

A Physics-Guided Feature Selection Machine Learning Approach for Cloud Probability Retrieval and Detection from TEMPO Imagery

Xiaohang Shi^{1,2#}, Yulong Fan^{1,3#}, Kai Yang⁴, Zhanqing Li⁴, Xiong Liu⁵, Jhoon Kim⁶,
Lin Sun³, Shulin Pang⁷, Chao Liu², Jing Wei^{1*}

1. MEEKL-AERM, College of Environmental Sciences and Engineering, Institute of Tibetan Plateau, and Center for Environment and Health, Peking University, Beijing, China.
2. State Key Laboratory of Climate System Prediction and Risk Management, School of Atmospheric Physics, Nanjing University of Information Science and Technology, Nanjing, China.
3. College of Geomatics, Shandong University of Science and Technology, Qingdao, China.
4. Department of Atmospheric and Oceanic Science, Earth System Science Interdisciplinary Center, University of Maryland, College Park, MD, USA.
5. Atomic and Molecular Physics Division, Center for Astrophysics | Harvard & Smithsonian, Cambridge, MA, USA
6. Department of Atmospheric Sciences, Yonsei University, Seoul, South Korea
7. Innovation Research Center of Satellite Application, Faculty of Geographical Science, Beijing Normal University, Beijing, China

[#]Co-first authors who contributed equally to this study

Corresponding author: Jing Wei (jingwei@pku.edu.cn)

Contents of this file

Figures S1 to S12

Tables S1 to S5

Introduction

This supporting information includes ten figures and four tables that complement the main manuscript by providing additional analyses and visualizations of cloud retrieval products from the TEMPO and GEMS instruments. Figures S1–S2 and S4–S5 present the diurnal evolution and comparative spatial distributions of TEMPO-retrieved cloud fractions and masks over North America for selected seasonal cases, while complementary continuous and binarized evaluations against the official CLDO4 product are provided in Figures S3 and S7. The algorithm’s spatial and temporal robustness is further validated using independent AERONET observations, including individual spatial case comparisons (Figure S9) and long-term tracking of diurnal cloud occurrence frequency (Figure S8). In addition, to demonstrate the model’s architectural generalizability, Figures S10–S12 extend the hourly retrieval framework to Asia using GEMS hyperspectral observations, benchmarked against Himawari-8 data, AERONET, and CALIPSO observations. Tables S1 and S5 summarize the optimized spectral channels selected for TEMPO and GEMS cloud retrievals, respectively. A targeted performance assessment over highly challenging snow- and ice-covered surfaces is presented in Table S2. Tables S3 and S4 further demonstrate the statistical robustness of the PGML model through spatial and temporal 10-fold cross-validations.

Collectively, these materials support the discussion of the diurnal variability, spatial consistency, and validation of cloud retrievals from the TEMPO and GEMS geostationary instruments.

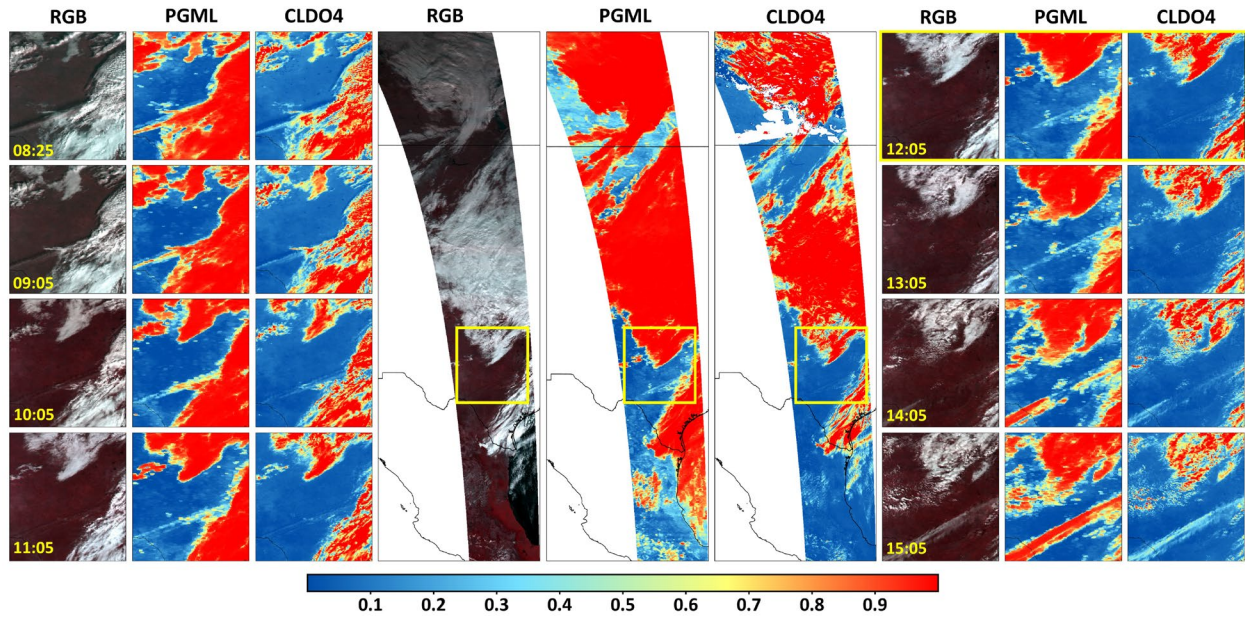


Figure S1. Diurnal evolution of TEMPO-retrieved results over North America on 20 November 2023 (Autumn case). Shown are the full-scene (central) and zoom-in (surrounding) color-composite images (Channels: 2041-1039-1143), PGML retrievals, and the official CLDO4 product.

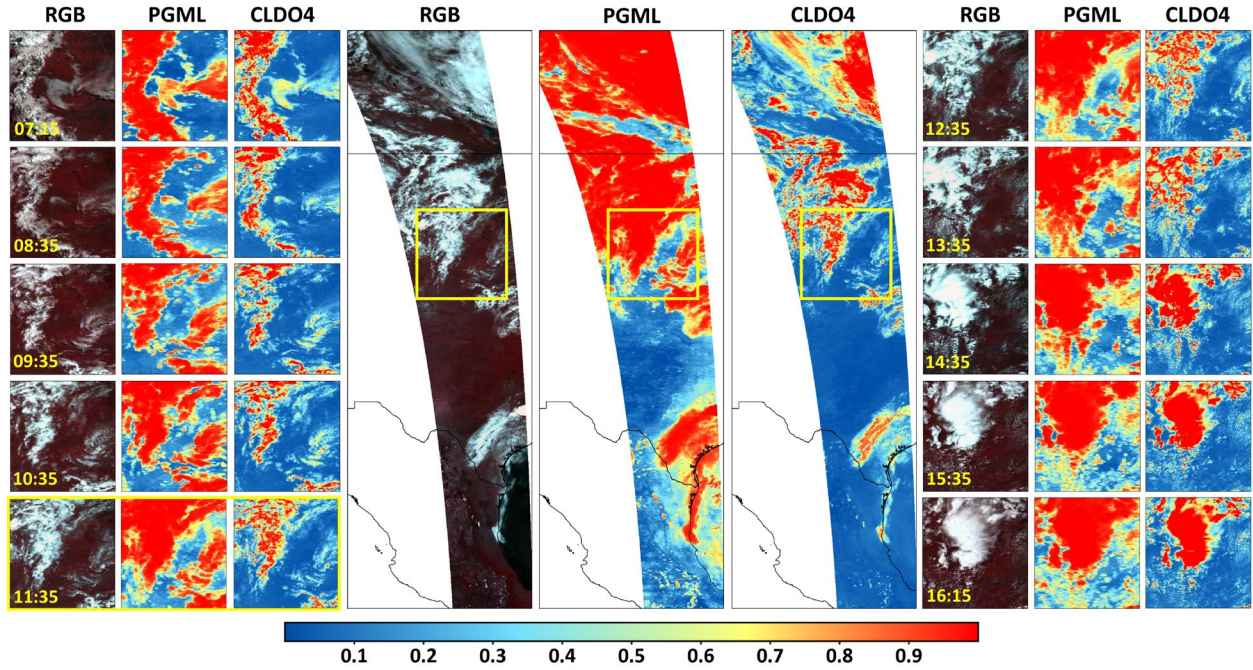


Figure S2. Diurnal evolution of TEMPO-retrieved results over North America on 26 May 2024 (Spring case). Shown are the full-scene (central) and zoom-in (surrounding) color-composite images (Channels: 2041-1039-1143), PGML retrievals, and the official CLDO4 product.

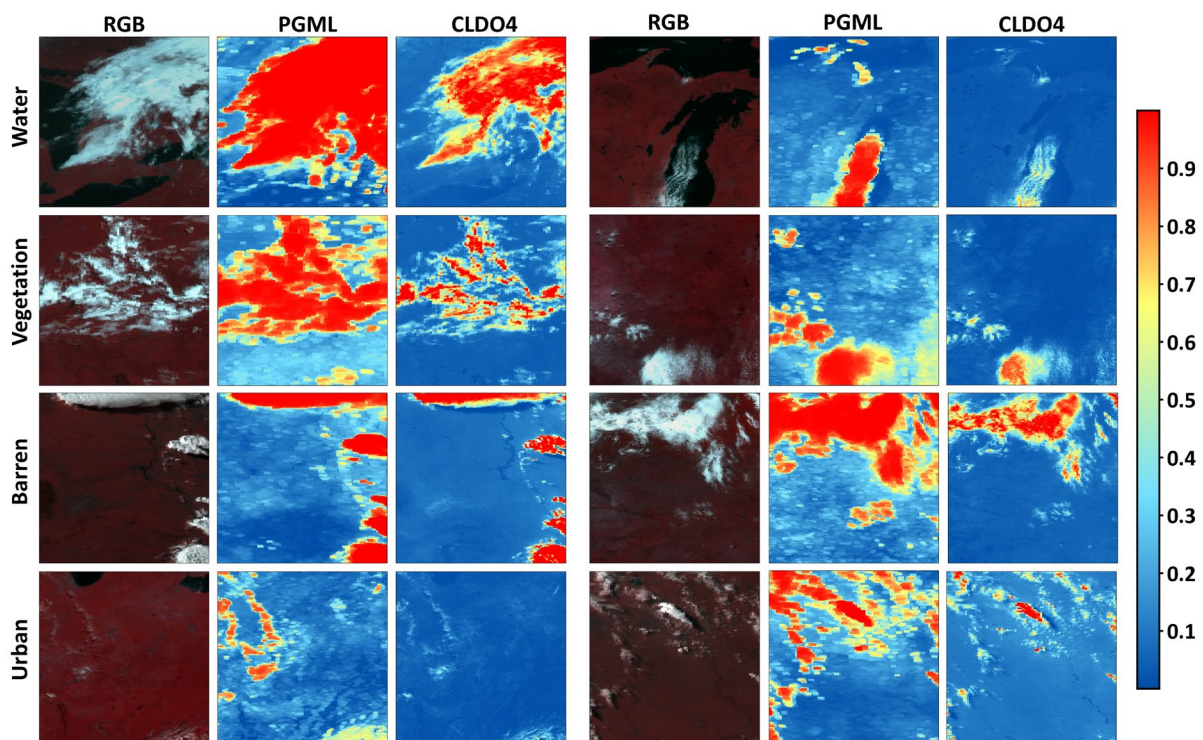


Figure S3. Comparison of PGML results with the TEMPO official algorithm (CLDO4) across different surface types.

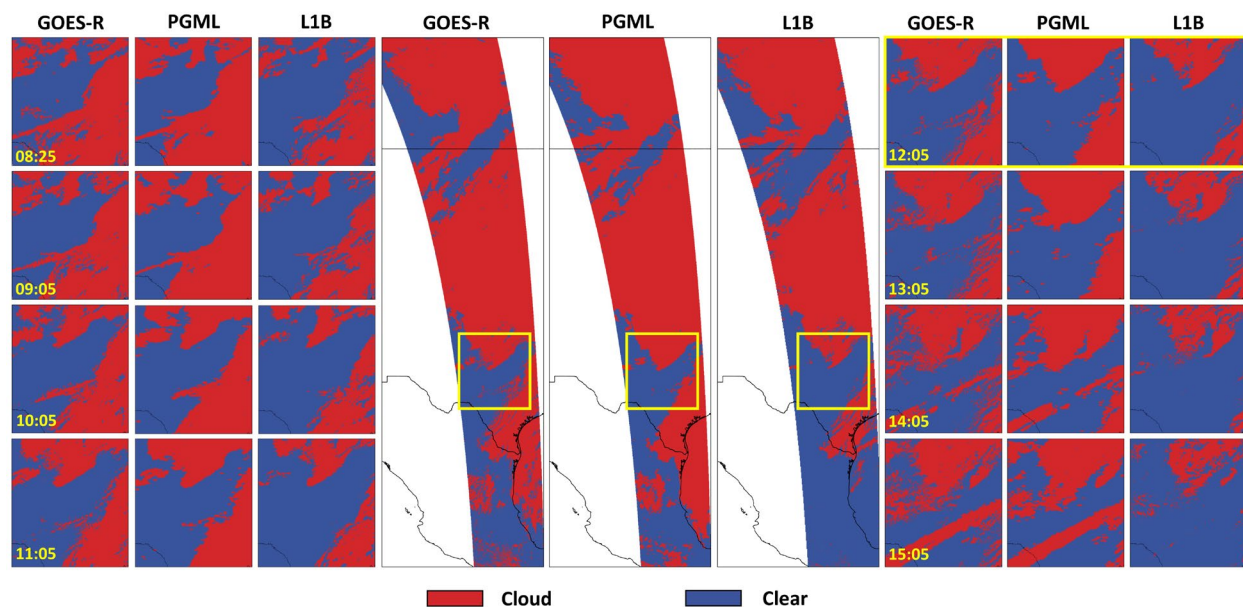


Figure S4. Comparison of the diurnal variations in cloud-mask results corresponding to Figure S1, derived from the GOES-R reference cloud mask, PGML retrievals, and the TEMPO L1B cloud mask over North America on 20 November 2023 (Autumn case).

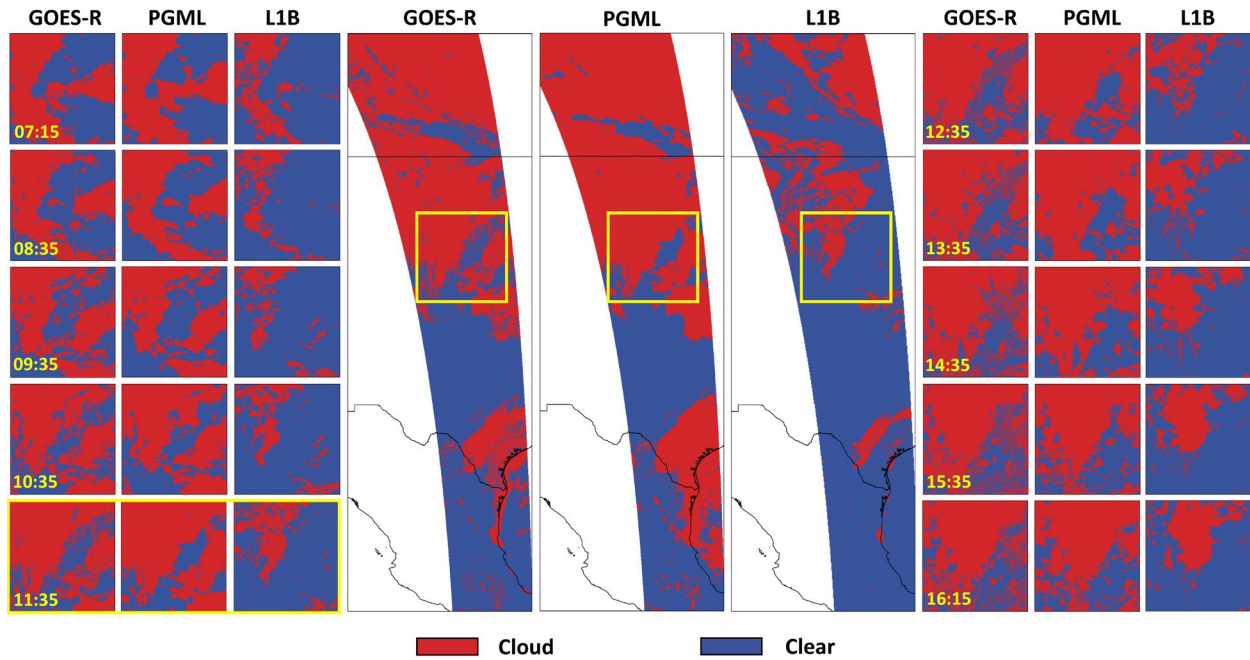


Figure S5. Comparison of the diurnal variations in cloud-mask results corresponding to Figure S2, derived from the GOES-R reference cloud mask, PGML retrievals, and the TEMPO L1B cloud mask over North America on 26 May 2024 (Spring case).

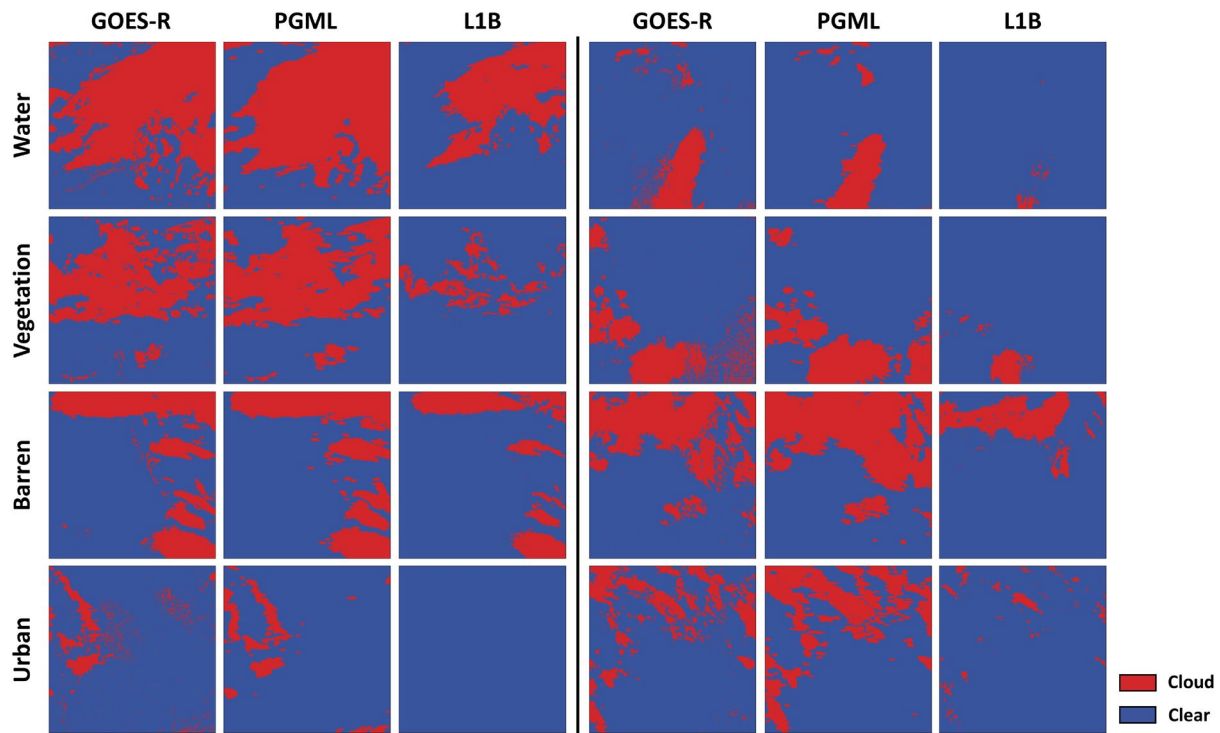


Figure S6. Comparison of cloud masks corresponding to Figure S3, derived from the GOES-R reference cloud mask, PGML retrievals, and the TEMPO L1B cloud mask across different surface types.

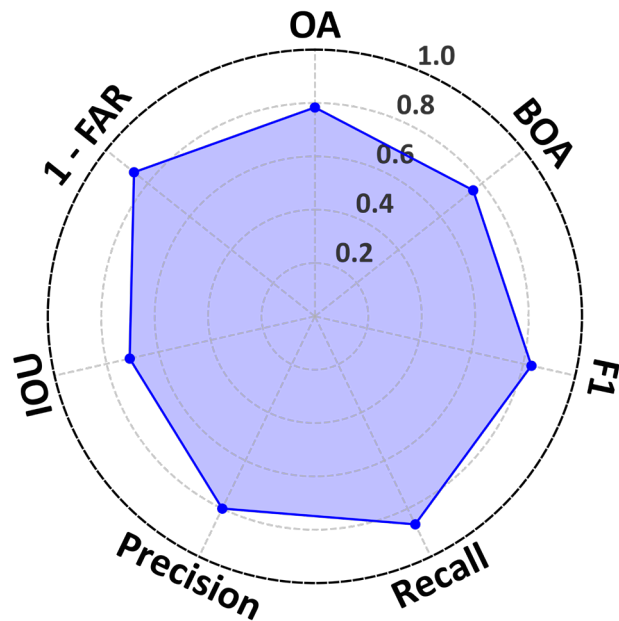


Figure S7. Evaluation of the binary cloud mask derived from the TEMPO CLDO4 product (using a cloud fraction threshold of 0.1) against the GOES-R reference cloud masks.

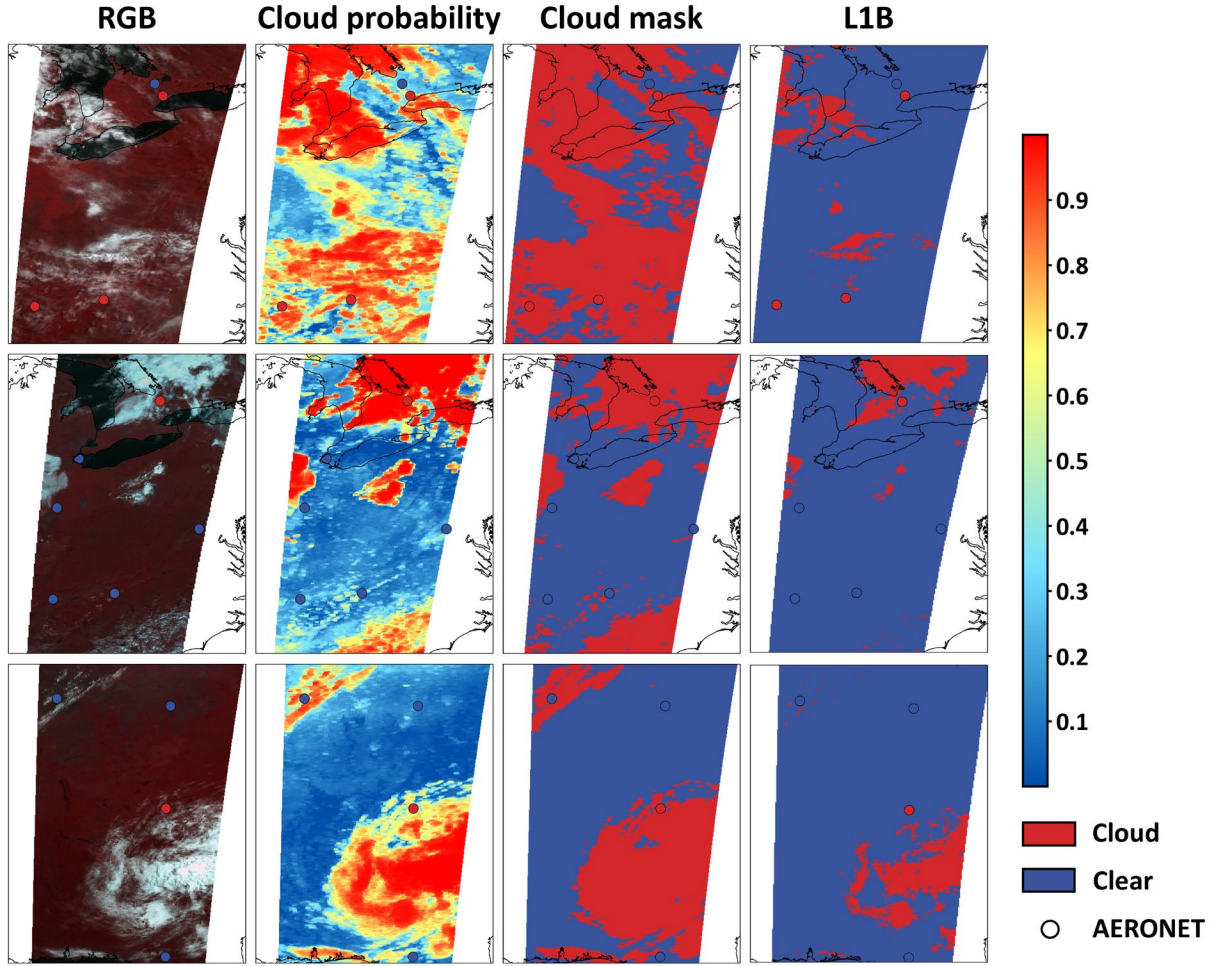


Figure S8. Spatial distribution of our TEMPO-retrieved cloud fraction and cloud mask, compared with the TEMPO L1B cloud mask, using AERONET observations as reference.

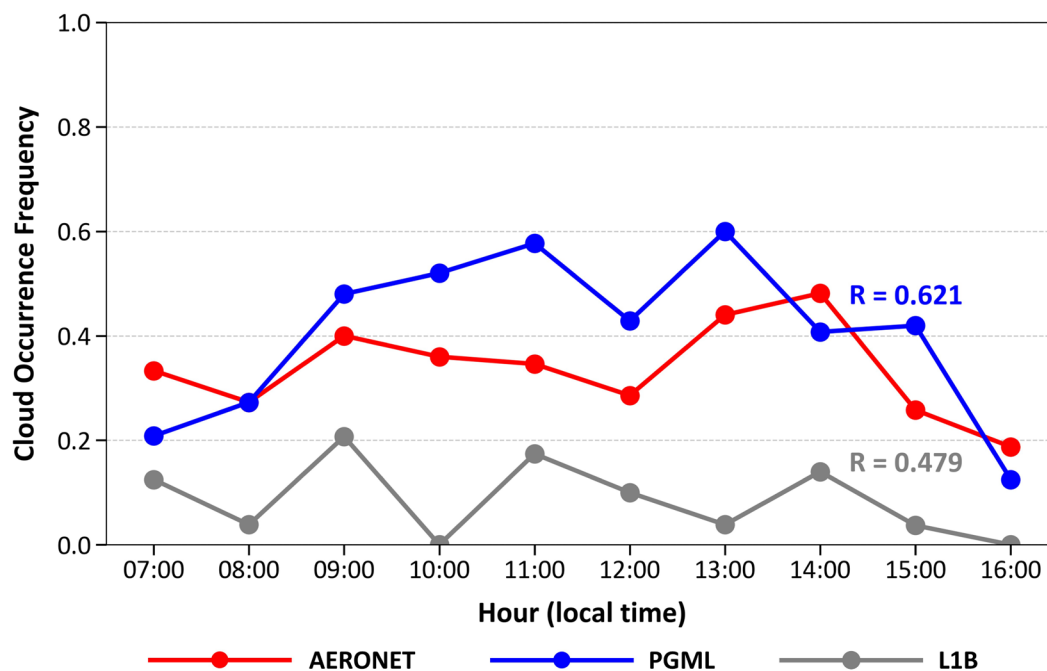


Figure S9. Comparison of diurnal cloud occurrence frequency among PGML, TEMPO L1B, and AERONET observations.

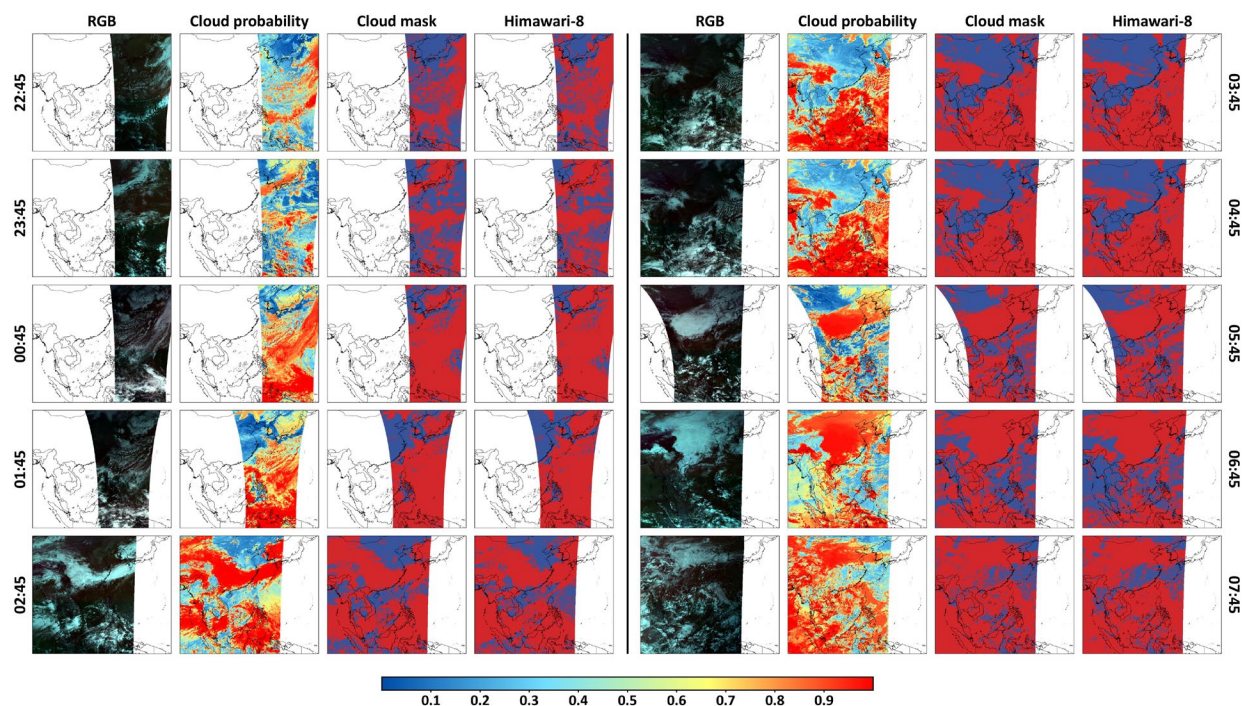


Figure S10. Diurnal evolution of GEMS-retrieved cloud probability and cloud mask over Asia at hourly intervals from 22:45 to 07:45 (UTC time). Shown are the full-scene color-composite images (Channels: 300-400-500), our cloud fraction retrievals, our cloud masks, and the Himawari-8 reference cloud masks.

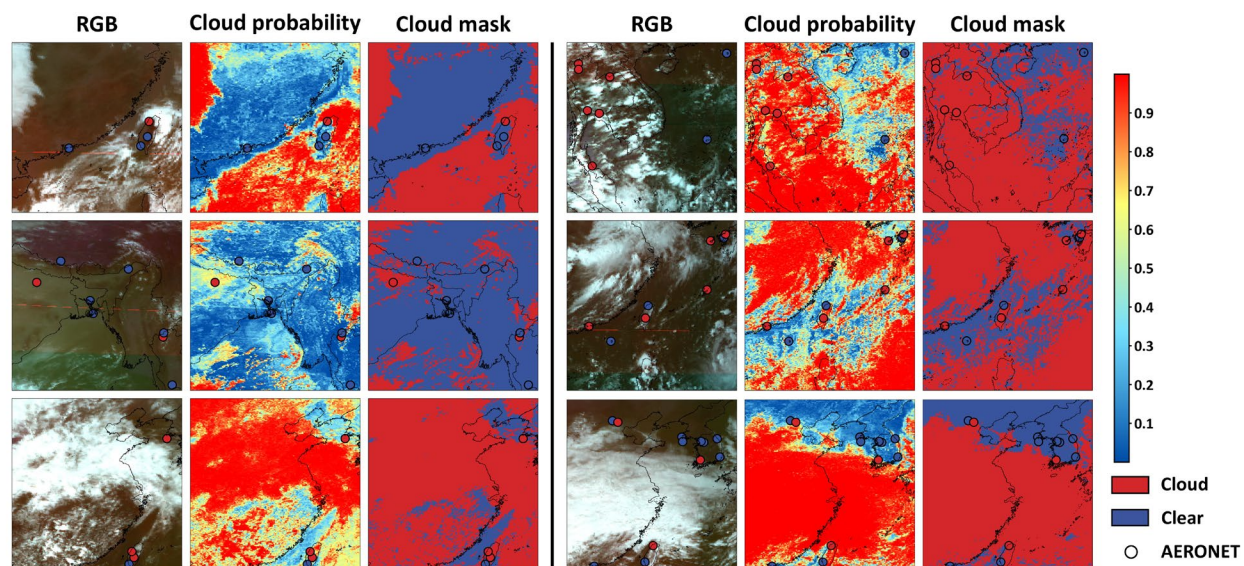


Figure S11. Spatial distribution of the accuracy assessment comparing GEMS-retrieved cloud probability and cloud mask with AERONET observations.

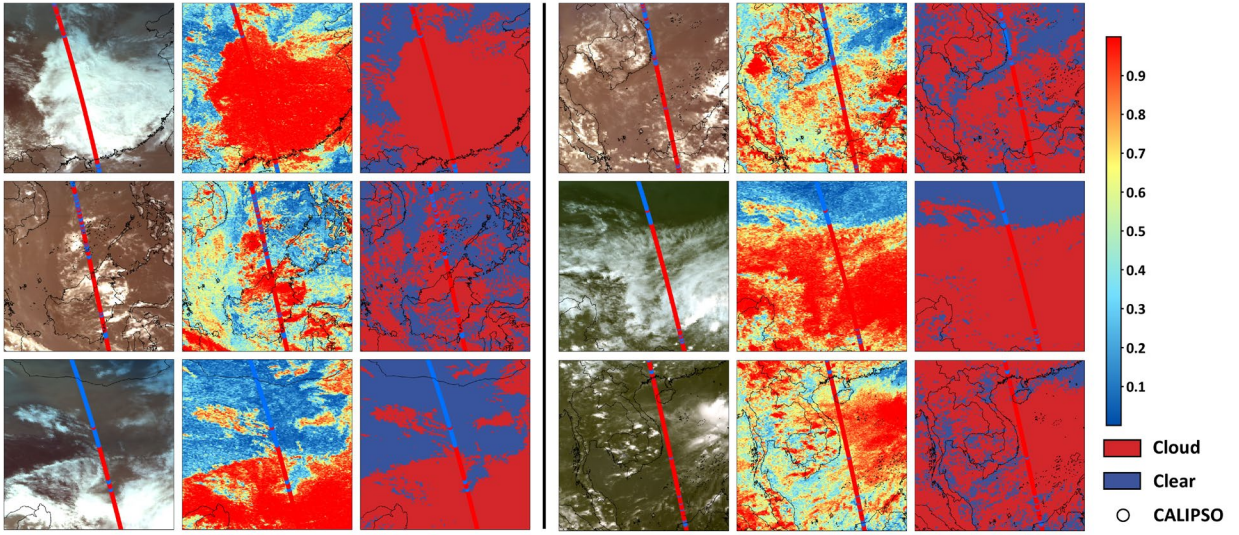


Figure S12. Spatial distribution of the accuracy assessment comparing GEMS-retrieved cloud probability and cloud mask with CALIPSO observations.

Table S1. Selected TEMPO spectral channels used for cloud fraction retrieval in this study.

Band	Wavelength (nm)	Band	Wavelength (nm)	Wavelength (nm)
UV	297.04 355.63 380.36	Visible	393.42	557.84
			396.79	561.21
			400.35	576.65
			432.61	584.16
			434.59	597.22
			452.40	617.01
			471.99	619.98
			486.24	651.82
			489.80	687.22
			540.62	708.37
			544.98	738.81

Table S2. Performance comparison of the PGML model for cloud detection over snow and ice.

Model	OA (%)	BOA (%)	F1 (%)	Recall (%)	Precision (%)	IoU (%)	FAR (%)
PGML	98.00	62.50	98.98	100.00	97.99	97.98	2.01
L1B	97.60	50.00	98.78	100.00	97.60	97.60	2.40

Table S3. Performance of the PGML model for cloud detection using the spatial-based 10-fold cross-validation approach.

Fold	OA (%)	BOA (%)	F1 (%)	Recall (%)	Precision (%)	IoU (%)	FAR (%)
1	84.12	82.72	87.62	87.67	87.58	77.97	12.41
2	84.52	83.38	87.63	88.03	87.23	77.98	12.76
3	84.55	83.57	87.54	87.70	87.38	77.83	12.62
4	84.50	83.44	87.58	87.75	87.41	77.90	12.59
5	84.42	82.94	87.86	88.26	87.47	78.35	12.53
6	84.05	83.03	87.16	87.28	87.04	77.24	12.96
7	84.47	83.69	87.22	87.35	87.09	77.34	12.90
8	84.30	83.14	87.59	87.51	87.67	77.92	12.32
9	84.79	84.07	87.54	87.28	87.79	77.84	12.20
10	84.38	83.09	87.70	87.90	87.51	78.10	12.49
All	84.41	83.32	87.55	87.68	87.42	77.85	12.58

Table S4. Performance of the PGML model for cloud detection using the temporal-based 10-fold cross-validation approach.

Hour	OA (%)	BOA (%)	F1 (%)	Recall (%)	Precision (%)	IoU (%)	FAR (%)
07:00	79.62	77.49	84.03	86.21	81.95	72.46	18.04
08:00	83.95	83.38	86.18	87.27	85.12	75.72	14.88
09:00	85.52	83.87	89.20	88.60	89.81	80.51	10.18
10:00	84.11	82.96	87.15	88.18	86.16	77.23	13.84
11:00	82.49	81.05	86.05	87.11	85.01	75.51	14.98
12:00	83.96	81.30	87.96	90.22	85.80	78.50	14.19
13:00	82.39	80.46	86.13	88.83	83.58	75.64	16.41
14:00	81.87	79.49	86.22	87.52	84.96	75.78	15.04
15:00	82.21	81.13	85.73	85.45	86.01	75.02	13.99
16:00	81.88	80.81	85.35	85.36	85.33	74.43	14.67
All	82.96	81.36	86.57	87.64	85.53	76.33	14.47

Table S5. Selected GEMS spectral channels used for cloud fraction retrieval in this study.

Band	Wavelength (nm)	Band	Wavelength (nm)	Wavelength (nm)
UV	316.8	Visible	388.8	432.0
	324.0		393.6	439.8
	349.8		397.2	446.4
	361.8		400.8	448.8
	367.8		405.0	477.0
	373.8		411.6	480.6
	378.6		423.6	497.4
			427.2	