Continental scale downscaling of AWRA-L analysed soil moisture using random forest regression

Y. Yu [a], **L.J. Renzullo** [a] **and S. Tian** [a]

[a] *Fenner School of Environment & Society, The Australian National University, Australian Capital Territory*
Email: u6726739@anu.edu.au

Abstract: The Australian Water Resource Assessment Landscape (AWRA-L) model as used by the Bureau of Meteorology (BoM) provides daily continental scale soil moisture (SM) estimates (among other landscape water variables) at ~5-km resolution. At such a coarse scale these data cannot represent the high spatiotemporal variability of SM across heterogeneous land surfaces. Downscaling of coarse SM products based on machine learning (ML) has become increasingly popular due to its robust predictions and potential for large-scale applications. As a first step towards high-resolution daily Australia-wide SM estimation, a downscaling framework was developed to generate monthly SM with 500-m spatial resolution using analysed SM from AWRA-L and multisource geospatial predictors in random forest (RF) regression. Candidate predictors include digital elevation model (DEM), soil properties from the Australian soil and landscape grids, and several retrievals from the MODerate-resolution Imaging Spectroradiometer (MODIS). Ten experiments were conducted to decide the best combination of predictors. In the chosen model, DEM and available water capacity (AWC) were consistently identified as the most important predictors based on the ranking of variable importance.

The downscaled SM shows greatly enhanced spatial details at the local scale while maintaining consistent patterns with AWRA-L analysis at the continental scale. Validations against in-situ measurement networks using Pearson correlation coefficient (R) show that there is very little difference in the performance between the downscaled and AWRA-L SM. Average R values for the downscaled SM against CosmOz, OzFlux and OzNet were 0.87, 0.68 and 0.75, respectively, while the original AWRA-L SM average R were 0.86, 0.68 and 0.76, respectively. Furthermore, the time series comparison based on a wetness unit shows that the downscaled SM can well catch up the fluctuations of in-situ SM. In general, this study explores the potential of ML approach for the SM downscaling applications at the continental scale. It could be a promising direction to exploit the modelling capability of integrating multisource geospatial data including satellite retrievals, land surface models (LSM) and interpolated ground observation data. Future directions should concentrate on integrating this approach into an operational framework with a daily frequency. Exploration of the relationships between SM and auxiliaries under difference scales would be essential, in order to better understand the dominant physical controls on spatial variability of SM.

Keywords: *Soil moisture, Australian Water Resource Assessment Landscape (AWRA-L) model, downscaling, machine learning, random forest*

1. INTRODUCTION

Surface soil moisture (SM) usually refers to a measure of water in the uppermost part of the soil profile (Romano, 2014), which is known to be crucial in the partitioning of precipitation into runoff, evaporation and infiltration by affecting the distribution of water and radiation at the land-atmosphere interface (Kolassa et al., 2017; Maggioni and Houser, 2017). Satellite-based SM retrievals and land surface model (LSM) data are able to provide spatiotemporally continuous SM data at a global or continental scale, but their applications are usually constrained by the coarse spatial resolution in the order of 10 km (Djamai et al., 2016). Approaches therefore have been developed to downscale SM products by accounting the impact of different environmental variables (Peng et al., 2017), aiming at providing better spatial details while maintaining good correlation with in-situ measurements.

Downscaling based on machine learning (ML) approaches has become increasingly popular due to its robust predictions and its potential for large-scale applications. There exist several studies that have applied ML in the downscaling of SM at regional scales, most of which have indicated random forest (RF) regression as an outstanding performance against other ML approaches (Ahmad et al., 2010; Im et al., 2016; Long et al., 2019; Mao et al., 2019). However, few of them focused on a continental scale and most concentrated on the downscaling of satellite SM products. Furthermore, most existing research utilised satellite SM coarser than 25-km resolution as the model response, which means the training set was usually far too small (only a few thousand training samples or even less) for a ML model.

In this study, the Australian Water Resource Assessment Landscape model (AWRA-L) analysed SM is utilised as the model response for the downscaling. Specifically we use the analysed top soil layer SM generated by Tian et al. (2021) which assimilated two satellite SM products simultaneously into the operationalised AWRA-L model. Compared to satellite SM products, it fills in the spatial gaps caused by a 3-day revisit cycle of sensors, thus can provide continuous spatiotemporal information. However, its application potential is still constrained by its relatively coarse spatial resolution (5-km). Due to the lack of fine-scale forcing and other auxiliary data, the AWRA-L analysis cannot run at a finer scale with confidence. As a first step towards Australia-wide available water estimation, the application of downscaling technologies on AWRA-L analysis would help well demonstrate the high spatiotemporal variability of SM across heterogeneous land surfaces. The downscaling based on AWRA-L analysis (about 260,000 grid cells for Australia) can also meet the size requirement of the training set for a typical ML model.

The main objective of this study is to generate monthly SM with 500-m spatial resolution for Australia using 5-km AWRA-L SM analysis and multisource geospatial data including satellite retrievals (e.g., albedo, vegetation indices), topographic data and soil textures in ML approach. This study evaluates the downscaled SM by spatiotemporal comparisons with AWRA-L analysis, and validations against in-situ SM data from three individual sites. This study provides an approach of generating moderate-resolution SM at a continental scale, and is expected to derive insights into how downscaling can be incorporated as part of the AWRA-L operational system to produce fine-scale soil moisture estimates.

2. DATASETS

2.1. AWRA-L soil moisture analysis

The AWRA-L model is designed to support the monitoring and assessment of water resources and water accounting, which is currently being operationalised by the Bureau of Meteorology (BoM) (Frost et al., 2018). AWRA-L is a grid-based landscape hydrological model with a 5-km spatial resolution and daily temporal resolution. AWRA-L simulates SM at three layers (upper: 0–10 cm, lower: 10–100 cm, and deep: 1–6 m). In this study, we used the SM analysis for the upper soil layer after the assimilation of satellite SM retrievals from the Soil Moisture Active Passive (SMAP) and Soil Moisture and Ocean Salinity (SMOS) using a Kalman filter type sequential state updating process (Tian et al., 2021). The depth of AWRA-L upper layer is different from the general depth of the surface SM (5-cm). Pinnington et al. (2021) found the assimilation results of LSM are usually consistent while running with either a 10-cm top layer or a 5-cm top layer. The comparison between AWRA-L upper layer and other independent surface SM can be considered fair. The AWRA-L SM analysis shows improved agreement with in-situ SM measurements compared to the model open-loop simulations.

2.2. MODIS retrievals

Products from MODerate-resolution Imaging Spectroradiometer (MODIS) onboard Terra and Aqua satellites have been widely applied in monitoring the dynamics of landscape, hydrology and lower atmosphere. We chose a total of six MODIS products for this study including 500-m daily Albedo (MCD43A3), 500-m 8-day

evapotranspiration (ET; MOD16A2), 500-m 16-day enhanced vegetation indices (EVI; MOD13A1), 500-m 8-day leaf area index (LAI; MCD15A2H) and 1-km 8-day land surface temperature (LST; MOD11A2 and MYD11A2). We acquired the relevant data from corresponding MODIS collections. All of them can be freely accessed through the NASA's Earth Observing System Data and Information System (EOSDIS; https://search.earthdata.nasa.gov/search).

2.3. Topographic and soil texture data

We used the Smoothed Digital Elevation Model (DEM-S) data from the Geosciences Australia (https://www.ga.gov.au/). DEM-S excludes the influences of ground vegetation features and has been smoothed to reduce the impacts of noise. Its spatial resolution is 1-arcsecond (about 30-m).

Soil texture can describe the spatial variations of various soil attributes from different depths at a regional or continental scale. We used four soil attributes including available water capacity (AWC), clay, sand and silt from soil and landscape grid data (available at https://data.csiro.au/). The resolution is 3-arcsecond (about 90-m) and the depth of soil texture data chosen for this study is 0-5cm to be consistent with AWRA-L analysis.

2.4. In-situ SM

We used the in-situ measured data from three SM measurement networks across Australia to conduct validation and comparison with the modelling data. These SM monitoring networks include (1) the Australian Cosmic-Ray Neutron Soil Moisture Monitoring Network (CosmOz; https://cosmoz.csiro.au/) that uses fast neutrons to measure soil moisture over a 40-hectare (0.4-km^2) region for each site; (2) the Australian and New Zealand Flux Research and Monitoring (OzFlux; http://www.ozflux.org.au/) that uses flux towers and; (3) the OzNet Hydrological Monitoring Network (OzNet; http://www.oznet.org.au/). Among them, CosmOz and OzFlux are national scale networks, while OzNet is a regional scale network located in the Murrumbidgee Catchment. Figure 1 gives the distribution of the networks used in this study. The depth of the in-situ data used for comparison is 0-5cm.

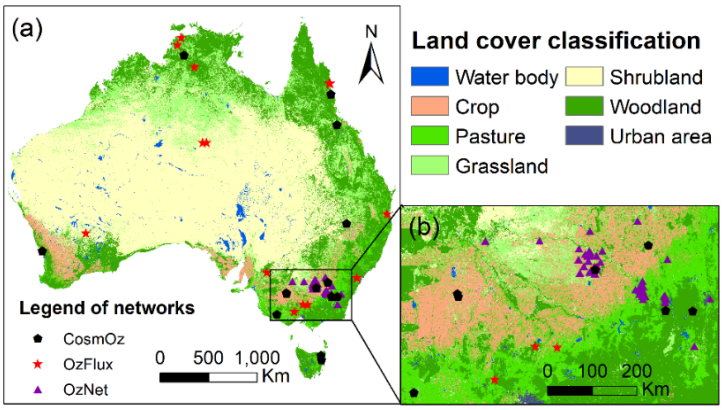

Figure 1. (a) The distribution of three in-situ SM networks across Australia; (b) the locations of in-situ sites distributed in the Murrumbidgee Catchment, southeast Australia.

3. METHODOLOGY

3.1. Random forest regression

The Random Forest (RF), proposed by Breiman (2001), is an ensemble machine learning approach based on decision trees. It can be used for both classification and regression tasks and has been widely applied in various fields due to its robust predictions and strength in reducing overfitting. The principle of RF is to build a series of decision trees based on the bootstrapping sampling of the training set, and make the final prediction by either choosing the class selected by most trees (classification) or averaging the individual predictions of all trees (regression). There are some key hyperparameters in the RF modelling, including the number of variables selected at each split (mtry) and the number of trees (ntree). In this study, we chose a mtry of 7 and a ntree of 800 based on the results of model tuning. RF is also able to estimate the contribution of different predictors using two variable importance function, including permutation importance and Gini index importance, both of which are presented in percentage (%). In this study, we chose the permutation importance to rank the contribution of predictors. The percent of the permutation importance means that how much the out-of-bag error of the model would increase when that predictor is randomly permuted.

3.2. Downscaling framework

The idea behind ML downscaling is to fit a regression model between SM and predictors (resampled to the same resolution with SM) at coarse resolution, and predict downscaled SM using fine-resolution predictors.

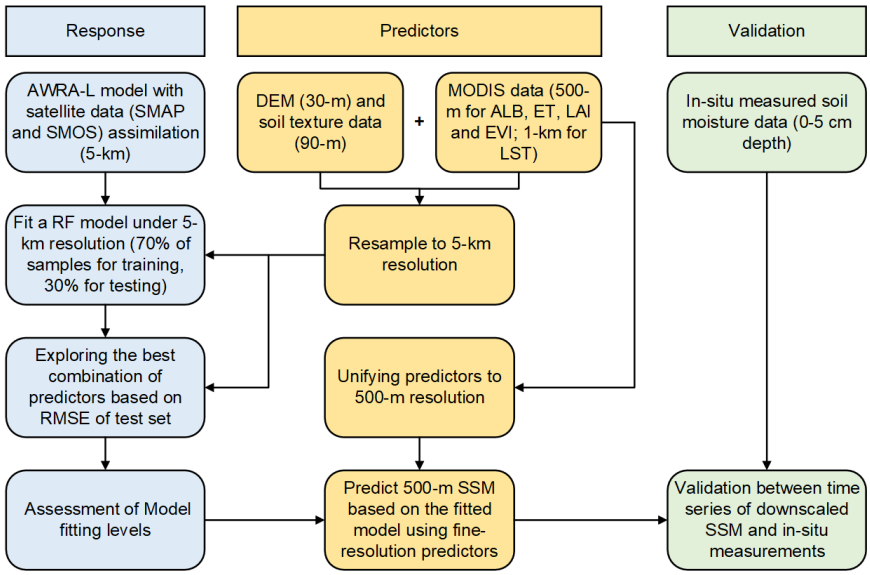

Figure 2. The flow chart of the downscaling framework.

The detailed process is as follows: Firstly, we composited the monthly mean values of MODIS and SM data. This is to avoid the potential of large-scale data missing in MODIS products and unify the temporal resolution of different products. We then resampled all the predictors to the same resolution with AWRA-L analysis data (5-km) using the nearest neighbour (NGB) function, based on which we built the database for RF model for every month. For each database, we randomly chose 70% of the samples as the training data, and the remaining 30% of the samples were used as the testing data. The same seed was used for each model to make the randomised sampling process repeatable. We conducted ten experiments to decide a best combination of the predictors based on the root mean squared error (RMSE) of test set. The strategy was to start from five static predictors (i.e., DEM and four soil texture) and then incorporate new predictors accordingly, in order to observe when the RMSE of model can become stable. The order of incorporation is albedo, ET, LAI, LST Aqua daytime (LSTAD), LST Aqua nighttime (LSTAN), LST Terra daytime (LSTTD), LST Terra nighttime (LSTTN), EVI and NDVI. Finally, we built the RF regression between AWRA-L analysis and all the chosen predictors. Once the model training was completed, it predicted the monthly downscaled SM based on the 500-m rasters of different predictors, which were also resampled using NGB function. Among them, resampling the coarse LST (1-km) data to a fine resolution using NGB function cannot properly deliver the information at a 500-m resolution. The impacts of LST predictors are further measured based on the RF variable importance. Figure 2 shows the flow chart of the modelling process.

3.3. Statistical metrics

Two metrics were chosen to measure the model performance and validate the downscaled SM with in-situ measurements, including RMSE and Pearson correlation coefficient (R). The calculation of them is shown as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(\hat{x}_i - x_i)^2}{N}} \quad (1)$$

$$R = \frac{\sum(\hat{x}_i - \bar{\hat{x}})(x_i - \bar{x})}{\sqrt{\sum(\hat{x}_i - \bar{\hat{x}})^2 \sum(x_i - \bar{x})^2}} \quad (2)$$

where N represents the number of months in the study period; $\hat{x}_i$ and x_i represent modelled SM and reference SM, respectively, on the ith month; $\bar{\hat{x}}$ and $\bar{x}$ represent the mean values of the $\hat{x}$ and x, respectively.

4. RESULTS

The RSME of the different combination of predictors are shown in Figure 3 (a). The medium RMSE reduces first from around 0.43 to 0.33 mm with the incorporation of new predictors, and gradually levels off at around 0.33 mm. After the 9th experiment, despite the incorporation of a new predictor, RMSE does not reduce any more. We then chose the 9th experiment due to its best quartiles, with the combination of DEM, AWC, clay, silt, sand, albedo, ET, EVI, LAI, LSTAD, LSTAN, LSTTD and LSTTN. The ranking of variable importance of different predictors are shown in Figure 3 (b). DEM is always ranked as the most outstanding predictor in the regression with a medium value of about 53%. AWC is ranked second with a medium value of about 35%. The contribution of LAI, sand and EVI are ranked as lowest with medium values lower than 20%. Furthermore,

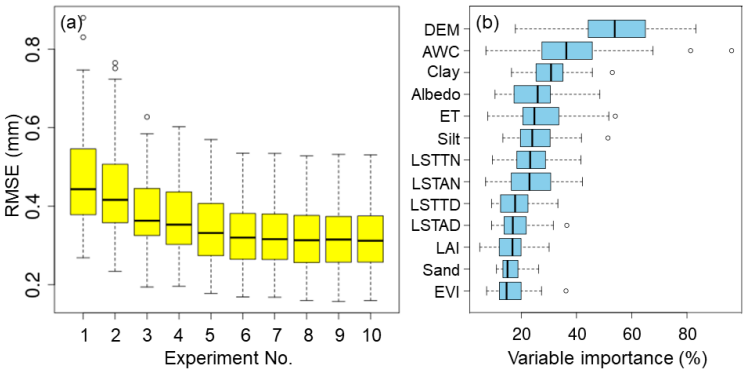

Figure 3. (a) The boxplot of the ten experiments to decide the best combination of predictors; (b) The boxplot of variable importance of different predictors for the chosen model.

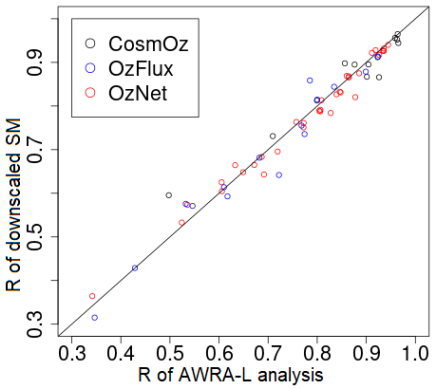

Figure 4. The comparison of R between AWRA-L analysis and downscaled SM against in-situ SM.

it should be noted that the importance of LST predictors are ranked between 7-10, revealing that their relatively lower contribution to the prediction. Then the adverse effects of resampling LST from 1-km to 500-m may be of little consequence.

The validation between AWRA-L analysis and downscaled SM against in-situ measurements are shown in Figure 4, indicating that the performance of AWRA-L analysis and downscaled SM are comparable. The average R of AWRA-L analysis are 0.86, 0.68 and 0.76 for CosmOz, OzFlux and OzNet, respectively; while the average R of downscaled SM are 0.87, 0.68 and 0.75, respectively. The 500-m downscaled SM maintains a same performance with 5-km AWRA-L analysis at most sites and does offer some improvements at a few sites, mostly from the two national-scale networks (i.e., CosmOz and OzFlux). Figure 5 gives six examples of the monthly time series comparisons between the AWRA-L analysis, downscaled SM and in-situ SM. All the datasets are using the monthly average value, while we also converted the SM unit to wetness (%) for both modelled and in-situ data due the difference in their original units. For these sites, the R between in-situ and AWRA-L analysis is ranging from 0.83 to 0.96, while R of downscaled SM is ranging from 0.84 to 0.96. The time series of downscaled SM match well with AWRA-L analysis, and are consistent with the seasonal fluctuations of in-situ data, especially perform well in Daly and Gnangara (Figure 5b and 5c).

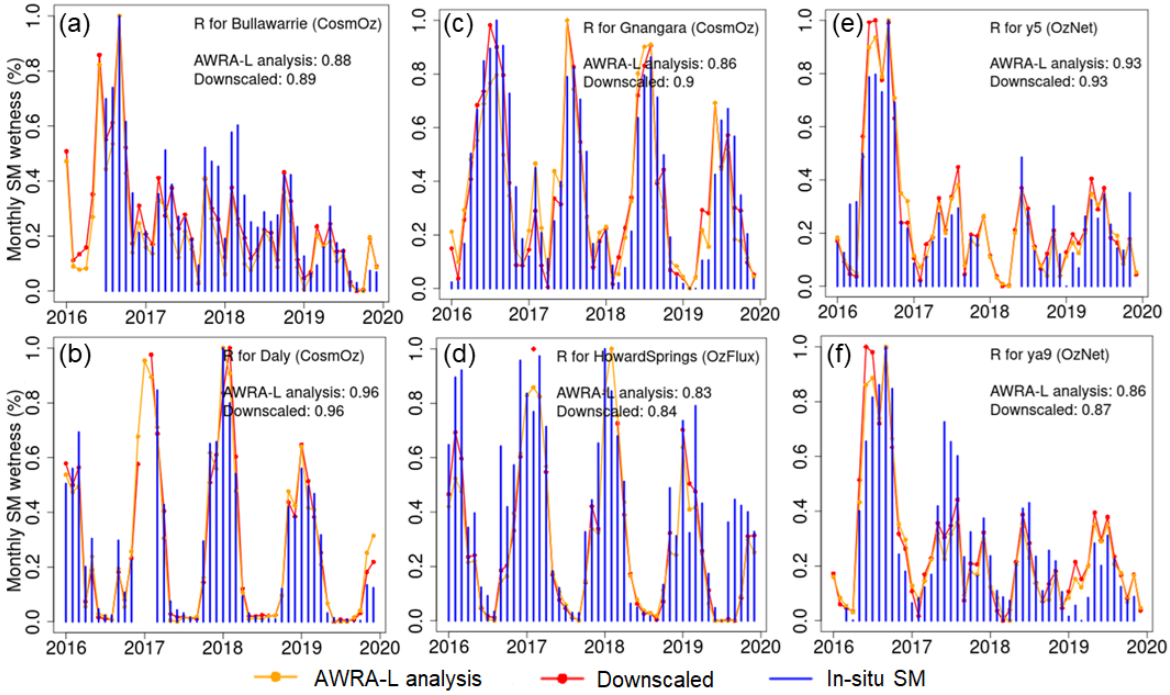

Figure 5. Six examples of the temporal comparison between monthly time series of AWRA-L analysis, downscaled SM and in-situ data from the network of (a-c) CosmOz; (d) OzFlux and; (e-f) OzNet.

Figure 6 shows an example of the continental scale spatial pattern of the downscaled SM in January 2016 and four zoom-in comparisons between downscaled SM and AWRA-L analysis from different locations. The continental scale SM has more spatial variabilities and higher values (mostly higher than 2-mm) at north, northeast and southeast; while the SM at central and western Australia shows relatively lower values (near 0-mm) and high homogeneity. At the local scale, the downscaled SM can provide significantly enhanced spatial details while maintaining consistent patterns with AWRA-L analysis. The downscaled one can well simulate both homogeneous groups and thin paths of SM at the 500-m scale.

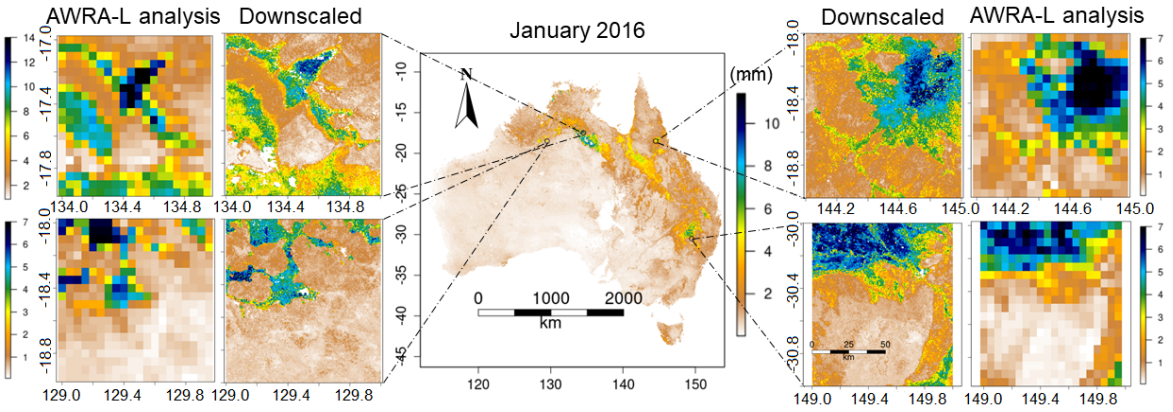

Figure 6. An example of the continental scale spatial pattern of the downscaled SM in January 2016 and four zoom-in comparisons between downscaled SM and AWRA-L analysis at different locations.

5. DISCUSSION AND CONCLUSION

In this research, we proposed a downscaling framework to generate 500-m resolution SM data across Australia with a monthly frequency. The framework builds a RF regression model using a 5-km resolution AWRA-L SM analysis and multiple geospatial predictors to make predictions at 500-m resolution.

Ten experiments based on different combinations of geospatial predictors were conducted. In the chosen model, based on RMSE, the contribution of different predictors was calculated using the variable importance function in the RF model. DEM was always ranked with the highest variable importance, followed by AWC. The validation with in-situ data show that the R between in-situ data and AWRA-L analysis are comparable with downscaled SM at most sites. The average R of AWRA-L analysis were 0.86, 0.68 and 0.76 for CosmOz, OzFlux and OzNet, respectively; while the average R of downscaled SM were 0.87, 0.68 and 0.75, respectively. Nevertheless, the downscaled SM showed better R in some sites (e.g., Gnangara, Robson, Calperum and Gingin) at the two national-scale networks. Moreover, in-situ data from CosmOz sites had the best agreement with both SM data among the three networks. This could be because CosmOz utilises fast neutrons generated from interactions between cosmic rays and the atmosphere and top soils to measure SM in a 0.4 km^2 area, which makes CosmOz advantageous in representativeness of a specific region than other networks.

Compared to AWRA-L analysis, the spatial details of downscaled SM at a local scale have been greatly enhanced. Given the downscaled SM can still maintain a same performance with AWRA-L analysis in the correlation against in-situ data, there could be several prospects for its potential. Firstly, it can guide the model development of AWRA-L at 500-m resolution or higher. The parameters of LSM are usually spatially varying and need to be calibrated. As the downscaled product can demonstrate a better representation of the spatial heterogeneity of SM, it can be applied for either parameter tuning of continental scale LSM or the improvement of surface sub-models that are spatially distributed differently. However, to apply the downscaled SM to a real-time analysis, a challenge would be the disaggregation of the downscaled SM from monthly to daily frequency. We used the monthly frequency in this study to unify the temporal resolution of different products, and avoid a large-scale missing area of predictors and overload of computing resources. In the future, some alternative strategies would be a direct application of the model at a daily frequency, or integrating this model into an operational model, which undoubtedly requires further explorations concentrated on spatial auxiliaries with high temporal resolution (e.g., albedo, LST from the Himawari geostationary satellite).

In general, this study explores the potential of ML approach for the SM downscaling applications at a continental scale. It could be a promising direction to exploit the modelling capability of integrating multisource geospatial data including satellite retrievals, LSM and interpolated ground observation data. The next step could be the establishment of an operational downscaling framework to generate even finer-resolution SM at a daily frequency. Further analyses should be conducted to explore the contribution of different

predictors to SM spatial patterns under different scales, in order to better understand the dominant physical controls on spatial variability of SM.

ACKNOWLEDGMENTS

This research was undertaken while supported by the Australian National University (ANU) University Research Scholarship and the Commonwealth Scientific and Industrial Research Organisation (CSIRO) and ANU Digital Agriculture Supplementary Scholarship through the Centre for Entrepreneurial Agri-Technology (CEAT). This research was supported with funds from the University of Sydney (USYD) and Grains Research and Development Corporation (GRDC) project *SoilWaterNow*. We acknowledge the resources and services provided by the National Computational Infrastructure (NCI), which is supported by the Australian Government through the National Collaborative Research Infrastructure Strategy.

REFERENCES

Ahmad, S., Kalra, A. & Stephen, H. 2010. Estimating soil moisture using remote sensing data: A machine learning approach. *Advances in water resources, 33*, 69-80.

Breiman, L. 2001. Random forests. *Machine learning, 45*, 5-32.

Djamai, N., Magagi, R., Goïta, K., Merlin, O., Kerr, Y. & Roy, A. 2016. A combination of DISPATCH downscaling algorithm with CLASS land surface scheme for soil moisture estimation at fine scale during cloudy days. *Remote Sensing of Environment, 184*, 1-14.

Frost, A., Ramchurn, A. & Smith, A. 2018. The Australian Landscape Water Balance Model. *Bureau of Meteorology: Melbourne, Australia*.

Im, J., Park, S., Rhee, J., Baik, J. & Choi, M. 2016. Downscaling of AMSR-E soil moisture with MODIS products using machine learning approaches. *Environmental Earth Sciences, 75*, 1120.

Kolassa, J., Reichle, R. & Draper, C. S. 2017. Merging active and passive microwave observations in soil moisture data assimilation. *Remote sensing of environment, 191*, 117-130.

Long, D., Bai, L., Yan, L., Zhang, C., Yang, W., Lei, H., Quan, J., Meng, X. & Shi, C. 2019. Generation of spatially complete and daily continuous surface soil moisture of high spatial resolution. *Remote Sensing of Environment, 233*, 111364.

Maggioni, V. & Houser, P. R. 2017. Soil moisture data assimilation. *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. III)*. Springer.

Mao, H., Kathuria, D., Duffield, N. & Mohanty, B. P. 2019. Gap Filling of High‐Resolution Soil Moisture for SMAP/Sentinel‐1: A Two‐Layer Machine Learning‐Based Framework. *Water Resources Research, 55*, 6986-7009.

Peng, J., Loew, A., Merlin, O. & Verhoest, N. E. 2017. A review of spatial downscaling of satellite remotely sensed soil moisture. *Reviews of Geophysics, 55*, 341-366.

Pinnington, E., Amezcua, J., Cooper, E., Dadson, S., Ellis, R., Peng, J., Robinson, E., Morrison, R., Osborne, S. & Quaife, T. 2021. Improving soil moisture prediction of a high-resolution land surface model by parameterising pedotransfer functions through assimilation of SMAP satellite data. *Hydrology and Earth System Sciences, 25*, 1617-1641.

Romano, N. 2014. Soil moisture at local scale: Measurements and simulations. *Journal of Hydrology, 516*, 6-20.

Tian, S., Renzullo, L. J., Pipunic, R. C., Lerat, J., Sharples, W. & Donnelly, C. 2021. Satellite soil moisture data assimilation for improved operational continental water balance prediction. *Hydrology and Earth System Sciences, 25*, 4567-4584.