficiency (KGE) scores
of 0.53 and 0.48 for the training and testing sites, respectively. While SMAP-HB median temporal correlations were 0.71 and 0.77 at the training and testing sites, respectively, the values for SMAP L4 were 0.73 and 0.74. SMAP-L4 generally performed better than SMAP-HB in terms of temporal correlation at mountainous and snow-dominated sites (e.g., at SNOTEL sites; see Fig. S1). The higher skill of SMAP L4 at these sites could be associated with the benefit of assimilation of in-situ precipitation observations into the meteorological forcings of the Catchment land surface model.14

To also quantify the added value of our merging scheme at the point level, we evaluated the temporal statistics spatially (Fig. 2b). SMAP-HB correlation, bias, and RMSE values across the CONUS are spatially homogeneous, with an overall improvement with respect to the baseline products (SMAP L3E and HydroBlocks). SMAP-HB showed a median improvement of 0.03 in temporal correlation with respect to the SMAP L3E product. However, the largest gains are observed in the KGE score, with a median improvement of 0.12 in comparison to SMAP L3. This KGE improvement consolidates overall improvements in temporal correlation, bias ratio, and variation ratio. Figs. S1 and S2 present additional temporal evaluation statistics stratified per soil moisture network, soil type, elevation, vegetation type, among others.

To assess the soil moisture products’ performance in representing spatial dynamics, the spatial correlation was calculated for each day by comparing the daily soil moisture products collocated grid-cell and daily in-situ observations over CONUS, New York Mesonet, and Oklahoma Mesonet. Aiming for statistical significance, the spatial correlation was only calculated when at least 60 in-situ observations and soil moisture products were available simultaneously at a given time step. As such, the spatial correlation aims to quantify at each time step to what extent are the soil moisture products representative of the soil moisture spatial variability. Our results show in Fig. 3 that HydroBlocks and SMAP-HB presented the highest spatial correlation across the CONUS, the New York Mesonet, and the Oklahoma Mesonet. SMAP-HB spatial correlation was 0.66 over CONUS, 0.42 over the New York Mesonet, and 0.54 over the Oklahoma Mesonet. The largest SMAP-HB improvement is observed at the NY-Mesonet, where HydroBlocks spatial correlation was 0.32 and SMAP L4 was 0.23. However, the caveat of this spatial correlation analysis is that it includes the training in-situ observations (also used to parameterize the merging scheme).

Usage Notes
Given its spatial detail, the SMAP-HB dataset will be useful for solving many physical processes and application at spatial scales that so far have been unresolved. These applications include mapping and understanding crop irrigation demands,4,6 farmer decision making and planting dates65, drought impacts62,72; and mapping of antecedent soil moisture conditions can help estimate the susceptibility to wildfires72, landslides6,10, flooding, and waterlogging conditions61,12. Detailed soil moisture information can aid and improve the quantification of biogeochemical cycles in wetlands and riparian zones66, as well as better inform the environmental conditions that facilitate epidemic outbreaks of, for example, West Nile virus67, malaria68, and locust69. SMAP-HB’s improved characterization of soil moisture spatial variability can inform the parameterization of atmospheric convection models69 directly supporting climate and weather predictions71. However, uncertainties still remain and some caveats should be considered:

- SMAP-HB estimates the volumetric surface soil moisture content of the top 5-cm of the soil based on SMAP-observed brightness temperature. As such, SMAP-HB retrievals are only available when and where SMAP has non-flagged brightness temperature observations.
- SMAP-HB showed lower temporal correlation at sites of high elevation (Fig. S2b), such as sites belonging to the SNOTEL network (Fig. S1). This could be due to (i) the confounding effects of topographic relief on the upwelling microwave brightness temperature observed by the radiometer; (ii) the likely more frequent presence of frozen or snow-covered soils that were not captured by quality control, but can affect both the in-situ measurements and the satellite retrievals; and (iii) the lower quality of the precipitation data (due to terrain blockage of radar beams, a lower rain gauge density, and a relatively high spatial heterogeneity in precipitation). In fact, Beck et al.72 demonstrated that the precipitation forcing can play a large role in driving the temporal correlation accuracy of the soil moisture products that were derived from merging approaches that include physically-based modeling.
- Although not quantified due to limited in-situ observation coverage, we expected high uncertainties near urban areas, given limitations in characterizing hydrological processes in urban and human-managed settings, as well as limited model capability in representing drainage networks. High uncertainties and NoData is expected in coastal areas and near large water bodies due to microwave signal contamination.
- With respect to irrigation, due to the large footprint of the SMAP sensor, SMAP-HB is limited to only capturing large-scale irrigation signals. To capture the impact of local-scale patchy irrigation, future work will include the assimilation of thermal sensors and an irrigation module into the HydroBlocks model. Such improvements on data and methods would benefit not only the spatial and temporal accuracy but may also enhance capabilities for local-scale applications.

Code availability
Source code for the HydroBlocks land surface model is available at https://github.com/chaneyn/HydroBlocks. The Random Forest model used to parameterize the merging scheme was implemented using the RandomForestRegressor class of the scikit-learn Python module. While not written as a portable library or toolset, code is available upon request.

Received: 4 March 2021; Accepted: 19 July 2021;
Published online: 11 October 2021
39. Reichle, R. et al. Smap l4 global 3-hourly 9 km ease-grid surface and root zone soil moisture analysis update, version 5 (2020).
40. O’Neill, P., Bindlish, R., Chan, S., Njoku, E. & Jackson, T. Algorithm theoretical basis document. level 2 & 3 soil moisture (passive) data products (2018).
41. Kumar, S. V., Dirmeyer, P. A., Peters-Lidard, C. D., Bindlish, R. & Bolten, J. Information theoretic evaluation of satellite soil moisture retrievals. Remote Sensing of Environment 204, 392–400, https://doi.org/10.1016/j.rse.2017.10.016 (2018).
42. Koster, R. D., Suarez, M. J., Ducharne, A., Stieglitz, M. & Kumar, P. A catchment-based approach to modeling land surface processes in a general circulation model: 1. model structure. Journal of Geophysical Research: Atmospheres 105, 24809–24822, https://doi.org/10.1029/2000jd900327 (2000).
43. Chaney, N. W., Metcalfe, P. & Wood, E. F. Hydroblocks: a field-scale resolving land surface model for application over continental extents. Hydrological Processes 30, 3543–3559, https://doi.org/10.1002/hyp.10891 (2016).
44. Chaney, N. W., Torres-Rojas, L., Vergopolan, N. & Fisher, C. K. Two-way coupling between the sub-grid land surface and river networks in earth system models. Geoscientific Model Development https://doi.org/10.5194/gmd-2020-291 (2020).
45. Chaney, N. W. et al. Harnessing big data to rethink land heterogeneity in earth system models. Hydrology and Earth System Sciences 22, 3311–3330, https://doi.org/10.5194/hess-22-3311-2018 (2018).
46. Danielson, J. J. & Gesch, D. B. Global multi-resolution terrain elevation data 2010 (GMTED2010) (US Department of the Interior, US Geological Survey, 2011).
47. Homer, C. G. et al. Completion of the 2011 national land cover dataset for the conterminous united states—representing a decade of land cover change information. Photogrammetric Engineering and Remote Sensing 81, 345–354 (2011).
48. Chaney, N. W. et al. Polarimetric soil properties: 30-m probabilistic maps of soil properties over the contiguous united states. Water Resources Research 55, 2916–2938, https://doi.org/10.1002/2018wr022797 (2019).
49. Piepmeier, J. R. et al. Smap l-band microwave radiometer: Instrument design and first year on orbit. IEEE Transactions on Geoscience and Remote Sensing 55, 1954–1966, https://doi.org/10.1109/igrs.2016.2631978 (2017).
50. Kraatz, S. et al. Evaluation of smap freeze/thaw retrieval accuracy at core validation sites in the contiguous united states. Remote Sensing 10, 1483, https://doi.org/10.3390/rs10091483 (2018).
51. Vergopolan, N. et al. Smap-hydroblocks: Hyper-resolution satellite-based soil moisture over the continental united states. Zenodo https://doi.org/10.5281/zenodo.5206725 (2021).
52. Bell, J. E. et al. 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).
53. Brotzge, J. A. et al. A technical overview of the new york state mesonet standard network. Journal of Atmospheric and Oceanic Technology 37, 1827–1845, https://doi.org/10.1175/jtech-d-19-0220.1 (2020).
54. McPherson, R. A. et al. Statewide monitoring of the mesoscale environment: A technical update on the oklahoma mesonet. Journal of Atmospheric and Oceanic Technology 24, 301–321, https://doi.org/10.1175/jtech1976.1 (2007).
55. Larson, K. M. et al. Use of gps receivers as a soil moisture network for water cycle studies. Geophysical Research Letters 35, https://doi.org/10.1029/2008gl036013 (2008).
56. Keefer, T. O., Moran, M. S. & Paige, G. B. Long-term meteorological and soil hydrology database, walnut gulch experimental watershed, arizona, united states. Water Resources Research 44, https://doi.org/10.1029/2006wr005702 (2008).
57. Bosch, D. D. et al. Little river experimental watershed database. Water Resources Research 43, https://doi.org/10.1029/2006wr005844 (2007).
58. Cosh, M. H., Jackson, T. J., Starks, P. & Heathman, G. Temporal stability of surface soil moisture in the little washita river watershed and its applications in satellite soil moisture product validation. Journal of Hydrology 323, 168–177, https://doi.org/10.1016/j.jhydrol.2005.08.020 (2006).
59. Seyfried, M. S., Murdoch, M. D., Hanson, C. L., Flerchinger, G. N. & Van Vactor, S. Long-term soil water content database, reynolds creek experimental watershed, idaho, united states. Water Resources Research 37, 2847–2851, https://doi.org/10.1029/2001wr000419 (2001).
60. Coopersmith, E. J., Cosh, M. H., Petersen, W. A., Prueger, J. & Niemeier, J. J. Soil moisture model calibration and validation: An ars watershed on the south fork iowa river. Journal of Hydrometeorology 16, 1087–1101, https://doi.org/10.1175/jhm-d-14-0145.1 (2015).
61. Collander, A. et al. Validation of smap surface soil moisture products with core validation sites. Remote Sensing of Environment 191, 215–231, https://doi.org/10.1016/j.rse.2017.01.021 (2017).
62. Ma, S., Baldocchi, D., Wolf, S. & Verfaillie, J. Slow ecosystem responses conditionally regulate annual carbon balance over 15 years in californian oak-grass savanna. Agricultural and Forest Meteorology 228–229, 252–264, https://doi.org/10.1016/agfmetorm.2016.07.016 (2016).
63. Dorigo, W. A. et al. The international soil moisture network: a data hosting facility for global in situ soil moisture measurements. Hydrology and Earth System Sciences 15, 1675–1698, https://doi.org/10.5194/hess-15-1675-2011 (2011).
64. Reichle, R. H. et al. The contributions of gauge-based precipitation and smap brightness temperature observations to the skill of the smap level-4 soil moisture product. Journal of Hydrometeorology 22, 405–424, https://doi.org/10.1175/jhm-d-20-0217.1 (2021).
65. Waldman, K. B. et al. Cognitive biases about climate variability in smallholder farming systems in zambia. Weather, Climate, and Society 11, 369–383, https://doi.org/10.1175/wcas-d-18-0050.1 (2019).
66. Dabrowska-Zielinska, K. et al. Assessment of carbon flux and soil moisture in wetlands applying sentinel-1 data. Remote Sensing 8, 756, https://doi.org/10.3390/rs8090756 (2016).
67. Keydel, A. C. et al. Seasonal temperatures and hydrological conditions improve the prediction of west nile virus infection rates in culex mosquitoes and human case counts in new york and connecticut. PLOS ONE 14, e0217854, https://doi.org/10.1371/journal. pone.0217854 (2019).
68. Bomblices, A., Duchemin, J.-B. & Eltahir, E. A. A mechanistic approach for accurate simulation of village scale malaria transmission. Malaria Journal 8, https://doi.org/10.1186/1475-2875-8-223 (2009).
69. Gomes, D., Salvador, P., Sana, J. & Casanova, J. L. Modelling desert locust pregnancy using 32-year soil moisture data on a large-scale. Ecological Indicators 117, 106655, https://doi.org/10.1016/j.ecolind.2020.106655 (2020).
70. Tawfik, A. B., Lawrence, D. M. & Dirmeyer, P. A. Representing subgrid convective initiation in the community earth system model. Journal of Advances in Modeling Earth Systems 9, 1740–1758, https://doi.org/10.1175/jamessd-10-00866 (2017).
71. Dirmeyer, P. A. & Halder, S. Sensitivity of numerical weather forecasts to initial soil moisture variations in cfsv2. Weather and Forecasting 31, 1973–1983, https://doi.org/10.1175/waf-d-16-0049.1 (2016).
72. Beck, H. E. et al. Evaluation of 18 satellite- and model-based soil moisture products using in situ measurements from 826 sensors. Hydrology and Earth System Sciences 25, 17–40, https://doi.org/10.5194/hess-25-17-2021 (2021).

Acknowledgements

We thank the researchers and funding agencies that maintain the open-source in situ soil moisture networks, and the researchers from the SMAP Science Team that shared data from the SMAP core calibration/validation sites. New York State Mesonet data access was made possible through funding from NOAA grant NA19OAR4310368. This work was supported by NASA Soil Moisture Cal/Val Activities as a SMAP Mission Science Team Member.
(grant number NNX14AH92G); by the “Modernizing Observation Operator and Error Assessment for Assimilating In-situ and Remotely Sensed Snow/Soil Moisture Measurements into NWM” project from NOAA (grant number NA19OAR4590199); and the High Meadows Environmental Institute at Princeton University through the Mary and Randall Hack ’69 Research Fund Award.

**Author contributions**

N.V. and E.F.W. conceived and designed the research. N.V. performed the analysis and lead the paper writing. N.C., M.P. and J.S. provide insights on the methodology development. All co-authors provided critical feedback and contributed to the writing.

**Competing interests**

The authors declare no competing interests.

**Additional information**

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41597-021-01050-2.

**Correspondence** and requests for materials should be addressed to N.V.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

[Open Access] This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

The Creative Commons Public Domain Dedication waiver http://creativecommons.org/publicdomain/zero/1.0/ applies to the metadata files associated with this article.

© The Author(s) 2021