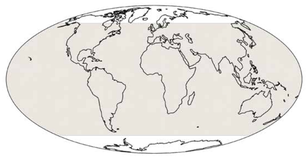
IRMerg: Enhancing Global Infrared Precipitation Estimates with Land Surface Variables and Contributing Factors Analysis Using Explainable Machine Learning

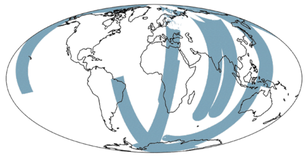
Ho Tin Hung and Li-Pen Wang
Department of Civil Engineering, National Taiwan University, Taipei 106, Taiwan

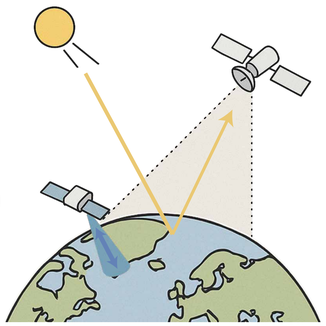
1. Introduction

Infrared observations (Passive):
Geostationary satellites
Broad spatio-temporal coverage
Weak relationship between cloud top temperatures and precipitation



Microwave observations (Active):
Low-orbit satellites
Limited spatio-temporal coverage
Accurate precipitation estimates

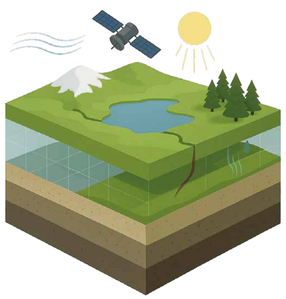




Research Questions:

1. Can Land Surface Variables enhance IR precipitation estimates?
2. What are the contributions of different land surface variables at spatial and temporal scales?

2. Data and Methodology



[Number of variables]

 # Temperature [7]
 # Radiation and Heat [7]
 # Wind and Pressure [3]

 # Soil Water [4]
 # Vegetation [2]

 # Lakes [7]
 # Evaporation and Runoff [10]
 # Snow [8]

Model predicts IR-error (IR - MW), where MW is the ground truth.

Infrared estimated precipitation + **Geo location** + **49 Land surface variables**

XGBoost model

Predicted error → **Calibrate & Evaluation**

SHAP value → **Contributing factors analysis**

Study Period: 2022 (Full Year)
Study Area: Quasi-Global (Latitude: 60°S to 60°N; Longitude: 180°W to 180°E)

3. Results and Contributing Factors Analysis

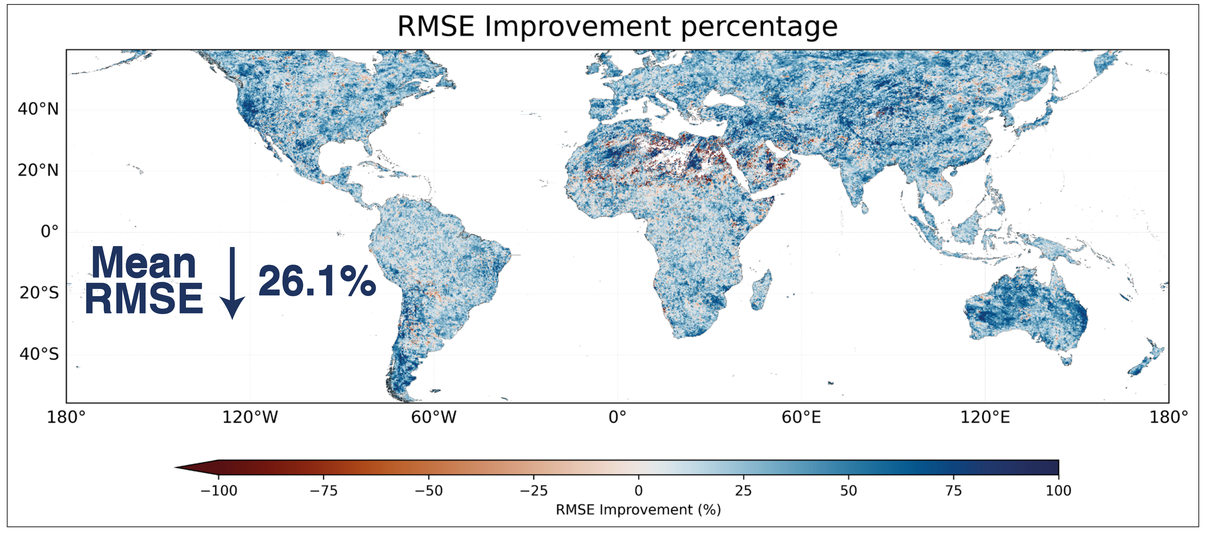
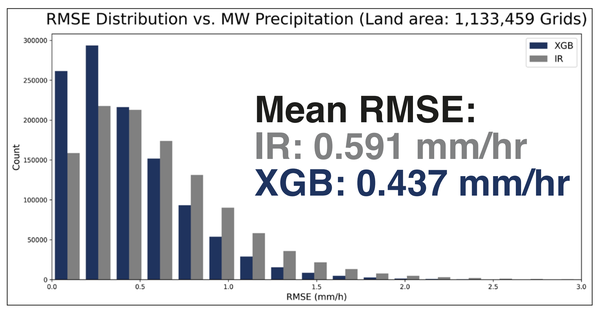


Figure 1. Global distribution of RMSE improvement (%) for precipitation estimates. Blue areas indicate regions with improved RMSE after calibration, while red areas show degradation. The map covers latitudes from 60°S to 60°N and longitudes from 180°W to 180°E, demonstrating widespread improvement across most land regions, especially in South America, Sub-Saharan Africa, and Australia.



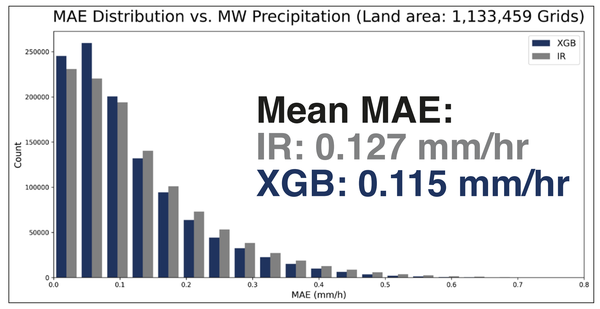


Figure 2. Histogram of RMSE and MAE for IR-based (gray) and XGBoost-corrected (blue) precipitation estimates, using microwave (MW) observations as the reference. The distribution illustrates a leftward shift for the XGBoost results, indicating improved accuracy compared to the original IR estimates.

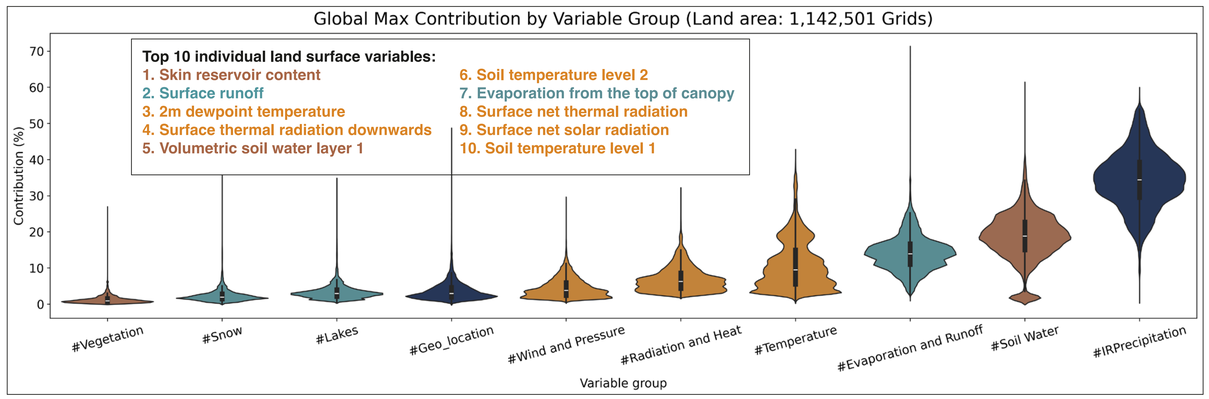


Figure 3. Violin plot of SHAP value-based maximum contributions by variable group. Each distribution represents how dominant a variable group is in contributing to the final precipitation prediction. The top contributing variables include skin reservoir content, surface runoff, and 2m dewpoint temperature, reflecting the critical role of surface and atmospheric water-related processes.

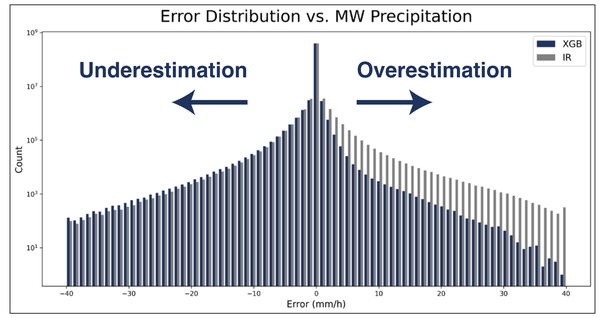


Figure 4. Histogram of prediction error (mm/h) for IR-based (gray) and XGBoost-corrected (blue) precipitation estimates, using microwave (MW) data as the ground truth.

4. Key findings

Significant Error Reduction:
The XGBoost model integrating land surface variables substantially reduced overall precipitation estimation errors (e.g., mean RMSE reduced by 26.1%, MAE also lowered). The method was especially effective at correcting overestimations, though less successful with underestimations.

Key Contributing Factors :
Explainable AI (SHAP) identified Soil Water content, Evaporation/Runoff, and Temperature variables as the key LSVs improving precipitation estimates. Given the Earth's dynamic systems, future calibration should consider incorporating additional variables specific to local conditions.

Spatially Clustered Importance:
Land surface variables importance varies regionally, showing distinct patterns linked to local land-surface processes influencing precipitation improvements.

Stable Temporal Variation:
Key land surface variables importance remained relatively stable across seasons (2022), suggesting their influence is consistent over time.

Variable Interactions Noted:
SHAP value variations across model runs suggest complex land surface variables interactions requiring further investigation to fully understand these relationships.

5. References

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

Crameri, F., Shephard, G. E., & Heron, P. J. (2020). The misuse of colour in science communication. Nature communications, 11(1), 5444.

Huffman, G. J., Bolvin, D. T., Joyce, R., Kelley, O. A., Nelkin, E. J., Portier, A., ... & West, B. J. (2023). IMERG V07 release notes. Goddard Space Flight Center: Greenbelt, MD, USA.

Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., ... & Thépaut, J. N. (2021). ERA5-Land: A state-of-the-art global reanalysis dataset for land applications. Earth system science data, 13(9), 4349-4383.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.

Ho Tin Hung I MSc I isaacgbhk@caece.net