PLOS Climate

RESEARCH ARTICLE

Disentangling key cloud properties for precipitation retrievals from geostationary satellite data using machine learning

Hwayon Choi, Yong-Sang Choi*, Chang-Hoi Ho, Jinwon Kim

Department of Climate and Energy Systems Engineering, Ewha Womans University, Seoul, Republic of Korea

* ysc@ewha.ac.kr

Check for updates

OPEN ACCESS

Citation: Choi H, Choi Y-S, Ho C-H, Kim J (2026) Disentangling key cloud properties for precipitation retrievals from geostationary satellite data using machine learning. PLOS Clim 5(3): e0000860. https://doi.org/10.1371/journal.pclm.0000860

Editor: Jingyu Wang, Nanyang Technological University, SINGAPORE

Received: November 18, 2025

Accepted: March 1, 2026

Published: March 30, 2026

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pclm.0000860

Copyright: © 2026 Choi et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution,

Abstract

Cloud properties governing precipitation formation remain poorly understood due to their intrinsic complexity and the difficulty of identifying physically consistent predictors. This study identifies the key variables in precipitation retrievals over the East Asian mid-latitude and tropical warm pool regions using the geostationary satellite data collected in the year 2023, in conjunction with machine learning and Shapley additive explanations. In the East Asian mid-latitudes, light-to-moderate precipitation is associated with the ice water path and the cloud particle growth characteristics, as indicated by the spectral reflectance (R) difference between 1.61 μm and 0.47 μm ($R_{0.47} - R_{1.61}$) and between 1.61 μm and 0.64 μm ($R_{0.64} - R_{1.61}$), respectively. Heavy precipitation events are linked to the presence of high-altitude cirrus clouds indicated by $R_{1.37}$ and the upper-tropospheric moisture content characterized by the brightness temperature (BT) difference between 6.2 μm and 9.6 μm ($BT_{6.2} - BT_{9.6}$). Over the tropical warm pool region, light-to-moderate precipitation is related to ice water path and cloud optical thickness, while heavy precipitation is predominantly related to the surface reflectance contrast, represented by $R_{0.47} - R_{0.86}$ and $R_{0.64} - R_{0.86}$, which modulates the initiation and intensity of convection. These results demonstrate that the optimal cloud properties and environmental predictors for satellite-based precipitation retrievals vary by regions and intensities.

1. Introduction

Geostationary satellites play an important role in monitoring atmospheric phenomena as they produce high-resolution data in near real time. Multispectral radiation measurements from infrared and visible channels [1–4] allow the retrieval of atmospheric variables at high spatiotemporal resolutions. Geostationary satellite observations are valuable for precipitation estimation because they provide continuous monitoring of clouds and atmospheric conditions over large spatial domains. Previous studies have developed various precipitation estimation models based on machine learning

and reproduction in any medium, provided the original author and source are credited.

Data availability statement: The GPM IMERG datasets are available from National Aeronautics and Space Administration (NASA) website at https://gpm1.gesdisc.eosdis.nasa.gov/data/GPM_L3/GPM_3IMERGHH.07/. The GK2A data are available from Korea Meteorological Administration (KMA) data website at https://data.kma.go.kr/data/rmt/rmtList.do?code=21&pgmNo=683. All codes used for train the RF model and compute SHAP values in this study are available in public GitHub repository: https://github.com/hyc-atmos/rain-shap-paper.

Funding: This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (RS-2018-NR031078) and by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2021-NR058849). The first author was supported by the Specialized university program for confluence analysis of Weather and Climate Data of the Korea Meteorological Institute (KMI) funded by the Korean government (KMA). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

(ML) using data from various geostationary satellites [5–21]. While these models yield good performance in general, with probability of detection (POD) of up to 0.9, their results often vary across regions and depend on the choice of input variables. Furthermore, since cloud-precipitation transitions involve complex physical and dynamical interactions among hydrometeors within clouds, it is difficult to quantify the influence of individual predictors on model accuracy.

Although statistical significance of selected input variables is fundamental in ML-based precipitation estimations [22–24], physical interpretability of the selected variables is also crucial for developing robust and transferable algorithms [25–27]. However, most of existing ML-based precipitation estimation studies rely on data-driven models of limited physical interpretability, since the contribution of individual input variables to the estimation is not explicitly quantified. This lack of transparency hampers understanding of the underlying physical processes related to the precipitation estimation and limits the confidence in applying ML models across various climatic regimes. Recently, explainable artificial intelligence (XAI) techniques have gained increasing attention in atmospheric sciences as a means of improving interpretability of complex ML models to enhance trust in their predictions. By explicitly attributing ML model outputs to individual input variables, XAI presents a pathway to link statistical model performance to physically meaningful processes.

This requirement for interpretability stems from the fact that precipitation formation is governed by complex physical and dynamical interactions among hydrometeors within clouds. In addition, key precipitation formation mechanisms exhibit pronounced regional variations driven by local climate, orography, and thermodynamic environments. Thus, physically consistent variable selection strategies must be tailored for specific regions. Given the intrinsic coupling between cloud physical properties and precipitation processes, many studies incorporated cloud-related parameters as input variables into satellite precipitation algorithms [28–36]. Commonly employed variables include the cloud optical thickness, the effective radius of cloud particles, cloud top temperature, ice water path, and various spectral indices that capture cloud the phase and size distribution of cloud particles [5,14,17]. However, the unprecedented volume and diversity of modern geostationary satellite observations complicate the selection of optimal input variable combinations for region-specific precipitation estimations [37]. This complexity arises from highly nonlinear interactions among multiple cloud properties, the variations of key cloud physical processes across climate regimes, and the scale-dependent nature of precipitation formation.

This study develops region-specific precipitation retrieval algorithms using the Korean geostationary meteorological satellite Geo-KOMPSAT-2A (GK2A) daytime data in 2023 in conjunction with ML models. The algorithms are constructed separately for the two precipitation regimes, light-to-moderate (0.5 to 10 mm h^{-1}) and heavy (> 10 mm h^{-1}) according to the World Meteorological Organization, 2018 [38], across two climatologically-distinct regions, the East Asian mid-latitude and the tropical warm pool (TWP) regions. The SHapley Additive exPlanations (SHAP) algorithm is employed to quantify the relative importance of each satellite-derived variable according to the selected precipitation-intensity categories and regions. By identifying

key physical mechanisms which drive precipitation formation in different climatic regimes related to key input satellite variables, this study establishes a scientific basis for the development of robust and transferable ML algorithms for precipitation monitoring and forecasting.

2. Materials and methods

Fig 1 provides an overview of the overall research methodology adopted in this study, including the training datasets derived from geostationary satellite observations and target precipitation data, the experimental domains, and the modeling and analysis workflow. The training datasets and precipitation categorization into light-to-moderate and heavy rain are shown in Fig 1a. The temporal and spatial domains of the experiment—along with the global distribution of the 99th percentile of precipitation for each region over the 10-year period from 2015 to 2024—are presented in Fig 1b. The modeling and analysis workflow is summarized in Fig 1c, including region- and intensity-specific random forest (RF) models, performance evaluation using standard validation metrics, SHAP-based feature importance analysis, and model optimization.

2.1. Data

This study uses all 16 channels of the GK2A Advanced Meteorological Imager (AMI) for comprehensive analyses (https://data.kma.go.kr/data/rmt/rmtList.do?code=21&pgmNo=683.) [4,39]. We calculate and use reflectance (R) values from six

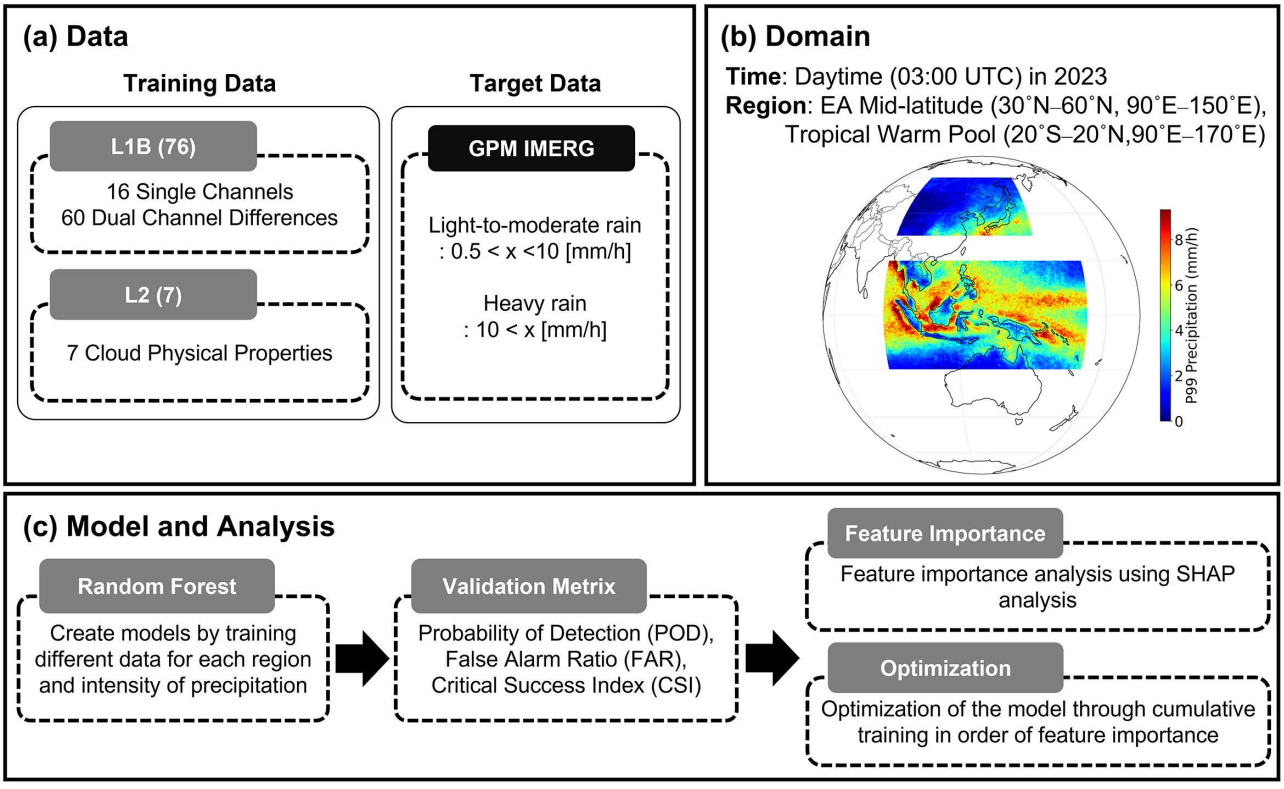


Fig 1. The framework diagram of the overall research process. (a) shows the training data and target data used in the experiment, **(b)** shows the temporal and spatial domain of the experiment and the globe map shows 99th percentile of precipitation in each region over a 10-year (2015–2024) period. **(c)** shows the model and analysis. Panel b base map coastlines were generated using the GSHHS (Global Self-consistent Hierarchical High-resolution Shorelines) dataset (https://www.ngdc.noaa.gov/mgg/shorelines/gshhs.html), which is distributed under the GNU Lesser General Public License (LGPL) (https://www.gnu.org/licenses/lgpl-3.0.html).

https://doi.org/10.1371/journal.pclm.0000860.g001

channels, where the four visible channels (centered at 0.47, 0.51, 0.64, and 0.86 μm) provide information on the daytime cloud cover, aerosols, and surface reflectance, and the two near-infrared channels (1.37 and 1.61 μm) are sensitive to the phase and effective radius of clouds (Fig 1a). In addition to the single-channel reflectance values, dual-channel reflectance differences ($R_{\lambda1}$ minus $R_{\lambda2}$; $R_{\lambda1}$–$R_{\lambda2}$, where λ1 and λ2 represent the center wavelength of channels) are computed to capture relative spectral contrasts between channels, yielding a total of fifteen reflectance difference variables. Among the other ten channels using brightness temperature (BT) values, the shortwave infrared band (3.8 μm) is useful for detecting fire hotspots and cloud top phases. The water vapor channels (6.2, 6.9, and 7.3 μm) provide information on mid- to upper-tropospheric moisture distribution. The other infrared channels (8.6, 9.6, 10.4, 11.2, 12.4, and 13.3 μm) are essential for retrieving atmospheric and surface temperatures, and cloud-top properties. Similar to the reflectance data, dual-channel brightness temperature differences are calculated across all infrared channels, resulting in an additional forty-five BT difference variables ($BT_{\lambda1}$ minus $BT_{\lambda2}$; $BT_{\lambda1}$–$BT_{\lambda2}$). By incorporating and combining the full spectral range, this study utilizes the comprehensive information content of AMI to enhance the accuracy of the analysis. The primary applications, including the center wavelength of each single channel, along with the channel number, spectral classification (visible, near-infrared, and infrared), and channel notation, are summarized in Table 1 [4]. In addition, we employ seven Level-2 (L2) cloud-related variables—cloud top height (CTH), cloud top temperature (CTT), cloud top pressure (CTP), cloud optical thickness (COT), cloud effective radius (CER), ice water path (IWP), and liquid water path (LWP)—by processing the data from these 16 channels to analyze the physical characteristics of clouds that are directly related to precipitation estimation. Cloud phase was excluded from this analysis because it is a categorical variable rather than a discrete numerical variable.

The target data for training the ML model (Fig 1a) are used from the Integrated Multi-Satellite Retrievals for Global Precipitation Measurement (IMERG) V07B data developed by the National Aeronautics and Space Administration (NASA) (https://gpm1.gesdisc.eosdis.nasa.gov/data/GPM_L3/GPM_3IMERGHH.07/). As shown in Fig 1b, all data are collected for

Table 1. Physical characteristics of the GK2A AMI single channel observations used as input variables for the machine learning model.

Channel number	Spectral band	Central wavelength (μm)	Reflectance / Brightness temperature	Primary application
1	Visible	0.4708	$R_{0.47}$	Aerosols/ Clouds
2		0.5086	$R_{0.51}$	
3		0.6394	$R_{0.64}$	Clouds
4	Near-Infrared	0.8630	$R_{0.86}$	Vegetation
5		1.3740	$R_{1.37}$	Cirrus
6		1.6092	$R_{1.61}$	Snow/ice, cloud phase, cloud particle size
7	Infrared	3.8316	$BT_{3.8}$	Fog, low clouds, fire
8		6.2104	$BT_{6.2}$	Upper-level water vapor features
9		6.9413	$BT_{6.9}$	Midlevel water vapor features
10		7.3266	$BT_{7.3}$	Lower-level water vapor
11		8.5881	$BT_{8.6}$	Cloud-top phase
12		9.6210	$BT_{9.6}$	Ozone
13		10.3539	$BT_{10.4}$	Clouds
14		11.2288	$BT_{11.2}$	
15		12.3664	$BT_{12.4}$	
16		13.2908	$BT_{13.3}$	Carbon dioxide

This table includes channel number, spectral band classification (visible, near-infrared, and infrared), central wavelength, corresponding R or BT variables, and their primary meteorological and cloud-related applications.

https://doi.org/10.1371/journal.pclm.0000860.t001

the daytime precipitation (03:00 UTC) in the test areas (the East Asian mid-latitude: 30˚N–60˚N, 90˚E–150˚E, the tropical warm pool: 20˚S–20˚N, 90˚E–170˚E) in 2023. Since the IMERG data are available at 0.1˚ × 0.1˚ (approximately 10 km at nadir) resolutions [40], they are rescaled using the nearest-neighbor interpolation method to match the 2 km resolution of the GK2A.

Although IMERG precipitation products are widely used in satellite-based precipitation studies, they are known to show region- and regime-dependent uncertainties. Previous evaluations have shown that IMERG generally performs well in capturing the spatial and temporal variability of precipitation in both the tropics and mid-latitudes. However, uncertainties exist regarding extreme rainfall events and complex land surfaces. A comprehensive global assessment reported that IMERG demonstrates robust performance over oceanic regions, including the TWP, and reasonable agreements with ground-based observations in the East Asian mid-latitudes, supporting its applicability for large-scale precipitation analyses in these regions [41]. We also acknowledge that the current analysis focuses on 2023 due to the availability of complete, high-quality coincident GK2A and IMERG datasets for that year.

2.2. Machine learning model

This study employs and optimizes a RF model (Fig 1c) for precipitation estimation [42]. RF models have been widely used in previous environmental and remote sensing studies due to their robustness to nonlinear relationships, capability to handle high-dimensional input variables, and relatively stable performance without extensive tuning [5,9,16]. A preliminary test in preparation of this study showed that the RF model yields the most balanced estimation performance among various ML models employed in previous studies (S1 Fig, S2 Fig), albeit small inter-model differences [6,14,19]. Based on these results and the widespread adoption of RF models in similar applications, an RF model is selected as a representative ML model, while the focus of this study remains on the role of ML-based feature selection rather than model-specific optimization.

The SHAP model (Fig 1c) is used to evaluate the contribution and priority of various input variables [43]. SHAP quantifies a consistent and interpretable measure of feature influence by decomposing model predictions into additive contributions of individual input variables. The variables with positive SHAP values increase the model's prediction, while negative values indicate a decrease. The colors in SHAP graph represent the magnitude of the values, with red (blue) indicating high (low) values. By combining RF with SHAP analysis, this study uses the predictive capability of ML while maintaining physical interpretability [19,44,45]. Finally, it enables efficient feature selections and stable precipitation retrievals from satellite data with high accuracy.

2.3. Validation metrics

To evaluate the model performance (Fig 1c), this study employs a widely-used validation metrics including the probability of detection (POD), false alarm ratio (FAR), and critical success index (CSI) [46], derived from True Positive (TP: correctly predicted events), True Negative (TN: correctly predicted non-events), False Positive (FP: incorrectly predicted events), and False Negative (FN: missed events) values. These validation metrics are calculated as:

$$\mathrm{POD} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{1}$$

$$\mathrm{FAR} = \frac{\mathrm{FP}}{\mathrm{TP} + \mathrm{FP}}, \tag{2}$$

$$\mathrm{CSI} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}. \tag{3}$$

Defined in this way, POD reflects the proportion of correctly predicted events, while FAR indicates the frequency of false alarms. In addition, CSI shows a measure of the model's overall accuracy by balancing TP, FP, and FN. All indices vary between 0 and 1; larger values indicate higher accuracy for POD and CSI, while smaller values indicate higher accuracy for FAR.

3. Results

3.1. Key variables for precipitation in East Asian mid-latitude region

The SHAP model is employed in conjunction with the RF model to explore the key factors that influence precipitation estimation. The SHAP importance rankings represent the physical mechanisms that govern precipitation formations. Note that the key variables must correspond to the fundamental physical properties of clouds. The SHAP importance, calculated as $\frac{|SHAP_i|}{\sum_i |SHAP_i|}$ for the variable i, is used to quantify the extent to which each variable influences precipitation estimation in the model.

Fig 2 displays the SHAP values illustrating how different satellite-derived variables influence light-to-moderate (Fig 2a) and heavy (Fig 2b) precipitation amounts over East Asian mid-latitude region. Light-to-moderate precipitation in this region is strongly affected by both the IWP and spectral reflectance differences involving the 1.61 μm channel. The importance of IWP represents the critical role of ice-phase processes in precipitation formations, as ice water content is a key indicator of cloud glaciation and the availability of condensed water for precipitation formation [47]. Larger IWP values indicate large ice particle concentrations which can grow rapidly through vapor deposition, aggregation and collection, to form precipitation particles. The spectral reflectance differences involving the 1.61 μm channel include the reflectance differences between the 1.61 μm channel ($R_{1.61}$) and multiple visible channels ($R_{visible;}$ $R_{0.47}$, $R_{0.51}$, and $R_{0.64}$). At 1.61 μm, where absorption by the liquid and ice cloud water is large, reflectance decreases with increasing particle size. Thus, microphysically mature clouds characterized by larger droplets and enhanced ice-phase developments tend to exhibit low reflectance [48,49]. Hence, large positive differences between $R_{1.61}$ and the visible channels (i.e., $R_{visible}-R_{1.61}$) indicate the presence of thick clouds of large ice contents and/or droplet populations.

These results demonstrate the limitation of relying solely on thermodynamic indicators for precipitation estimation, and emphasize the importance of incorporating cloud information. Traditional thermodynamic variables, such as infrared BT or CTT, primarily reflect cloud-top properties and without unique representation of precipitation processes. While low BT or CTT values indicate high cloud tops, these signatures are not exclusive to precipitating cumulonimbus systems; thin cirrus layers and anvil clouds also exhibit similarly low CTT without producing surface precipitation. In contrast, the 1.61 μm channel is sensitive to the size distribution and thermodynamic phase of cloud particles, and its reflectance is also influenced by the cloud water path. Thus, it can distinguish precipitation-capable clouds from non-precipitating high altitude clouds.

For heavy precipitation in this region, $R_{1.37}$ and the difference between $R_{1.37}$ and $R_{1.61}$ ($R_{1.37}-R_{1.61}$) are the most important predictors (Fig 2b). The $R_{1.37}-R_{1.61}$ shows both negative and positive values in cloudy regions. Note that the locations of large positive $R_{1.37}-R_{1.61}$ values correspond to heavy rainfall areas. This relationship stems from the detection of ice particles embedded within the cirrus layer [50]. The $R_{1.37}$ channel serves as a cirrus-detection band because it is sensitive to the radiation reflected from the high-altitude clouds. While cirrus clouds themselves rarely produce precipitation directly, it plays a crucial role as an indicator of heavy convective precipitation within the synoptic context of this region. In the frontal and baroclinic systems, one of characteristic East Asian weather patterns, extensive cirriform layers frequently overlie deep convective systems within the warm conveyor belt [26]. This overlying cirrus indicates vigorous upward motion and intense moisture transport beneath, both intrinsically linked to heavy rainfall. In addition, the BT difference $BT_{6.2}-BT_{9.6}$, which characterizes cloud-top properties and upper tropospheric moisture [51–53], is another crucial predictor. This variable provides constraints for identifying meteorologically active environments with moist environment and strong dynamical forcing.

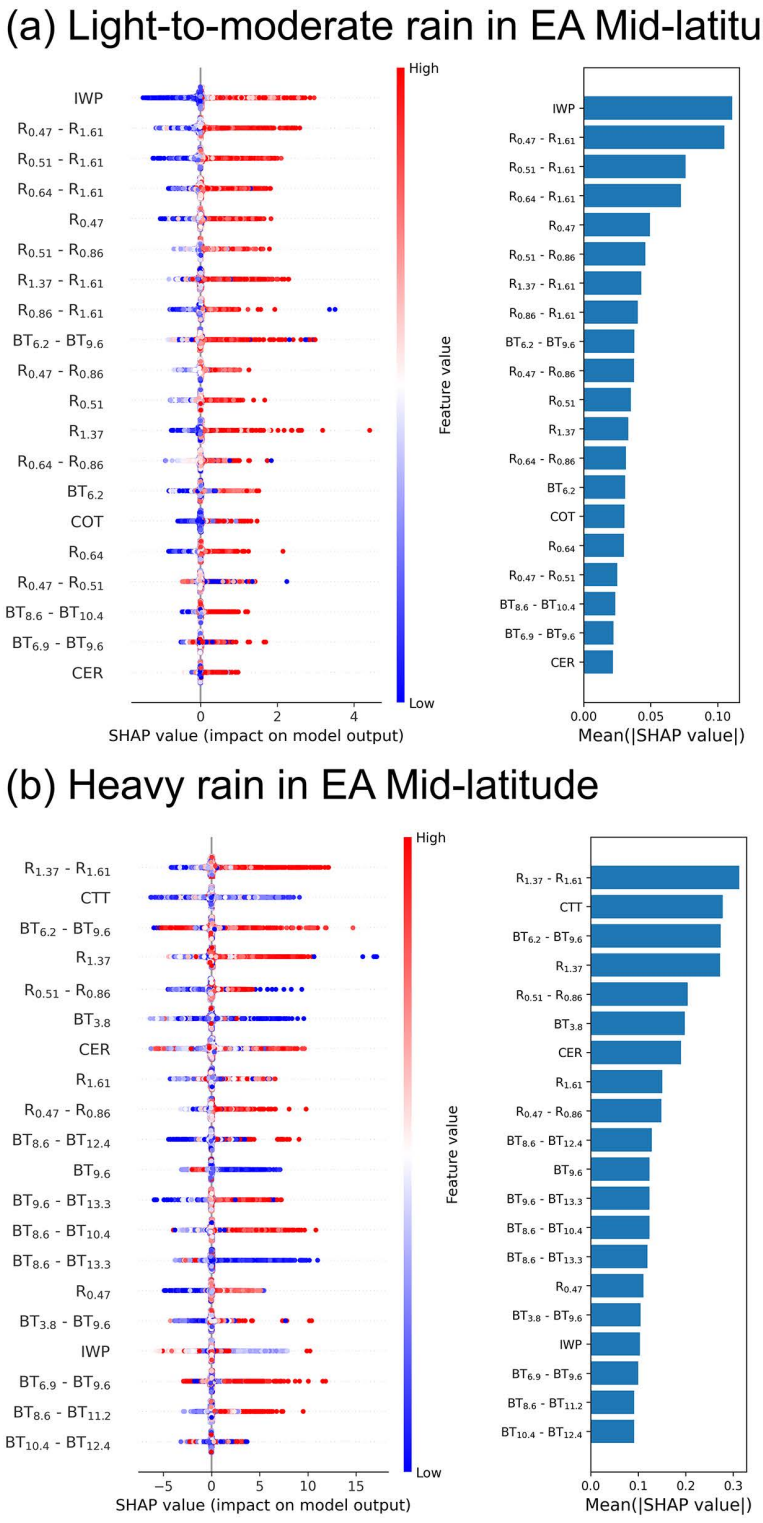


Fig 2. The contribution of each variable to the precipitation estimation in the East Asian mid-latitude region demonstrated by SHAP value. The importance of variables is ranked by the average absolute SHAP value of the variables. Only the top 20 variables with high contributions are shown. The distribution of SHAP values for light-to-moderate rain estimation **(a)** and heavy rain estimation **(b)**.

https://doi.org/10.1371/journal.pclm.0000860.g002

Fig 3 shows the spatial distribution of the key geostationary satellite variables and precipitation in the East Asian mid-latitude region, where IWP (Fig 3a) has the strongest influence on estimating light-to-moderate rainfall, and the $R_{1.37}-R_{1.61}$ values (Fig 3b) serve as the primary predictor for the heavy precipitation events. The distribution of these variables and the observed rainfall areas (Fig 3c) show reasonable correspondences; for the overall rainfall and the heavy events, the precision values are 0.57 and 0.40 and the recall scores are 0.50 and 0.40, respectively. This regional consistency indicates that the variables which contribute most to the precipitation estimation in the ML model are statistically significant and physically consistent, reflecting their connection to the physical and dynamical processes that govern precipitation formation. In particular, $R_{1.37}-R_{1.61}$ demonstrates exceptional specificity for detecting heavy rainfalls, as large positive values occur solely in the areas of intense precipitation. These results confirm that the spatial patterns of IWP and $R_{1.37}-R_{1.61}$ align closely with the actual distributions of light-to-moderate and heavy rainfalls across the East Asian mid-latitudes, thus validating their physical relevance and predictive capability for the two different precipitation intensities.

3.2. Key variables for the TWP region

Fig 4 presents the SHAP values for the retrieval of light-to-moderate (Fig 4a) and heavy rain (Fig 4b) over the TWP region. IWP is substantially more important predictor than other variables for the light-to-moderate rainfall (Fig 4a), similar to the East Asian mid-latitude region (Fig 2a), likely due to abundant tropical moisture supplies promoting deep convections and ice particle formations within clouds [54]. The second highest-ranked predictor is the COT (Fig 4a), which is more important in TWP than in the East Asian mid-latitudes. This regional contrast reflects the differences in regional environmental characteristics, as well-developed deep convective clouds with large COT that are associated with increased likelihood of precipitation occurs more frequently in the TWP than in the mid-latitudes [55]. Overall, COT tends to be more relevant for light-to-moderate precipitation in the TWP, reflecting variations in convective cloud development, while its relative contribution decreases for heavy rainfall associated with already well-developed convective clouds. Fig 4a also shows that the reflectance differences between visible and near-infrared channels, $R_{0.51}-R_{0.86}$, $R_{0.47}-R_{0.86}$, $R_{0.64}-R_{1.61}$, and $R_{0.47}-R_{1.61}$, which are related to mature cloud structures, follow IWP and COT similarly as those found for the light-to-moderate precipitation in the East Asian mid-latitude region (Fig 2a).

For heavy precipitation estimations in the TWP region, the differences between the visible channels ($R_{0.51}$ and $R_{0.47}$) and the near-infrared channel ($R_{0.86}$), i.e., $R_{0.51}-R_{0.86}$ and $R_{0.47}-R_{0.86}$, which represent the surface reflectance contrast such as land-ocean contrast, are the most important predictors (Fig 4b). These results are consistent with the well-established understanding that intense rainfall events in the TWP originate over the ocean where sea surface temperatures are sufficiently high. While the surface reflectance contrast is expected to be a dominant driver of intense precipitation in the TWP, the emergence of $R_{1.37}-R_{1.61}$ as a significant predictor highlights a common physical mechanism in the East Asian mid-latitude region. As discussed for the East Asian mid-latitudes (Fig 3b), this spectral difference demonstrates the presence of extensive cirriform clouds associated with deep convective systems. This similarly indicates the occurrence of vigorous upward motions and enhanced moisture transports underlying heavy precipitation events in the TWP. In addition, CTT derived from multi-band BT data is more directly linked to vertical cloud development. Together with CER, it shows strong associations with deep convection and intense precipitation. While BT from individual channels shows wavelength-specific information, CTT integrates BT over multiple spectral bands to estimate the actual cloud-top temperature, providing a more direct relationship with convection intensity. Lower CTTs in the TWP are generally related to stronger updrafts and deeper convection, which result in enhanced rainfall [33].

Fig 5 shows the spatial distribution of the key variables and precipitation in the TWP region. The spatial distribution of IWP (Fig 5a), which has the strongest influence on the light-to-moderate rainfall estimation, closely matches the precipitation pattern shown in Fig 5c, with precision of 0.61 and recall of 0.55. The close correspondence further confirms that IWP is the most important predictor for retrieving light-to-moderate rainfall in the TWP region. Fig 5b illustrates the spatial distribution of $R_{0.51}-R_{0.86}$, which is identified as the primary predictor of the heavy rainfall. Given that intense precipitation

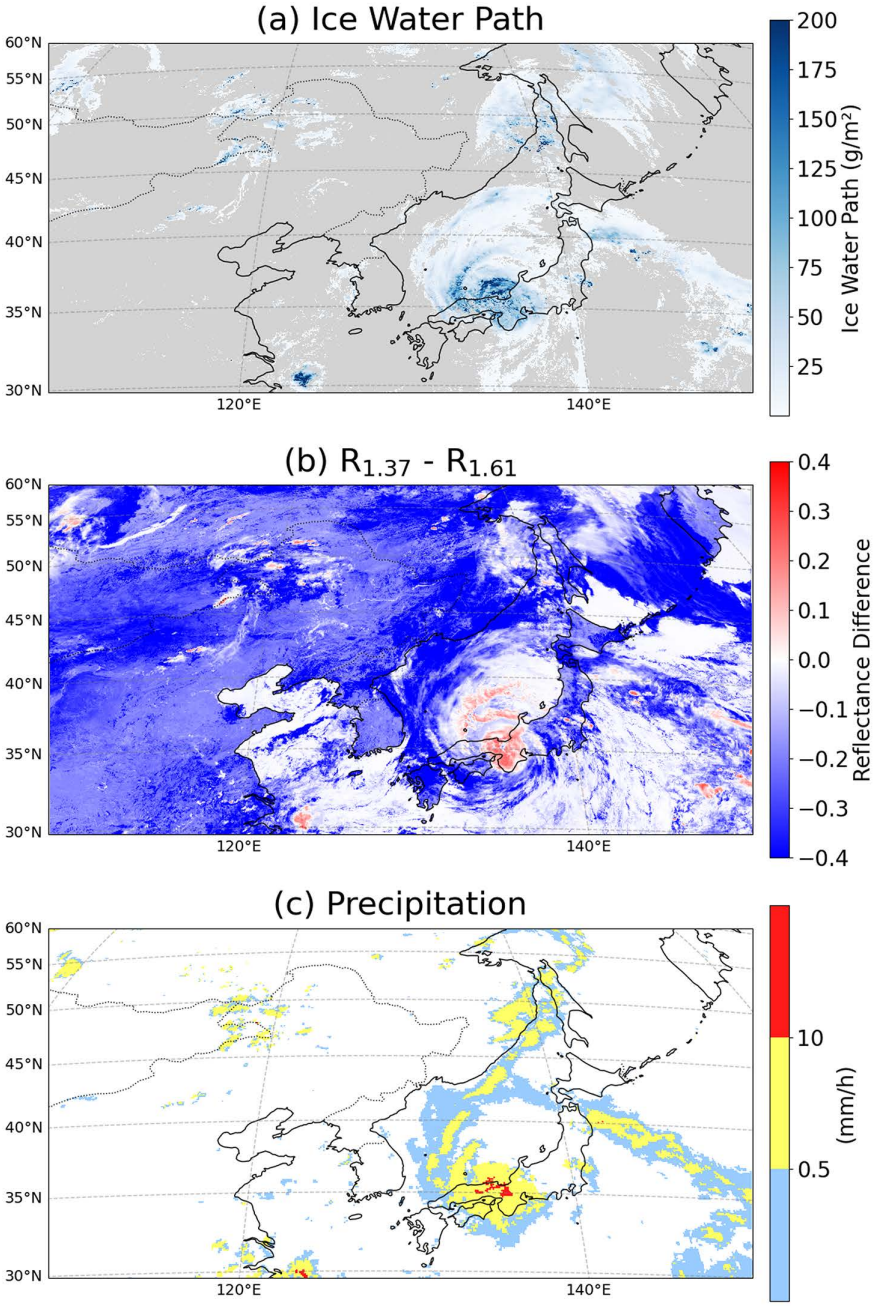


Fig 3. Spatial distribution of variables and precipitation distribution in the East Asian mid-latitude region. An arbitrarily selected case on August 15, 2023 at 03:00 UTC. **(a)** Spatial distribution of IWP, the most influential variable for light-to-moderate rain estimation. **(b)** Spatial distribution of the reflectance difference between 1.37 μm and 1.61 μm, the most influential variable for heavy rain estimation. **(c)** Distribution of target precipitation. Base map coastlines were generated using the GSHHS dataset (https://www.ngdc.noaa.gov/mgg/shorelines/gshhs.html), which is distributed under the GNU LGPL (https://www.gnu.org/licenses/lgpl-3.0.html).

https://doi.org/10.1371/journal.pclm.0000860.g003

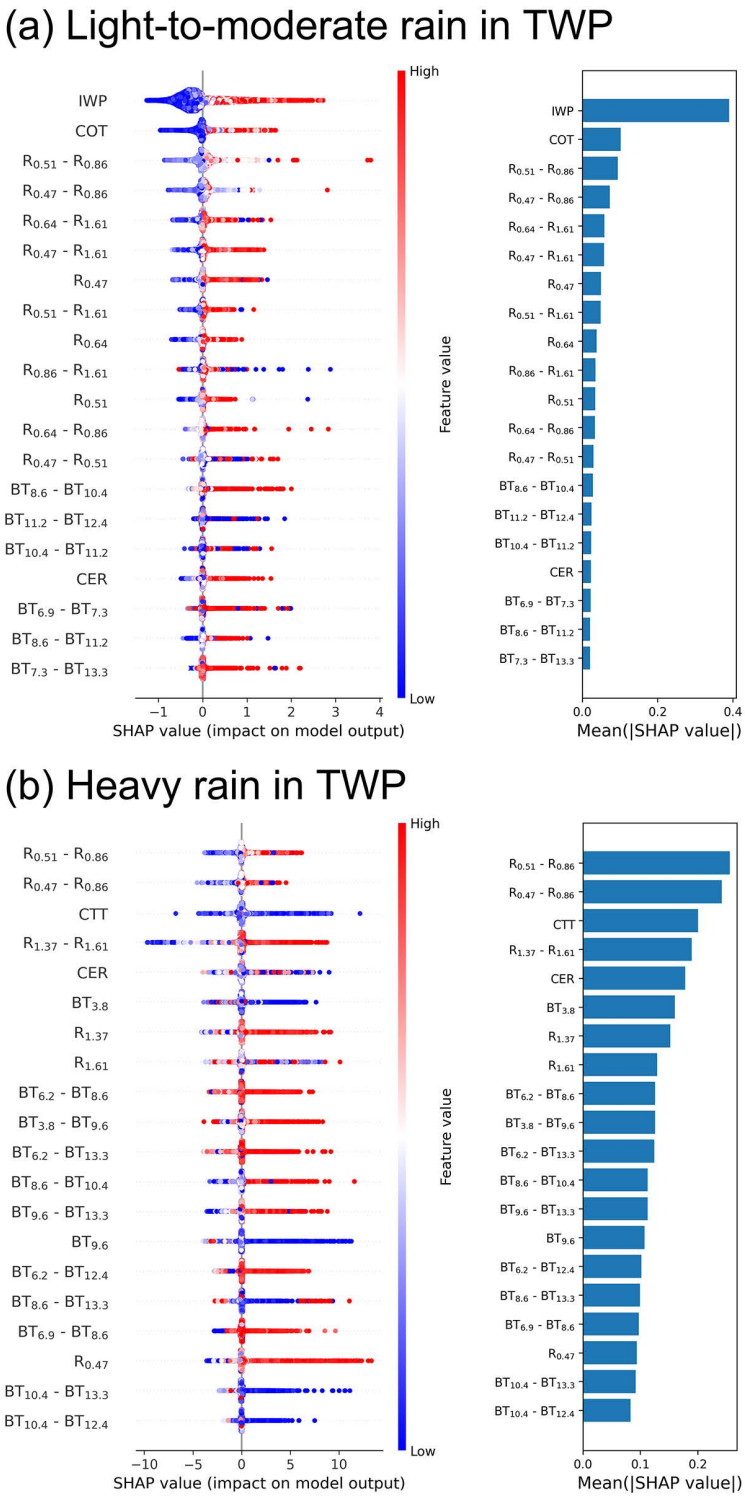


Fig 4. The same as Fig 2 except for the tropical warm pool region.

https://doi.org/10.1371/journal.pclm.0000860.g004

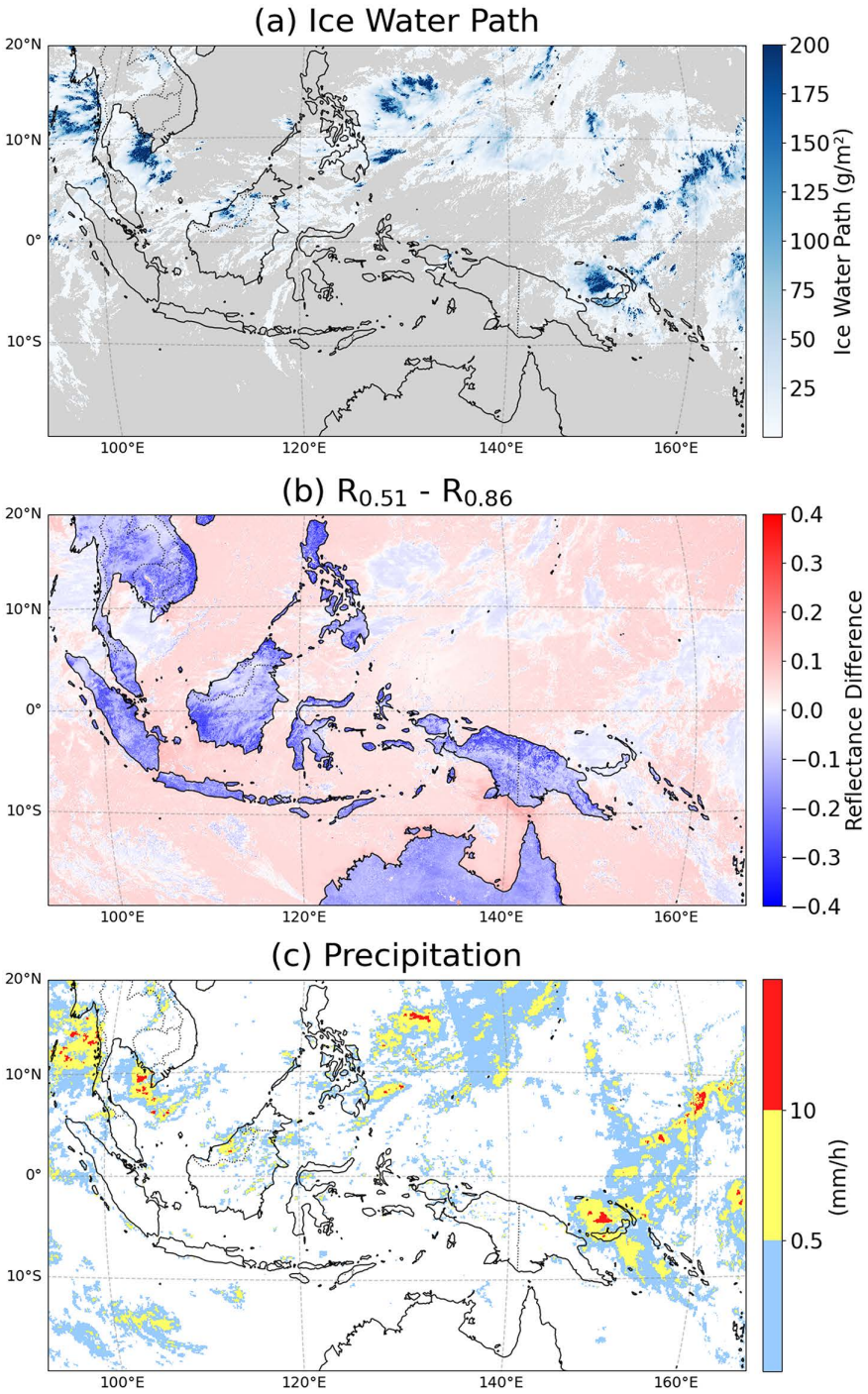


Fig 5. The same as Fig 3 except for the tropical warm pool region. Base map coastlines were generated using the GSHHS dataset (https://www.ngdc.noaa.gov/mgg/shorelines/gshhs.html), which is distributed under the GNU LGPL (https://www.gnu.org/licenses/lgpl-3.0.html).

https://doi.org/10.1371/journal.pclm.0000860.g005

in the TWP occurs predominantly over warm ocean areas, the variable effectively captures land-ocean contrast and thus attains high predictive importance.

3.3. Integrated interpretation of results

Precipitation results from various physical and dynamical processes associated with clouds. In this section, we integrate the results using SHAP-based interpretation to identify key satellite-derived variables and elucidate their physical links to precipitation formation. By determining the most important predictors according to regions and intensities, the results clarify the characteristics of clouds and the environmental factors that are most closely linked to regional precipitation. For the light-to-moderate precipitation, the East Asian mid-latitude and TWP region show similar key predictors (Fig 2a and Fig 4a), with IWP representing the contents of ice particles in clouds, as the most important variable. In contrast, the primary predictors for intense rainfall differ significantly between these regions: high-altitude cloud properties dominate in the East Asian mid-latitudes (Fig 2b), while the surface reflectance contrast is crucial for TWP (Fig 4b).

In the East Asian mid-latitudes, the primary variables are $R_{1.37}$ and $R_{1.37} - R_{1.61}$, indicating the presence of extensive cirriform layers over deep convection within the warm conveyor belt. For TWP, the dominant predictors are the reflectance differences between the visible and near-infrared channels, $R_{0.51} - R_{0.86}$ and $R_{0.47} - R_{0.86}$, which effectively distinguish the land from the ocean. The reflectance differences also indicate environmental contrasts in modulating convection and reflect that heavy precipitation in TWP occurs predominantly over warm oceanic surfaces. Although cirrus-related $R_{1.37}$ and $R_{1.61}$ are significant for intense precipitation in both regions, these are less important in TWP than in the East Asian mid-latitudes due to the overwhelming influence of warm sea surface temperatures. These contrasting predictors indicate the fundamental differences in the underlying precipitation mechanisms between the two regions.

4. Discussion

Increasing availability of various high-resolution satellite data allows us to develop more sophisticated precipitation estimation models that can achieve higher retrieval accuracy than legacy model. By exploiting wide spectrum of information from geostationary satellites, in conjunction with the feature importance rankings determined in this study, it is possible to construct more effective and reliable algorithms for retrieving precipitation in various regions. Importantly, the use of SHAP analysis enables XAI to explicitly quantify the contribution of individual satellite-derived variables to the ML-based precipitation estimation. The interpretability allows physically-meaningful assessments of the physical relevance of a specific variable in estimating precipitation by mitigating the limits in black-box nature of ML models.

Based on the SHAP analysis, the predictor variables are incorporated incrementally in the descending order of importance to determine the optimal combination for obtaining an accurate yet compact precipitation estimation model (Fig 1c). As seen in Fig 6, model performance for both the light-to-moderate and heavy precipitation is improved noticeably as additional variables are included, but gradually approaches saturation beyond which additional predictors yield minimal improvements for both regions. This result indicates the importance of achieving optimal balance between model complexity and prediction accuracy for operational applications which require both the computational efficiency and real-time performance are important.

In the East Asian mid-latitude region shown in Fig 6a, the performance indices, including POD, FAR, and CSI, are saturated at around 25–30 variables for light-to-moderate rain. Most of these variables are derived from visible-channel reflectance values as well as the differences between individual reflectance values, with the difference between $R_{0.86}$ and $R_{1.61}$ showing particularly high importance. Among single-channel infrared brightness temperatures, $BT_{6.2}$ which represents upper-level water vapor, ranks highest, followed by brightness temperature differences that capture cloud altitude and atmospheric conditions. Among the L2 variables, only IWP and COT contribute meaningfully, with COT being less important than the visible reflectance values. For heavy rain in the East Asian mid-latitude region, high-precision estimation is achieved using only 8 highest-ranked variables, including $R_{1.37}$, $R_{1.61}$, and $BT_{3.8}$. These channels provide information on

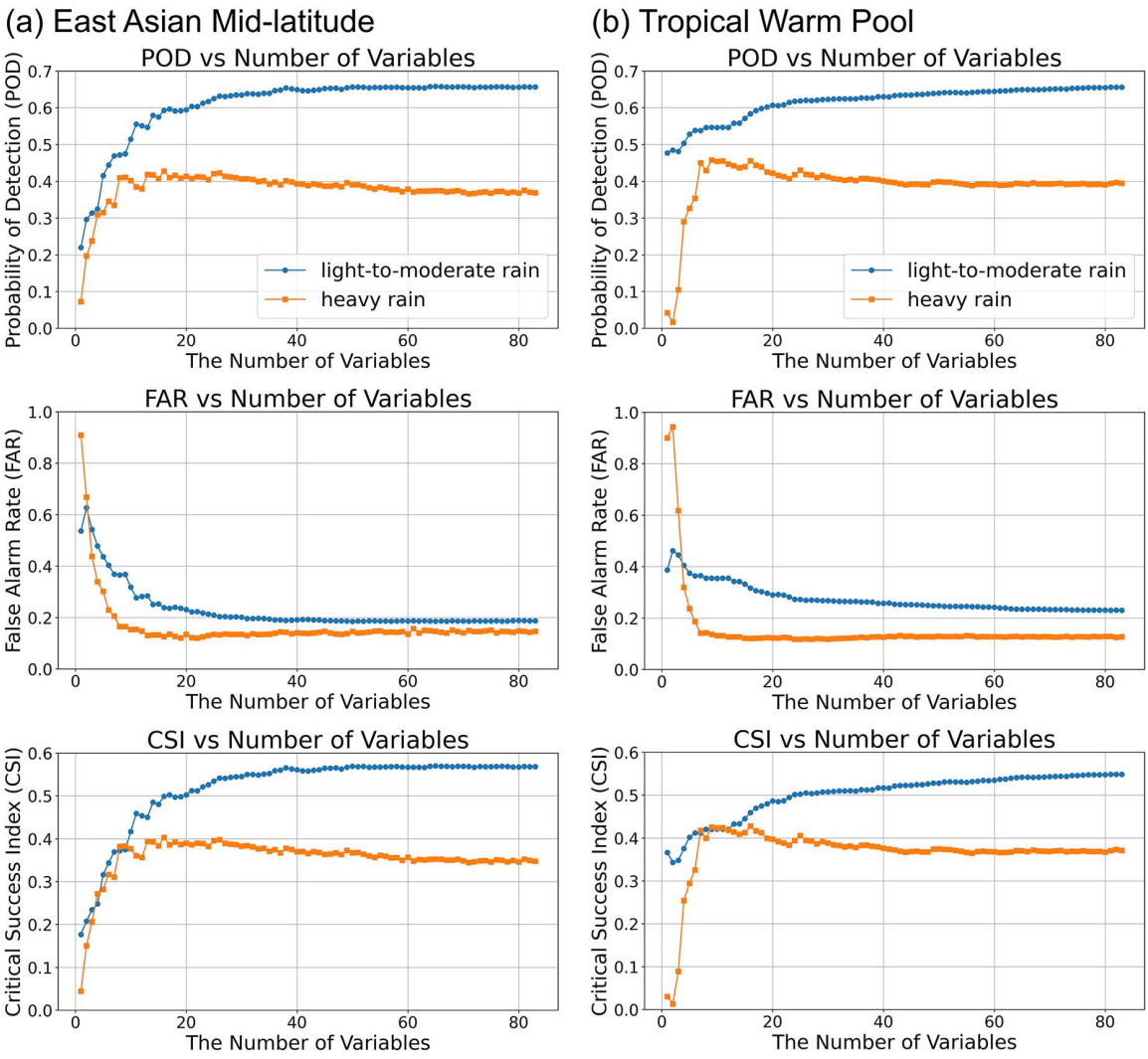


Fig 6. Accuracy graph of the model trained cumulatively in order of feature importance. Graphs showing how the model's POD, FAR, and CSI change as it trains by sequentially accumulating the most important features from geostationary satellite data. **(a)** and **(b)** show results for the East Asian mid-latitude and TWP regions, respectively.

https://doi.org/10.1371/journal.pclm.0000860.g006

cirrus ($R_{1.37}$), cloud phase and particle size ($R_{1.61}$), and low clouds ($BT_{3.8}$,). In addition, combined indices capture broader atmospheric properties: $R_{1.37} - R_{1.61}$ for high-altitude clouds, $BT_{6.2} - BT_{9.6}$ for upper-tropospheric humidity, $R_{0.51} - R_{0.86}$ for surface reflectance contrast, and CTT and CER further enhancing heavy rain estimation.

Similarly, in the TWP region shown in Fig 6b, optimal performance is achieved with about 20 and 8 variables for the light-to-moderate rain and the heavy rain, respectively. The variables needed for the light-to-moderate rain estimation in the TWP mostly correspond to those in the East Asian mid-latitude region; however, the L2 data, particularly IWP and COT, are more prominent. For the TWP heavy rain estimation, high accuracy is ensured with much fewer variables than for the light-to-moderate precipitation estimation, similarly as in the East Asian mid-latitude region. Variables such as $R_{0.51} - R_{0.86}$ and $R_{0.47} - R_{0.86}$ used for this purpose are effective in detecting tropical heavy rainfall over the ocean. Subsequent key variables exhibit sensitivity to upper-level clouds, enabling reliable detection of optically thick clouds commonly associated with precipitation.

The present results demonstrate that selecting only the most relevant features can achieve the highest classification accuracy while maintaining a relatively simple model structure. This finding is expected to contribute to improving the computational efficiency of real-time satellite-based precipitation estimates, with important implications in developing high-resolution precipitation estimation algorithms on the basis of the geostationary satellite data for real-time applications. This feature selection can be advanced in future studies by incorporating more systematic ML strategies, such as automated hyperparameter optimization and model selection within RF model–based approaches [56]. These extensions would help to accommodate the increasing volume and complexity of satellite observations and to enhance model robustness. Ultimately, such efforts are expected to contribute to the development of scalable and efficient precipitation estimation systems using geostationary data for real-time applications [57].

This study is limited to the East Asian mid-latitudes and the TWP because of the field of view of GK2A; however, results in the study can be applied to other regions covered by different geostationary satellites. All geostationary satellites provide similar atmospheric radiation information, especially those using the imagers of the absorption channels similar to GK2A, for example, the Advanced Baseline Imager on the United States Geostationary Operational Environmental Satellite (GOES-16/17/18) [58], the Flexible Combined Imager on the Meteosat Third Generation [59], and the Advanced Himawari Imager on Japan's Himawari-8/9 satellites [2]. Therefore, the experimental method developed in this study can be applied to the areas beyond the field of view of GK2A by using other geostationary satellites.

Supporting information

S1 Fig. Comparison of precipitation retrieval performance across 12 machine learning models for heavy rain (upper row) and light-to-moderate rain (lower row) in the East Asian mid-latitude region. Each boxplot summarizes the distribution of POD, FAR, and CSI computed from daily evaluations over the year 2023 at 0300 UTC. For heavy rain, most models exhibit low POD, with NB showing exceptionally high POD but at the cost of near-perfect FAR, indicating severe overprediction. RF (highlighted in red) achieves the lowest FAR among all models, yielding competitive CSI despite its moderate POD. For light-to-moderate rain, RF demonstrates the best overall performance with the highest POD, the lowest FAR, and the highest CSI.
(TIF)

S2 Fig. Same as S1 Fig, but for the tropical warm pool region. A consistent pattern is observed across both regions, with RF maintaining the most balanced detection skill for both rain intensity categories, supporting its selection as the optimal model in this study.
(TIF)

Author contributions

Conceptualization: Hwayon Choi, Yong-Sang Choi.

Data curation: Hwayon Choi.

Formal analysis: Hwayon Choi.

Funding acquisition: Yong-Sang Choi.

Methodology: Hwayon Choi.

Resources: Hwayon Choi.

Software: Hwayon Choi.

Supervision: Yong-Sang Choi.

Validation: Hwayon Choi.

Visualization: Hwayon Choi.

Writing – original draft: Hwayon Choi.

Writing – review & editing: Yong-Sang Choi, Chang-Hoi Ho, Jinwon Kim.

References

1. Ouaknine J, Gode S, Napierala B, Viard T, Foerster U, Fray S. MTG flexible combined imager optical design and performances. In: Earth observing systens XVIII. 2013.
2. Da C. Preliminary assessment of the Advanced Himawari Imager (AHI) measurement onboard Himawari-8 geostationary satellite. Remote Sens Letters. 2015;6(8):637–46.
3. Schmit TJ, Lindstrom SS, Gerth JJ, Gunshor MM. Applications of the 16 spectral bands on the Advanced Baseline Imager (ABI). 2018.
4. Kim D, Gu M, Oh T-H, Kim E-K, Yang H-J. Introduction of the advanced meteorological imager of Geo-Kompsat-2a: In-orbit tests and performance validation. Remote Sensing. 2021;13(7):1303.
5. Kühnlein M, Appelhans T, Thies B, Nauss T. Improving the accuracy of rainfall rates from optical satellite sensors with machine learning—A random forests-based approach applied to MSG SEVIRI. Remote Sens Environ. 2014;141:129–43.
6. Meyer H, Kühnlein M, Appelhans T, Nauss T. Comparison of four machine learning algorithms for their applicability in satellite-based optical rainfall retrievals. Atmosph Res. 2016;169:424–33.
7. Tebbi MA, Haddad B. Artificial intelligence systems for rainy areas detection and convective cells' delineation for the south shore of Mediterranean Sea during day and nighttime using MSG satellite images. Atmosph Res. 2016;178:380–92.
8. Mohia Y, Ameur S, Lazri M, Brucker J. Combination of spectral and textural features in the MSG satellite remote sensing images for classifying rainy area into Different classes. J Indian Soc Remote Sens. 2017;45(5):759–71.
9. Hirose H, Shige S, Yamamoto MK, Higuchi A. High temporal rainfall estimations from Himawari-8 multiband observations using the random-forest machine-learning method. Journal of the Meteorological Society of Japan Ser II. 2019;97(3):689–710.
10. Liu Q, Li Y, Yu M, Chiu LS, Hao X, Duffy DQ, et al. Daytime rainy cloud detection and convective precipitation delineation based on a deep neural Network method using GOES-16 ABI images. Remote Sens. 2019;11(21):2555.
11. Moraux A, Dewitte S, Cornelis B, Munteanu A. Deep learning for precipitation estimation from satellite and rain gauges measurements. Remote Sens. 2019;11(21):2463.
12. Sadeghi M, Asanjan AA, Faridzad M, Nguyen P, Hsu K, Sorooshian S, et al. PERSIANN-CNN: precipitation estimation from remotely sensed information using artificial neural networks–convolutional neural networks. J Hydrometeorol. 2019;20(12):2273–89.
13. Lazri M, Labadi K, Brucker JM, Ameur S. Improving satellite rainfall estimation from MSG data in Northern Algeria by using a multi-classifier model based on machine learning. Journal of Hydrology. 2020;584:124705.
14. Zhang Z, Wang D, Qiu J, Zhu J, Wang T. Machine learning approaches for improving near-real-time IMERG rainfall estimates by integrating cloud properties from NOAA CDR PATMOS-x. J Hydrometeorol. 2021;22(10):2767–81.
15. Gao Y, Guan J, Zhang F, Wang X, Long Z. Attention-unet-based near-real-time precipitation estimation from fengyun-4A satellite imageries. Remote Sens. 2022;14(12):2925.
16. Kumah KK, Maathuis BH, Hoedjes J, Su Z. Near real-time estimation of high spatiotemporal resolution rainfall from cloud top properties of the MSG satellite and commercial microwave link rainfall intensities. Atmosph Res. 2022;279:106357.
17. Afzali GV, Petković V, Arulraj M, Nguyen P, Hsu K-l, Sorooshian S, et al. Integrating LEO and GEO observations: toward optimal summertime satellite precipitation retrieval. J Hydrometeorol. 2023;24(11):1939–54.
18. Berthomier L, Perier L. Espresso: a global deep learning model to estimate precipitation from satellite observations. Meteorology. 2023;2(4):421–44.
19. Lin X, Fan J, Hou ZJ, Wang J. Machine learning of key variables impacting extreme precipitation in various regions of the contiguous United States. J Adv Model Earth Syst. 2023;15(3):e2022MS003334.
20. Lyu Y, Yong B. Using an explainable machine learning approach to produce high-resolution hourly precipitation estimates for a typical data-deficiency basin. Journal of Geophysical Research: Machine Learn Comput. 2025;2(1):e2024JH000489.
21. Mosaffa H, Ciabatta L, Filippucci P, Sadeghi M, Brocca L. HR-PrecipNet: a machine learning framework for 1-km high-resolution satellite precipitation estimation. J Hydrol. 2025;658:133217.
22. Blum AL, Langley P. Selection of relevant features and examples in machine learning. Artificial intelligence. 1997;97(1-2):245–71.
23. Guyon I, Eliseeff A. An introduction to variable and feature selection. J Mach Learn Res. 2003;3(Mar):1157–82.
24. Liu H, Motoda H. Feature selection for knowledge discovery and data mining. Springer Science & Business Media; 2012.
25. Cotton WR, Anthes RA. Storm and cloud dynamics. Academic Press; 1992.
26. Houze RA. Cloud dynamics. Academic Press; 2014.

27. Takemi T. Convection and precipitation under various stability and shear conditions: Squall lines in tropical versus midlatitude environment. Atmosph Res. 2014;142:111–23.
28. Rosenfeld D, Gutman G. Retrieving microphysical properties near the tops of potential rain clouds by multispectral analysis of AVHRR data. Atmosph Res. 1994;34(1-4):259–83.
29. Stephens GL, Kummerow CD. The remote sensing of clouds and precipitation from space: a review. J Atmosph Sci. 2007;64(11):3742–65.
30. Rapp AD. Observational evidence linking precipitation and mesoscale cloud fraction in the southeast Pacific. Geophy Res Lett. 2016;43(13):7267–73.
31. Sun J, Shi Z, Chai J, Xu G, Niu B. Effects of mixed phase microphysical process on precipitation in a simulated convective cloud. Atmosphere. 2016;7(8):97.
32. Liu Y, Luo R, Zhu Q, Hua S, Wang B. Cloud ability to produce precipitation over arid and semiarid regions of Central and East Asia. Inter J Climatol. 2020;40(3):1824–37.
33. Braga RC, Rosenfeld D, Krüger OO, Ervens B, Holanda BA, Wendisch M, et al. Linear relationship between effective radius and precipitation water content near the top of convective clouds: measurement results from ACRIDICON–CHUVA campaign. Atmosph Chem Phy. 2021;21(18):14079–88.
34. Murakami Y, Kummerow CD, van den Heever SC. On the Relation among Satellite-Observed Liquid Water Path, Cloud Droplet Number Concentration, and Cloud-Base Rain Rate and Its Implication to the Autoconversion Parameterization in Stratocumulus Clouds. J Climate. 2021;34(20):8165–80.
35. Zhao P, Xiao H, Liu J, Zhou Y. Precipitation efficiency of cloud and its influencing factors over the Tibetan plateau. Inter J Climatol. 2022;42(1):416–34.
36. Tang L, Gao W, Xue L, Zhang G, Guo J. Climatological characteristics of hydrometeors in precipitating clouds over eastern China and their relationship with precipitation based on ERA5 reanalysis. J Appl Meteorol Climatol. 2023;62(5):625–41.
37. Belgiu M, Drăguţ L. Random forest in remote sensing: A review of applications and future directions. ISPRS journal of photogrammetry and remote sensing. 2016;114:24–31.
38. Oke T. Guide to instruments and methods of observation. Geneva, Switzerland: WMO; 2018.
39. Choi Y-S, Ho C-H. Earth and environmental remote sensing community in South Korea: a review. Remote Sensing Applications: Soc Environ. 2015;2:66–76.
40. Huffman GJ, Bolvin DT, Braithwaite D, Hsu K, Joyce R, Xie P. NASA global precipitation measurement (GPM) integrated multi-satellite retrievals for GPM (IMERG). Algorit Theoret Basis Doc (ATBD). 2015;4(26):30.
41. Pradhan RK, Markonis Y, Godoy MRV, Villalba-Pradas A, Andreadis KM, Nikolopoulos EI, et al. Review of GPM IMERG performance: a global perspective. Remote Sens Environ. 2022;268:112754.
42. Breiman L. Random forests. Machine learning. 2001;45:5–32.
43. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inform Process Syst. 2017;30.
44. He Z, Yang Y, Fang R, Zhou S, Zhao W, Bai Y, et al. Integration of shapley additive explanations with random forest model for quantitative precipitation estimation of mesoscale convective systems. Front Environ Sci. 2023;10:1057081.
45. Chen C, Liu Y, Li Y, Chen D. Explainable artificial intelligence framework for urban global digital elevation model correction based on the SHapley additive explanation-random forest algorithm considering spatial heterogeneity and factor optimization. Inter J Appl Earth Observ Geoinform. 2024;129:103843.
46. Doswell C, Davies-Jones R, Keller DL. On summary measures of skill in rare event forecasting based on contingency tables. Weather Forecast. 1990;5(4):576–85.
47. Tubul Y, Koren I, Altaratz O, Heiblum RH. On the link between precipitation and the ice water path over tropical and mid-latitude regimes as derived from satellite observations. Atmosph Measure Techniq Discuss. 2017;2017:1–16.
48. Setvák M, Rabin RM, Doswell III CA, Levizzani V. Satellite observations of convective storm tops in the 1.6, 3.7 and 3.9 μm spectral bands. Atmosph Res. 2003;67:607–27.
49. Acarreta J, Stammes P, Knap W. First retrieval of cloud phase from SCIAMACHY spectra around 1.6 μm. Atmospheric research. 2004;72(1-4):89–105.
50. Choi YS, Ho CH, Ahn MH, Kim YS. Enhancement of the consistency of MODIS thin cirrus with cloud phase by adding 1.6 μm reflectance. Inter J Remote Sens. 2005;26(21):4669–80.
51. Fritz S, Laszlo I. Detection of water vapor in the stratosphere over very high clouds in the tropics. J Geophy Res Atmosph. 1993;98(D12):22959–67.
52. Schmetz J, Tjemkes S, Gube M, Van de Berg L. Monitoring deep convection and convective overshooting with METEOSAT. Adv Space Res. 1997;19(3):433–41.
53. Otkin JA, Greenwald TJ, Sieglaff J, Huang H-L. Validation of a large-scale simulated brightness temperature dataset using SEVIRI satellite observations. J Appl Meteorol Climatol. 2009;48(8):1613–26.

54. Del Genio AD, Kovari W. Climatic properties of tropical precipitating convection under varying environmental conditions. J Climate. 2002;15(18):2597–615.
55. Harrop BE, Hartmann DL. The relationship between atmospheric convective radiative effect and net energy transport in the tropical warm pool. J Climate. 2015;28(21):8620–33.
56. Addula SR, Sajja GS. Automated machine learning to streamline data-driven industrial application development. In: Second International conference computational and characterization techniques in engineering sciences (IC3TES). 2024.
57. Devarajulu VS. AI-driven mission-critical software optimization for small satellites: Integrating an automated testing framework. 2025.
58. Suominen V, Rantanen T, Venermo M, Saarinen J, Salenius J. Prevalence and risk factors of PAD among patients with elevated ABI. Eur J Vasc Endovasc Surg. 2008;35(6):709–14. https://doi.org/10.1016/j.ejvs.2008.01.013 PMID: 18313338
59. Durand Y, Hallibert P, Wilson M, Lekouara M, Grabarnik S, Aminou D. The flexible combined imager onboard MTG: from design to calibration. In: Sensors, systems, and next-generation satellites XIX, 2015.