Response to RC1:

The major contribution of this paper is the use of Bayesian Model Averaging (BMA) to combine the outputs from several different empirical ("machine learning") techniques for soil moisture downscaling. The authors test this methodological innovation by comparing to a large dataset of in-situ soil moisture sensors scattered across northern China. I have several comments that I hope the authors will address.

Response: Thank you for your insightful and constructive feedback, which has been invaluable in improving the quality of our manuscript. We have carefully addressed and responded to each comment in detail. Furthermore, we have introduced additional independent methods and regional comparisons to strengthen the robustness of our analysis and broaden the scope of our findings.

1. First, I believe that statistical derivation of BMA assumes that the models are independent of each other. It seems like the models developed here are likely not independent because they have been developed using the same inputs and the same training data. Have the authors tested whether this assumption applies to their models? If they are dependent, what is the impact on the results?

Response: We appreciate this comment and recognize that BMA indeed assumes independence among models. However, achieving complete independence among models can be challenging in practice, especially when models rely on the same meteorological input data. In our study, we found Pearson correlations among the four models' outputs to be approximately 0.75-0.85, indicating moderate dependency.

While this dependency might influence the BMA assumption of model independence, it does not necessarily violate it. Generally, BMA's independence assumption is not an absolute requirement for zero correlation but rather seeks models that contribute some level of unique information, thus reducing uncertainty in the ensemble [1, 2]. In practice, even with shared inputs, if the models exhibit unique structural differences such as ET and soil moisture [3, 4], Bayesian model can still effectively enhance ensemble accuracy by drawing on these differences.

Additionally, the observed dependencies in our models resemble convergence toward the true target rather than redundancy, meaning that each model's output aims to approximate the same objective (i.e., true soil moisture values) rather than merely replicating each other. As long as models are not entirely redundant (providing the same information), BMA can still yield effective integration and improvements in overall accuracy.

To further explore the impact of dependency, we also conducted additional analyses, which have been added in the main text section 4.5 and supplementary Fig. S8.

1) We assessed the sensitivity of the BMA ensemble by systematically removing one model at a time to evaluate any significant changes in accuracy. We observed accuracy reductions ranging from 7-12% when specific models were removed. This outcome suggests that each model contributes unique information that significantly impacts the ensemble's performance. If the models were highly dependent, removing one would not cause a notable change in accuracy, as the remaining models would effectively compensate. This sensitivity analysis indicates that while the datasets are correlated, they are not redundant, as each provides distinct characteristics or

features. The purpose of BMA is to leverage unique information from each model through weighted averaging to reduce uncertainty. If there were strong dependencies, BMA's weighting mechanism would be less effective. However, the observed impact on accuracy confirms that these dependencies do not compromise BMA's integrative performance. Therefore, this analysis reasonably demonstrates that any interdependence among the models has only a limited impact on BMA's effectiveness, and that each model contributes valuable information to the ensemble.

2) To further investigate the impact of dependencies, we also compared BMA results with those from a Hierarchical Bayesian Model (HBM) [5, 6], which explicitly accounts for dependency structures. HBM incorporates each data source as a distinct hierarchical level, introducing a "data source bias" random effect to model deviations of each source from the global mean. This framework allows HBM to account for dependencies by quantifying and mitigating inter-model biases. Our results showed that HBM achieved an accuracy improvement of less than 8% over BMA, indicating that while HBM accounts for dependencies, these dependencies and any associated systematic biases only moderately impact ensemble accuracy. This suggests that BMA remains a viable approach for practical applications, particularly in scenarios where data complexity or computational constraints make it preferable. Although Bayesian methods like HBM offer advantages over traditional techniques, they are sensitive to prior settings. The choice of priors can strongly influence posterior distributions, and if observational data does not fully represent true conditions—due to factors such as low data quality or limited spatial coverage—model calibration may affect overall accuracy.

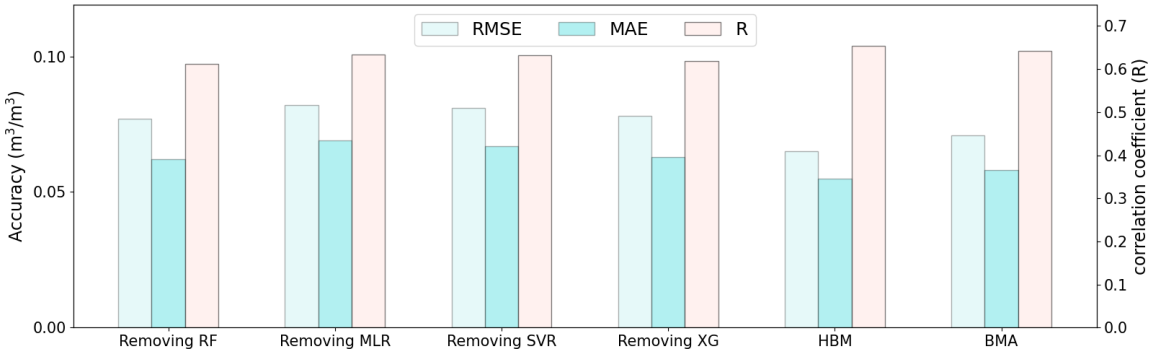

Figure S8. Sensitivity analysis of the Bayesian Model Averaging (BMA) ensemble, evaluating the impact of systematically removing one model at a time on accuracy. Accuracy reductions of 7–12% were observed when specific models were excluded, indicating that each model contributes unique information critical to the ensemble's performance. If the models were highly dependent, removing one would result in minimal accuracy changes, as the remaining models would compensate. This analysis demonstrates that while the datasets exhibit some correlation, they are not redundant, as each provides distinct and valuable features. To further explore the impact of model dependencies, we compared the BMA results with those from a Hierarchical Bayesian Model (HBM) (Sairam et al., 2019), which explicitly incorporates dependency structures. HBM treats each data source as a distinct hierarchical level and introduces a "data source bias" random effect to account for deviations of individual sources from the global mean. This approach enables HBM to quantify

and mitigate inter-model biases. Results showed that HBM achieved a modest accuracy improvement of less than 8% over BMA, suggesting that while HBM better accounts for dependencies, these dependencies and associated systematic biases have only a moderate effect on ensemble accuracy. These findings underscore BMA's robustness and practicality, particularly in scenarios where data complexity or computational constraints make it preferable. While advanced Bayesian approaches like HBM offer benefits, such as explicitly modeling dependencies, they are sensitive to prior settings. The choice of priors can significantly influence posterior distributions, and inaccuracies in observational data—stemming from low quality or limited spatial coverage—can impact model calibration and overall accuracy.

[1] Hollenbach, F. M., & Montgomery, J. M. (2020). Bayesian model selection, model comparison, and model averaging. *The sage handbook of research methods in political science and international relations*, 937-960.
[2] Raftery, A. E., Gneiting, T., Balabdaoui, F., & Polakowski, M. (2005). Using Bayesian model averaging to calibrate forecast ensembles. *Monthly weather review*, *133*(5), 1155-1174.
[3] Shao, X., Zhang, Y., Liu, C., Chiew, F. H., Tian, J., Ma, N., & Zhang, X. (2022). Can indirect evaluation methods and their fusion products reduce uncertainty in actual evapotranspiration estimates?. *Water Resources Research*, *58*(6), e2021WR031069.
[4] Chen, Y., Yuan, H., Yang, Y., & Sun, R. (2020). Sub-daily soil moisture estimate using dynamic Bayesian model averaging. Journal of Hydrology, 590, 125445.
[5] Kim, T. J., Kwon, H. H., & Lima, C. (2018). A Bayesian partial pooling approach to mean field bias correction of weather radar rainfall estimates: Application to Osungsan weather radar in South Korea. *Journal of Hydrology*, *565*, 14-26.
[6] Sairam, N., Schröter, K., Rözer, V., Merz, B., & Kreibich, H. (2019). Hierarchical Bayesian approach for modeling spatiotemporal variability in flood damage processes. Water resources research, 55(10), 8223-8237.

2. Second, all the downscaling methods considered provide very little improvement in the soil moisture estimates. A key goal of downscaling is to include fine scale spatial variability that is not present in the coarse resolution input. However, when I examine the histograms in Figure 6, I see no increase in the variability of soil moisture when the downscaling methods are applied. Some of the methods have less variability than the coarse resolution input. Are these methods successfully introducing any variability in the patterns? Also, the accuracy of the BMA method is only slightly better than the coarse resolution input. The exact improvement is difficult to see because Table 4 does not include the performance of the coarse resolution input nor the overall performance across all the datasets used. Those should be added). The authors seem satisfied with the improvement in their discussion and conclusions, but it seems like the improvements do not warrant the huge processing involved. The authors consider relatively few variables. Could better performance be achieved by using model inputs?

Response: Thank you for your constructive feedback, which has strengthened our study. 1) Our study's primary objective was to produce a 1-km soil moisture product by downscaling satellite-derived datasets and calibrating model bias with ground-based soil moisture measurements. This approach differs from traditional downscaling. Generally, in the remote sensing field, downscaling refers to using coarse-resolution data along with high-resolution auxiliary datasets to predict the target variable at finer resolutions. The primary focus here was not solely on increasing spatial variability but on generating reliable high-resolution soil moisture data for arid regions such as Northern China, where coarse-resolution ESA CCI data may overestimate surface moisture due to the limitations in capturing finer-scale processes [1, 2]. In arid and semi-arid regions, high surface exposure and low vegetation cover lead to rapid surface drying. Consequently, ESA CCI's microwave sensing, while sensitive to surface moisture, struggles to accurately capture deeper moisture levels, resulting in potential overestimation of surface soil moisture.

Additionally, machine learning techniques, particularly those constrained by field observations, tend to smooth extreme values, resulting in reduced spatial variability [3, 4]. Current machine learning models often exhibit limitations in capturing extreme soil moisture values, especially under drought conditions, and tend to make conservative predictions, which can lead to underestimation in dry areas.

It's important to clarify that a narrower range in soil moisture values after downscaling does not imply a reduction in spatial variability [5-7]. The spatial variability depends on the model's capability to capture local details and the heterogeneity present in the original data. Downscaling to higher resolution allows local features—such as topography, land cover, or moisture gradients—to emerge more clearly, thereby enhancing spatial variability where appropriate. In fact, as observed in the histograms and box plots, our downscaled product shows increased clustering in the middle range, with values in this range trending upward after downscaling. Using metrics like the coefficient of variation (CV) and Moran's I, we observe an increase in local spatial variability within the middle range, while the extreme ranges exhibit less pronounced variability.

In the new version, we have added the related context in the main text section 4.1 and supplementary Table S3.

Table S3. Coefficient of variation (CV) and Moran's I index

	ESA CCI	*RF*	*MLR*	*SVR*	*XG*	*BMA*
CV	*0.321*	*0.328*	*0.313*	*0.324*	*0.332*	*0.324*
Moran's I [0-100%]	*0.994*	*0.964*	*0.955*	*0.994*	*0.946*	*0.972*
Moran's I [0-15%]	*0.991*	*0.755*	*0.876*	*0.991*	*0.846*	*0.86*
Moran's I [15-85%]	*0.99*	*0.795*	*0.922*	*0.989*	*0.825*	*0.803*
Moran's I [85-100%]	*0.991*	*0.864*	*0.921*	*0.991*	*0.895*	*0.902*

Note: CV measures overall variability, with higher values indicating stronger heterogeneity. Moran's I quantifies spatial distribution patterns, where higher values reflect weaker heterogeneity and stronger spatial autocorrelation. The CV is calculated for the entire dataset. Moran's I index is determined using a simple four-neighborhood relationship, with brackets indicating different sample divisions. The 0-100% range represents the full sample, while the 0-15%, 15-85%, and 85-100% ranges correspond to low-value, mid-range, and high-value distributions, respectively.

Table 4. Comparison of BMA and individual machine learning

Stations	Num	R											
		RF		MLR		SVR		XG		BMA		ESA CCI	
		All	Mid	All	Mid	All	Mid	All	Mid	All	Mid	All	Mid
NZW	5044	0.323	0.383	0.338	0.411	0.321	0.398	0.325	0.399	0.342	0.424	0.321	0.375
CERN	263	0.567	0.664	0.617	0.693	0.629	0.705	0.573	0.672	0.642	0.721	0.586	0.647
QXZ	1204	0.477	0.571	0.480	0.593	0.478	0.583	0.468	0.554	0.514	0.610	0.479	0.531
Stations	Num	RMSE (m3/m3)											
		RF		MLR		SVR		XG		BMA		ESA CCI	
		All	Mid	All	Mid	All	Mid	All	Mid	All	Mid	All	Mid
NZW	5044	0.138	0.108	0.136	0.105	0.138	0.106	0.137	0.104	0.137	0.097	0.138	0.115
CERN	263	0.073	0.060	0.073	0.059	0.073	0.060	0.075	0.061	0.071	0.054	0.084	0.064
QXZ	1204	0.169	0.125	0.169	0.131	0.168	0.132	0.169	0.128	0.169	0.115	0.168	0.139
Stations	Num	MAE (m3/m3)											
		RF		MLR		SVR		XG		BMA		ESA CCI	
		All	Mid	All	Mid	All	Mid	All	Mid	All	Mid	All	Mid
NZW	5044	0.114	0.092	0.113	0.091	0.114	0.095	0.114	0.095	0.114	0.089	0.114	0.101
CERN	263	0.060	0.038	0.061	0.039	0.061	0.039	0.062	0.040	0.058	0.032	0.067	0.045
QXZ	1204	0.157	0.119	0.158	0.123	0.156	0.117	0.158	0.120	0.158	0.114	0.157	0.134

Note: "All" refers to the full set of sample points, whereas "Mid" denotes the subset of sample points that fall within the 15-85% range.

2) We have revised Table 4 to include performance metrics for the coarse-resolution input, enabling clearer comparison across all datasets. While accuracy validation is crucial, it represents just one aspect of our evaluation. Site-scale validations are subject to scale effects, so we conducted a comprehensive assessment that included drought event capture and product comparisons. Our Bayesian framework, combined with ground observations, successfully generated a stable high-resolution soil moisture dataset.

Although the overall accuracy gains may appear modest due to the large study area and site data scale effects, our work remains unique. Few studies attempt site-based soil moisture downscaling over a large area such as northern China. When focusing on specific regions, such as the Loess Plateau and the North China Plain—semi-arid areas with rich site data—the accuracy improvements become more pronounced, highlighting the robustness and utility of our dataset and approach.

3) Our choice of explanatory variables was guided by two main principles: (i) ensuring that we had stable, reliable remote sensing observations available at a large scale, thus allowing for future applications in other regions or even at a global scale; and (ii) selecting variables with strong correlations to soil moisture but minimal redundancy. The five variables we chose represent key drivers of soil moisture across meteorological, ecological, and hydrological dimensions, with low inter-correlation. Related context has been improved in the main text section 3.1 and Figure 3.

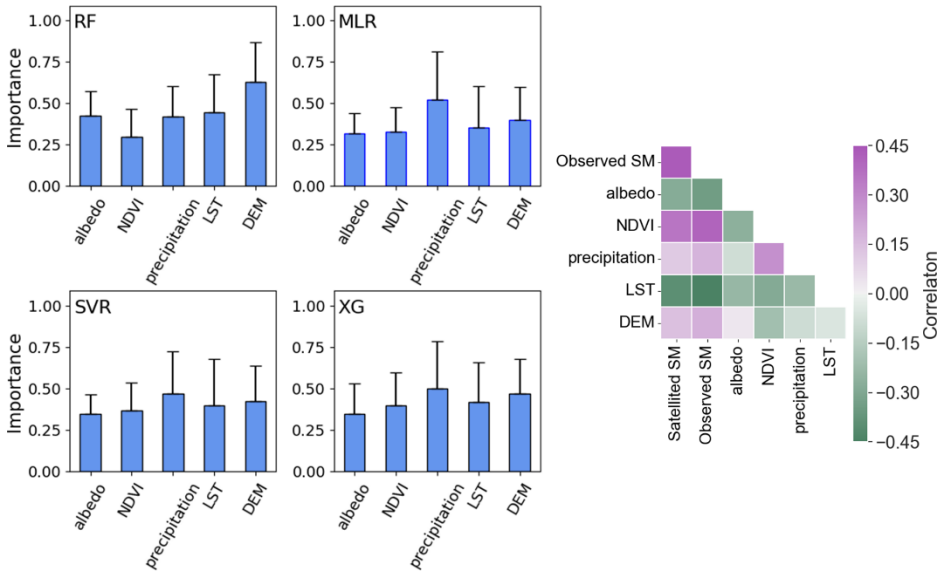

Figure 3: Assessment of explanatory variables' feasibility. (a) Average (blue bar) and standard deviation (error bar) of permutation-based importance of explanatory variables concerning soil moisture. (b) Average Pearson correlation coefficients among different explanatory variables, including correlations with two independent soil moisture data sources.

Other potential variables, such as vegetation indices (e.g., EVI), downwelling radiation, and evapotranspiration, were excluded due to their high correlation with the selected variables, limited efficacy in arid and semi-arid regions like northern China, and inconsistency in accuracy at daily and fine spatial scales. Soil attributes, such as texture and classification, are often critical in soil moisture modeling [8], yet in our study, they contributed less than 2% to overall accuracy improvements. This may be due to the relative homogeneity in soil texture across northern China, where sandy soils and loams predominate, offering little spatial variation to capture soil moisture heterogeneity. Moreover, the spatial partitioning employed before model implementation likely accounted for soil characteristics within each subregion, further diminishing the impact of texture. Consequently, soil texture added minimal explanatory value. In summary, while our choice of variables may omit certain minor features, the overall accuracy is robust and serves as a valuable reference for large-scale and global soil moisture studies.

[1] Zhang, G., Su, X., Ayantobo, O. O., & Feng, K. (2021). Drought monitoring and evaluation using ESA CCI and GLDAS-Noah soil moisture datasets across China. *Theoretical and Applied Climatology*, *144*, 1407-1418.
[2] Dorigo, W. A., Gruber, A., De Jeu, R. A. M., Wagner, W., Stacke, T., Loew, A., ... & Kidd, R. (2015). Evaluation of the ESA CCI soil moisture product using ground-based observations. *Remote Sensing of Environment*, *162*, 380-395.
[3] Sadayappan, K., Kerins, D., Shen, C., & Li, L. (2022). Nitrate concentrations predominantly driven by human, climate, and soil properties in US rivers. *Water Research*, *226*, 119295.
[4] Bo, Y., Li, X., Liu, K., Wang, S., Li, D., Xu, Y., & Wang, M. (2024). Hybrid theory-guided data driven framework for calculating irrigation water use of three staple cereal crops in China. *Water Resources Research*, *60*(3), e2023WR035234.
[5] Wang, F., & Tian, D. (2024). Multivariate bias correction and downscaling of climate models with trend-preserving deep learning. *Climate Dynamics*, *62*(10), 9651-9672.
[6] Maraun, D., Wetterhall, F., Ireson, A. M., Chandler, R. E., Kendon, E. J., Widmann, M., ... & Thiele-Eich, I. (2010). Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user. *Reviews of geophysics*, *48*(3).
[7] Latombe, G., Burke, A., Vrac, M., Levavasseur, G., Dumas, C., Kageyama, M., & Ramstein, G. (2018). Comparison of spatial downscaling methods of general circulation model results to study climate variability during the Last Glacial Maximum. *Geoscientific Model Development*, *11*(7), 2563-2579.
[8] Xu, M., Yao, N., Yang, H., Xu, J., Hu, A., de Goncalves, L. G. G., & Liu, G. (2022). Downscaling SMAP soil moisture using a wide & deep learning method over the Continental United States. *Journal of Hydrology*, *609*, 127784.

3. Third, little consideration is given as to whether the in situ dataset adequately captures 1-km spatial variations in soil moisture (which is the stated goal of the downscaling method). The measurement support is likely very small and the spacing is likely much larger than 1-km. Even

if the downscaling models reproduce this dataset exactly, have we really developed an accurate 1-km resolution soil moisture estimate? Can the authors provide some support that that a given in situ soil moisture observation is representative of its 1-km grid cell? Also, can the authors show that the collection of 1-km grid cells that have in situ observations capture the range of conditions that occur within the region? I believe some support along these lines would greatly strengthen the paper.

Response: Thanks for pointing out this issue.

1) We agree that scale effects are among the most significant challenges in remote sensing validation, especially for soil moisture downscaling to a 1-km resolution. Currently, there is no definitive solution to fully bridge the scale gap between in situ observations and satellite-based products. Capturing soil moisture variability at the 1-km scale is particularly challenging across northern China's extensive 3-million-square-kilometer study area, where diverse climate and surface characteristics further complicate validation. In light of these challenges, our study employs a multi-faceted validation approach. In addition to site-based validation, we incorporate drought event analysis and cross-product comparisons. This broader evaluation framework aligns with mainstream practices in current remote sensing research to address validation limitations from scale effects.

Furthermore, our results from a reduced-sample analysis suggest that the scale effect's impact on model outcomes is less significant than anticipated, supporting the robustness and reliability of our findings despite scale challenges. These approaches together reinforce the credibility of our model outputs by considering spatial variability within the constraints of available data.

2) One of the key strengths of our study is the integration of extensive ground data to calibrate remote sensing products and model outputs, reducing errors arising from surface heterogeneity and better aligning the model with actual ground conditions. However, while this integration helps minimize discrepancies, it can also introduce new mismatches between in situ and satellite data—an area that requires further attention in future research and remains a focus in many recent studies.

In this study, we address regional heterogeneity by dividing the study area into several subregions and calibrating the model with in situ data for each specific subregion. This process allows the model to learn distinct calibration parameters relevant to each area. Although this method effectively incorporates regional variations, it cannot fully eliminate scale-induced transfer effects at finer, localized scales. Moving forward, we plan to explore transfer learning techniques and develop specific loss functions designed to reduce scale-bias when calibrating 1-km satellite data with ground-based measurements. Such methods could enhance calibration accuracy and improve the model's adaptability across different spatial scales.

In the new version, we have added the related context in the main text section 4.5.

3) In response to these concerns, we conducted an additional experiment to examine the impact of scale effects on model accuracy. This experiment focused on the Maqu region [1, 2], located at the transition zone between the Tibetan Plateau and the Loess Plateau. Maqu's relatively flat terrain and predominantly grassland cover makes it suitable for comparative analysis, and the presence of

20 ground stations within a 5x5 grid enhances its suitability as a case study for evaluating scale effects.

Our model showed a significant accuracy improvement in this flat, homogeneous region of Naqu, highlighting the pronounced influence of scale effects in regions with minimal topographic variation. Furthermore, we conducted a sequential data reduction analysis, removing 10%, 20%, 30%, and 40% of ground training data while maintaining the same validation dataset. Although model accuracy was somewhat affected by the reduction in training data, the impact was relatively modest. This finding indicates that while sample data quantity influences the overall outcome, the scale effect on model validation remains relatively minor. Specifically, even with reduced training samples, the validation accuracy remained stable, suggesting that the scale information learned by the model from ground station data is sufficiently generalized to apply to the validation set. In essence, this stability implies that the scale difference between ground station data and 1-km remote sensing data does not introduce significant bias in model validation [3,4].

In the new version, we have added the related context in the main text section 4.5 and supplementary Fig. S7.

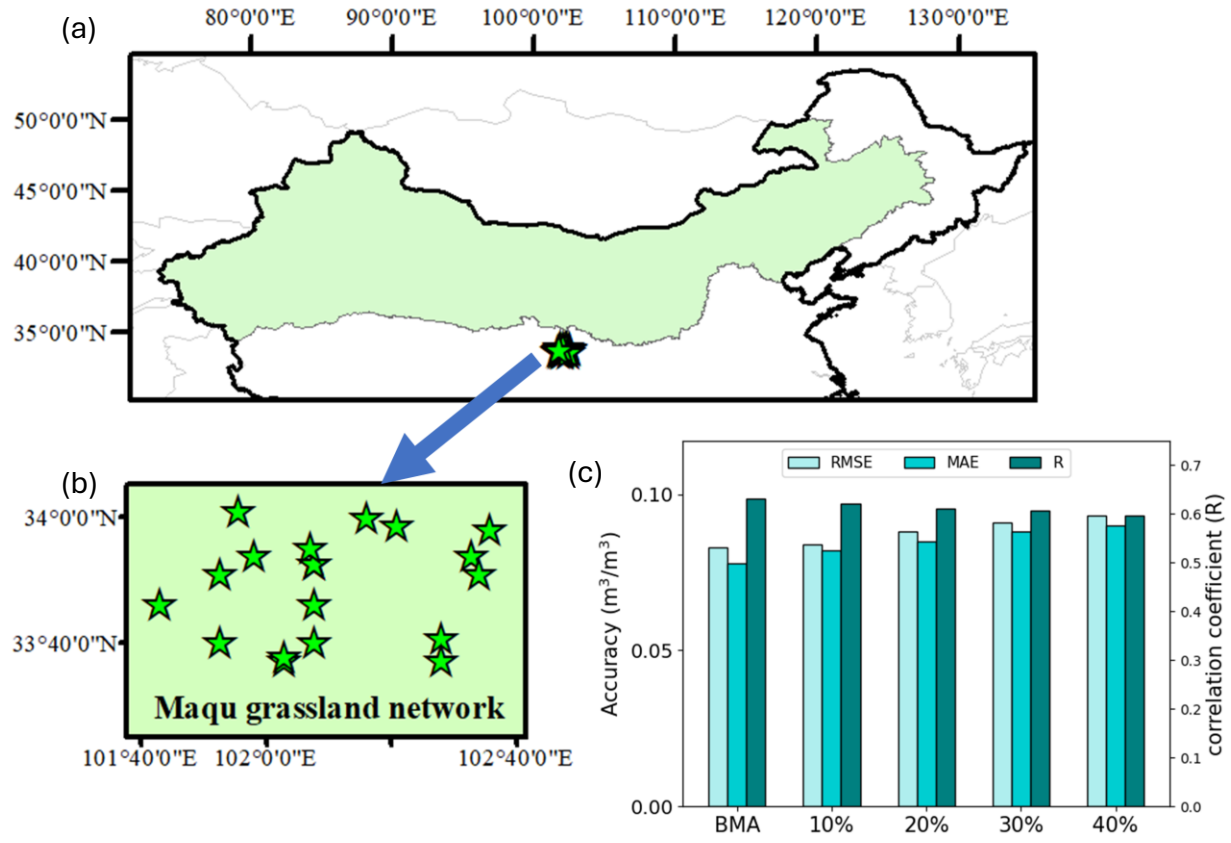

Figure S7. Additional experiment examining the impact of scale effects on model accuracy and the representativeness of in situ datasets in capturing soil moisture spatial variations. (a) The experiment was conducted in the Maqu region, a transitional zone between the Tibetan Plateau and the Loess Plateau, characterized by relatively flat terrain and predominantly grassland cover. (b) These features, combined with the presence of 20 ground stations arranged in a 5x5 grid, make

Maqu an ideal case study for evaluating scale effects. (c) The model demonstrated significant accuracy improvements in this flat, homogeneous region, underscoring the pronounced influence of scale effects in areas with minimal topographic variation. A sequential data reduction analysis was also performed, removing 10%, 20%, 30%, and 40% of the ground training data while maintaining the same validation dataset. Although the reduction in training data modestly impacted model accuracy, the effect was relatively minor. The validation accuracy remained stable even with fewer training samples, suggesting that the model effectively generalized the scale information learned from ground station data to the validation set. This stability indicates that the scale differences between ground station data and 1-km remote sensing data introduce negligible bias in model validation, reaffirming the robustness of the model's performance in addressing scale effects.

[1] Dente, L., Vekerdy, Z., Wen, J., & Su, Z. (2012). Maqu network for validation of satellite-derived soil moisture products. International Journal of Applied Earth Observation and Geoinformation, 17, 55-65.
[2] Liu, K., Li, X., Wang, S., & Zhang, H. (2023). A robust gap-filling approach for European Space Agency Climate Change Initiative (ESA CCI) soil moisture integrating satellite observations, model-driven knowledge, and spatiotemporal machine learning. Hydrology and Earth System Sciences, 27(2), 577-598.
[3] Dorigo, W. A., Gruber, A., De Jeu, R. A. M., Wagner, W., Stacke, T., Loew, A., ... & Kidd, R. (2015). Evaluation of the ESA CCI soil moisture product using ground-based observations. Remote Sensing of Environment, 162, 380-395.
[4] Brocca, L., Hasenauer, S., Lacava, T., Melone, F., Moramarco, T., Wagner, W., ... & Bittelli, M. (2011). Soil moisture estimation through ASCAT and AMSR-E sensors: An intercomparison and validation study across Europe. Remote Sensing of Environment, 115(12), 3390-3408.

4. I would suggest removing the Noah results because they really don't contribute to testing the innovation that is presented.

Response: We acknowledge that capturing soil moisture variability at a 1-km resolution is particularly challenging across northern China's extensive 3-million-square-kilometer study area, where diverse climate and surface conditions further complicate the validation process. Given these challenges, our study employs a multi-faceted validation approach. In addition to site-based validation, we incorporate drought event analysis and cross-product comparisons. This comprehensive evaluation framework aligns with mainstream practices in remote sensing research and is designed to address the limitations posed by scale effects in validating downscaled products.

In response to the feedback from the editor-in-chief and reviewers, we have retained the Noah results but revised this section to clarify its relevance. We also streamlined some parts of the manuscript by moving certain elements of the uncertainty analysis from the appendix to the main text, ensuring a clearer focus on the innovative aspects of our methodology.