Merging Multiple Satellite Precipitation Products over South Korea using Random Forest Model

Giang V. Nguyen (Presenter), Sungho Jung, Prof. Giha Lee

Water Disaster Research Lab

Introduction

Ensemble method

Precipitation data

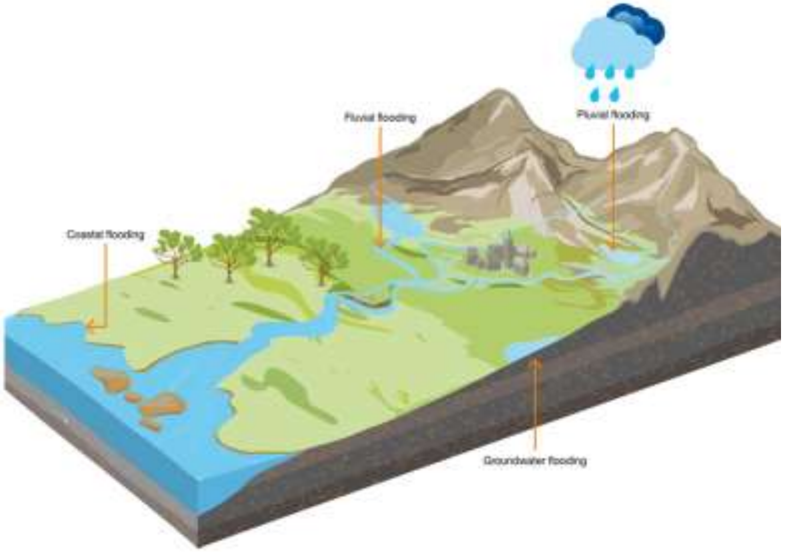
Results

Conclusion

- Precipitation (rain, snow, hail, …) is one of the key components as well as most challenging variables to estimate in the hydrology cycle.

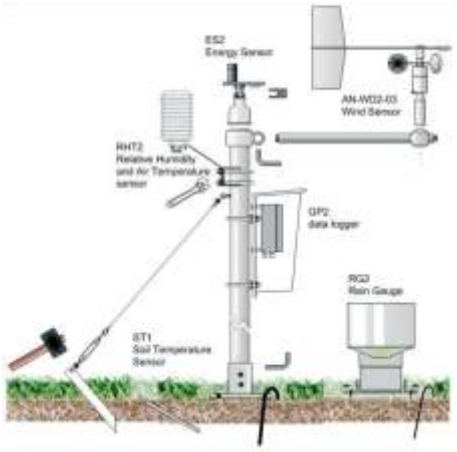
- It play a crucial role in studying of climate trend, water resources management and hydrological forecasting.

- Rainfall is one of the main causes of natural disasters, e.g. flooding, landside, drought, soil erosion, etc.



Accurate estimations of rainfall in different time scales (e.g. daily, monthly, season, annually) are extremely important quantities not only for researches but also for practical applications, water resources management as well as supporting decision makers.

Gauge-based



Advantages:

Rain gauge have a high accuracy at point-scale.

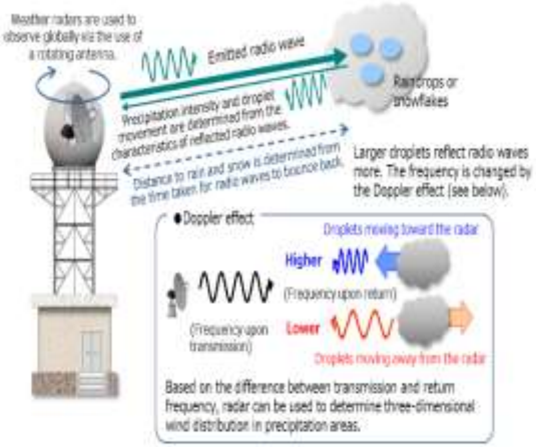
Disadvantages:

Scarcely and uneven distributed in most regions of the world.

Costly to construction and maintain.

→ The use of only ground-based measurement to estimate spatial distribution of precipitation is subject to large uncertainties.

Radar-based


Weather radars are used to observe globally via the use of a rotating antenna.
Emitted radio wave
Raindrops or snowflakes
Precipitation intensity and droplet movement are determined from the characteristics of reflected radio waves.
Distance to rain and snow is determined from the time taken for radio waves to bounce back.
Larger droplets reflect radio waves more. The frequency is changed by the Doppler effect (see below).
● Doppler effect
Droplets moving toward the radar
Higher
(Frequency upon return)
(Frequency upon transmission)
Lower
Droplets moving away from the radar
Based on the difference between transmission and return frequency, radar can be used to determine three-dimensional wind distribution in precipitation areas.

Advantages:

Radar precipitation provides highly resolution in both spatiotemporal scale.

Disadvantages:

Radar-derived data still has several drawbacks:
+ Coverage in limited areas.
+ Costly infrastructure construction. Especially, in mountainous areas.
+ Inaccuracy under complex atmosphere conditions.

Satellite-based

Advantages:

Provide rainfall estimated at the global scale.

Available at no cost and in near-real time (NRT).

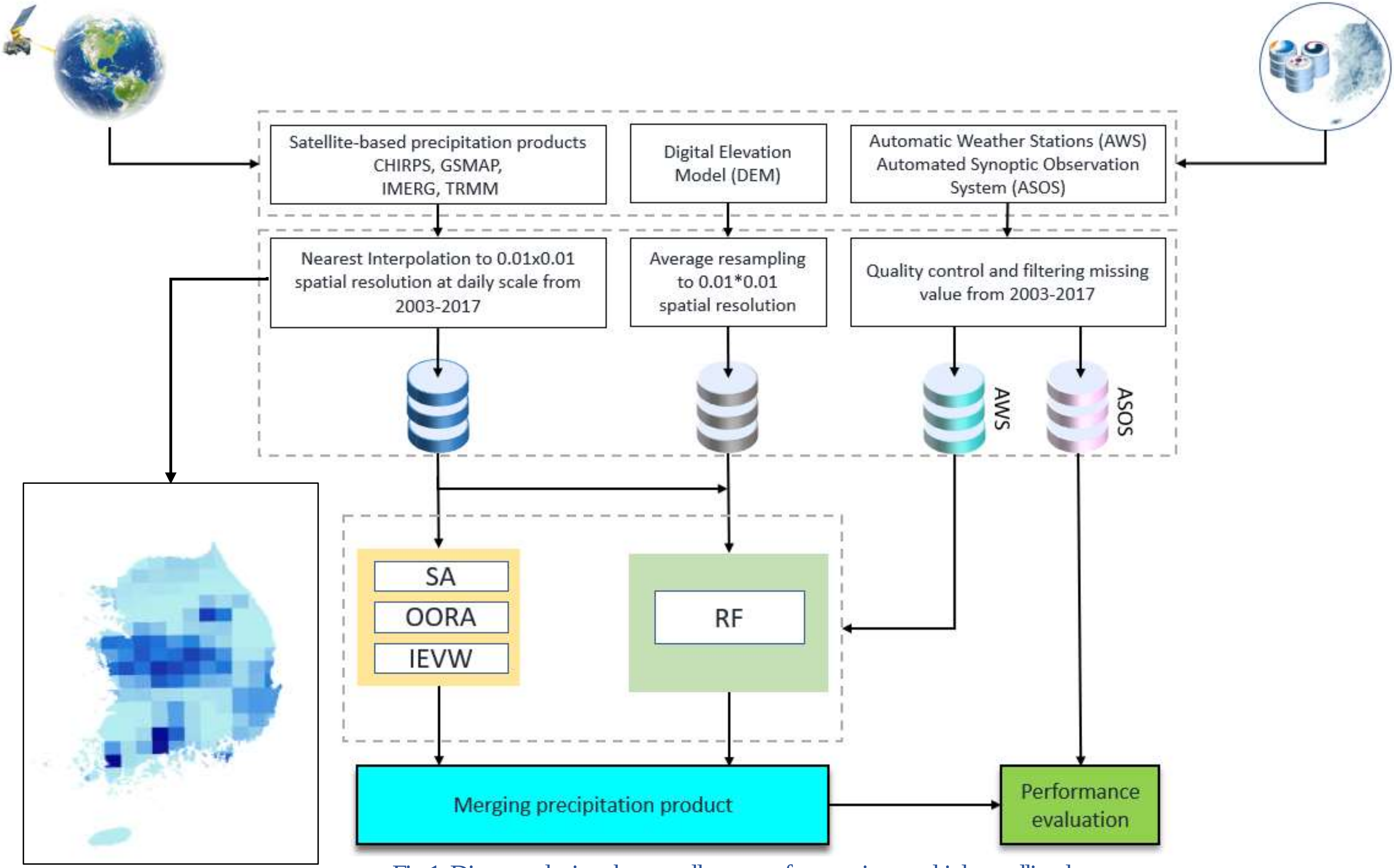
➡ Satellite-based information are highly valuable and have immense potential in water resource management, particularly for inaccessible transboundary and poorly-gauged river basins.

Disadvantages:

Multiple sources of errors are still present (e.g. false detection, systematic and random errors) and these products tend to perform worse at shorter time scales (e.g. daily and sub-daily).

The main objectives of the present study are to

1. Evaluate of the accuracy of precipitation from various satellite sources at daily scales over the whole South Korea.
2. Merge multiple sources of precipitation to obtain highly accurate rainfall data in the region of interest by using the Random Forest method.
3. Identify a suitable method that has high precision and easy to implement.



Satellite-based precipitation products CHIRPS, GSMAP, IMERG, TRMM

Digital Elevation Model (DEM)

Automatic Weather Stations (AWS) Automated Synoptic Observation System (ASOS)

Nearest Interpolation to 0.01x0.01 spatial resolution at daily scale from 2003-2017

Average resampling to 0.01*0.01 spatial resolution

Quality control and filtering missing value from 2003-2017

Merging precipitation product

Performance evaluation

Fig 1. Diagram depicts the overall process for merging multiple satellite data.

Statistical merging method:

Simple average:

$$P_{merge} = \frac{1}{N}\sum_{i=1}^{N} S_i$$

One-outlier-removed average :

$$P_{merge} = \frac{1}{N-1}\sum_{i=1}^{N-1} S_i$$

Inverse error variance weighting :

$$P_{merge} = \frac{1}{\sum_{i=1}^{N} 1/e_i^2}\sum_{i=1}^{N}\frac{1}{e_i^2} * S_i$$

Where : P is the merged precipitation

N is the number of SPPs

S_i is the i^{th} SPP

e is the error variance

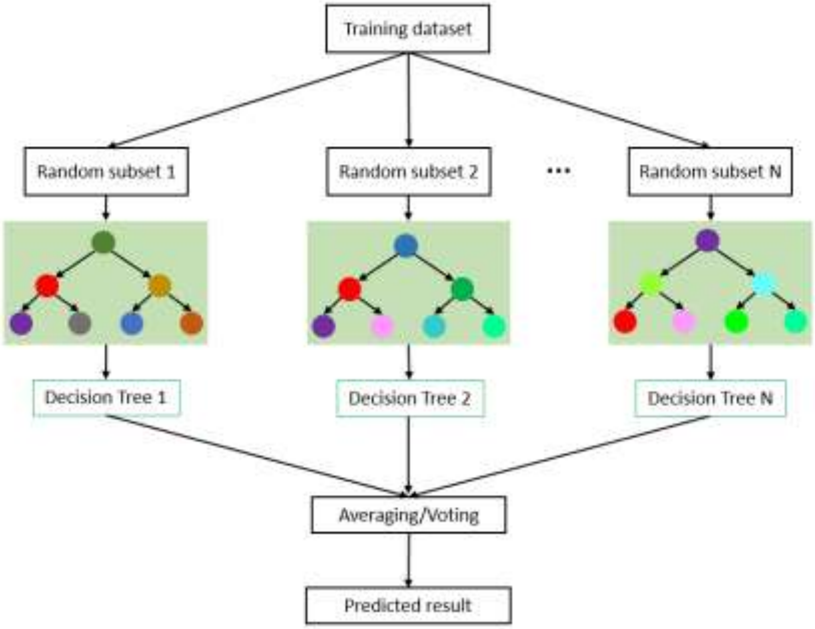
Random Forest (RF) is one of the most successful machine (statistical) learning algorithms for practical applications.

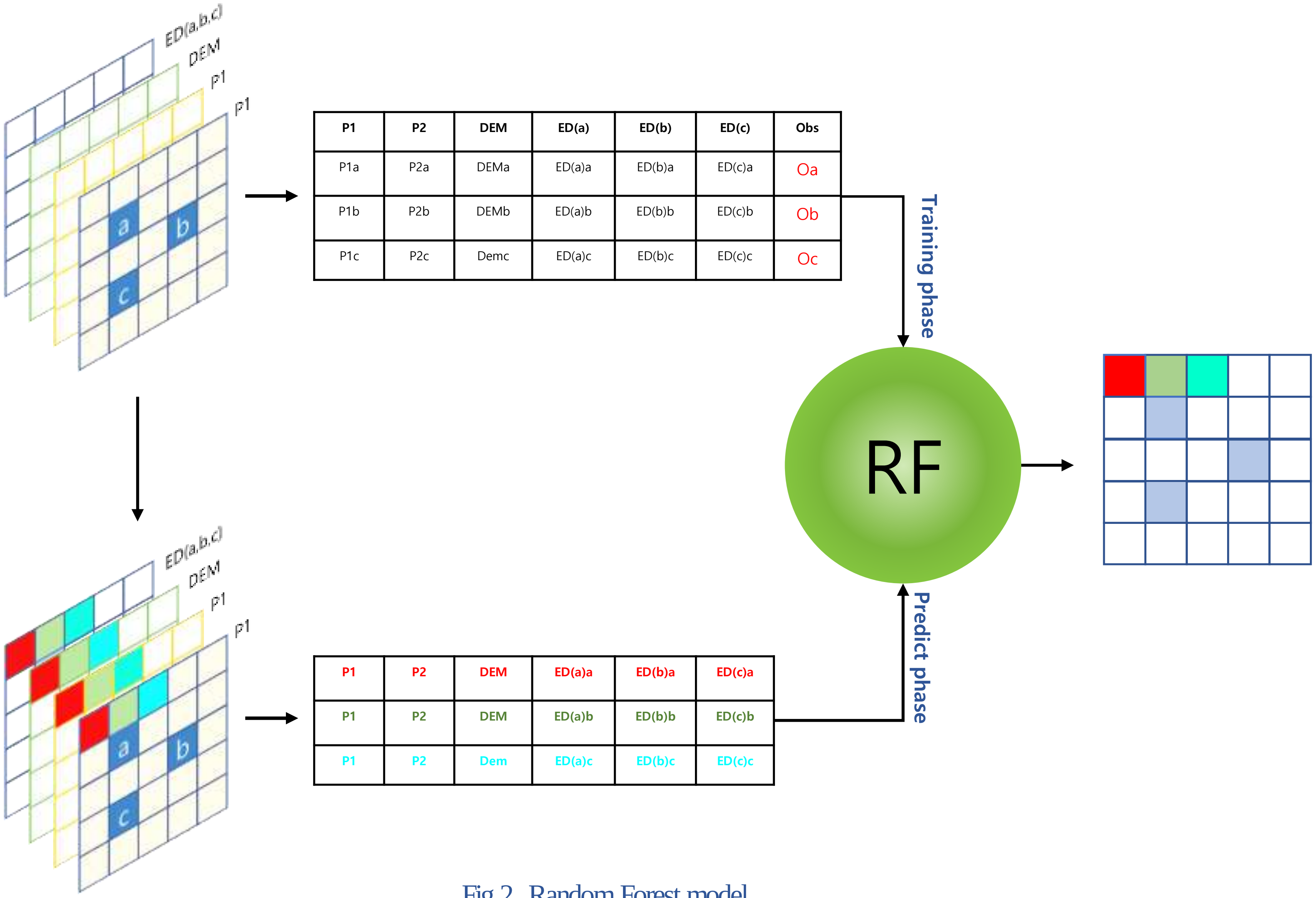
- ❖ It is less prone to overfitting than Decision Tree and other algorithms.
- ❖ RF can deal with complex or non-linear relationship between inputs and outputs and no need rigid assumption.
- ❖ Easy to implement.
- ❖ RF was successful applied to solve a some problems, eg: LULC classification satellite image or time series prediction.

➡ Random Forest (RF) was chosen for merging multiple satellite precipitation in this study.

Random Forest model:

- ❖ It is built on a lot of Decision Tree
- ❖ The input for each Decision Tree in the RF model were chosen randomness by bagging algorithm
- ❖ Final results, with regression problems, will be obtained by taking average results from all of the Decision Trees.





P1	P2	DEM	ED(a)	ED(b)	ED(c)	Obs
P1a	P2a	DEMa	ED(a)a	ED(b)a	ED(c)a	Oa
P1b	P2b	DEMb	ED(a)b	ED(b)b	ED(c)b	Ob
P1c	P2c	Demc	ED(a)c	ED(b)c	ED(c)c	Oc

P1	P2	DEM	ED(a)a	ED(b)a	ED(c)a
P1	P2	DEM	ED(a)b	ED(b)b	ED(c)b
P1	P2	Dem	ED(a)c	ED(b)c	ED(c)c

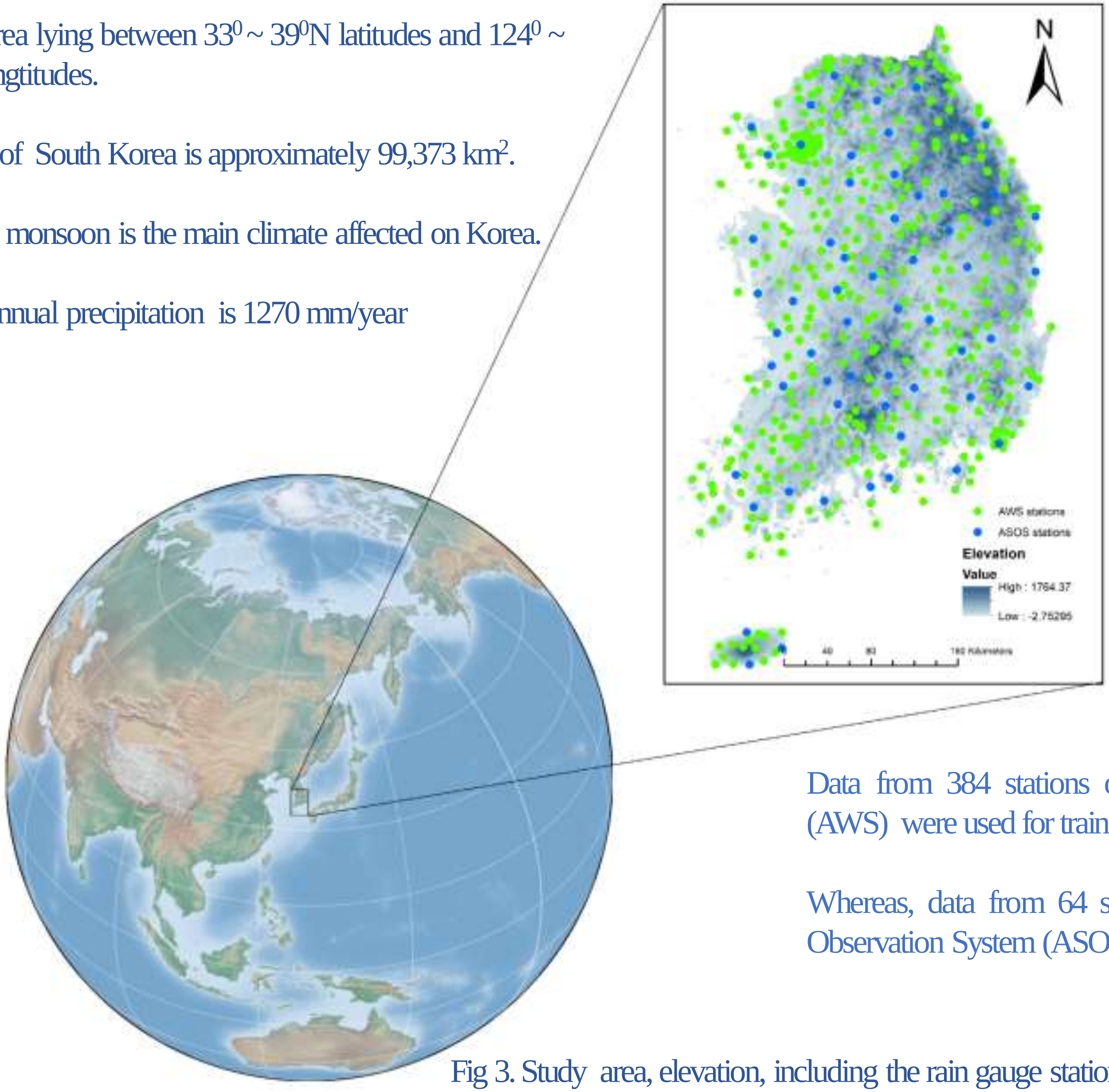
Fig 2. Random Forest model.

- South Korea lying between 33^0 ~ 39^0N latitudes and 124^0 ~ 130^0 E longtitudes.

- Total area of South Korea is approximately 99,373 km^2.

- The Asian monsoon is the main climate affected on Korea.

- Average annual precipitation is 1270 mm/year



Data from 384 stations of Automatic Weather Stations (AWS) were used for training.

Whereas, data from 64 stations of Automated Synoptic Observation System (ASOS)

Fig 3. Study area, elevation, including the rain gauge stations used in this research.

Table 1. Overview of 4 satellite data using for merge processing

Dataset	Full name	Latitudinal coverage	Spatial resolution	Temporal resolutions	Remark
CHIRPSv2	Climate Hazards group Infrared Precipitation with Station Version 2.0	50^0N-50^0S	0.05^0	Daily	Training
GSMaP-G	Global Satellite Mapping of Precipitation	60^0N-60^0S	0.1^0	Daily	Training
IMERG-L	Integrated Multi-satellitE Retrievals for GPM	60^0N-60^0S	0.1^0	Daily	Training
TRMM 3B42v7	TRMM Multi-satellite Precipitation Analysis research product 3B42 Version 7	50^0N-50^0S	0.25^0	Daily	Training
MSWEPv2.8	Multi-Source Weighted –Ensemble Precipitation	Global	0.1^0	Daily	Evaluation

01

Continuous indices of model performance:

Kling-Gupta Efficiency:

$$\boxed{KGE = 1 - \sqrt{(r-1)^2 + (\beta-1)^2 + (\gamma-1)^2}}$$

Linear correlation: $r = \dfrac{\sum_{i=1}^{N}(O_i-\overline{O})(S_i-\overline{S})}{\sqrt{\sum_{i=1}^{N}(O_i-\overline{O})^2}\sqrt{\sum_{i=1}^{N}(S_i-\overline{S})^2}}$

Bias ratio: $\beta = \dfrac{\mu_s}{\mu_o}$

Variability ratio: $\gamma = \dfrac{CV_s}{CV_o} = \dfrac{\sigma_s/\mu_s}{\sigma_o/\mu_o}$

Root mean square error:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(O_i - S_i)^2}\,(mm/d)$$

Mean Absolute Error:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|O_i - S_i|(mm/d)$$

O_i is observed data, S_i is satellite product, μ_s is the mean of satellite data, μ_o is the mean of observation data. σ_s is standard deviation of satellite data and σ_o is standard deviation of observation data.

Contingency Table Scores

Satellite product	Observed rainfall		
	Yes	No	Total
Yes	Hit (H)	False alarm (F)	H+F
No	Miss (M)	Correct negative (C)	M+C
Total	H+M	F+C	Ne

Categorical metrics	Formula	Optimal Value
Probability of Detection (POD)	$\mathrm{POD} = \frac{H}{H+M}$	1
False Alarm Ratio (FAR)	$\mathrm{FAR} = \frac{F}{F+H}$	0
Critical Success Index (CSI)	$\mathrm{CSI} = \frac{H}{H+M+F}$	1

Rainfall event	Intensity (i) [mm.d^{-1}]
No rain	[0, 1)
Light rain	[1, 5)
Moderate rain	[5, 20)
Heavy rain	[20, 40)
Violent rain	>= 40

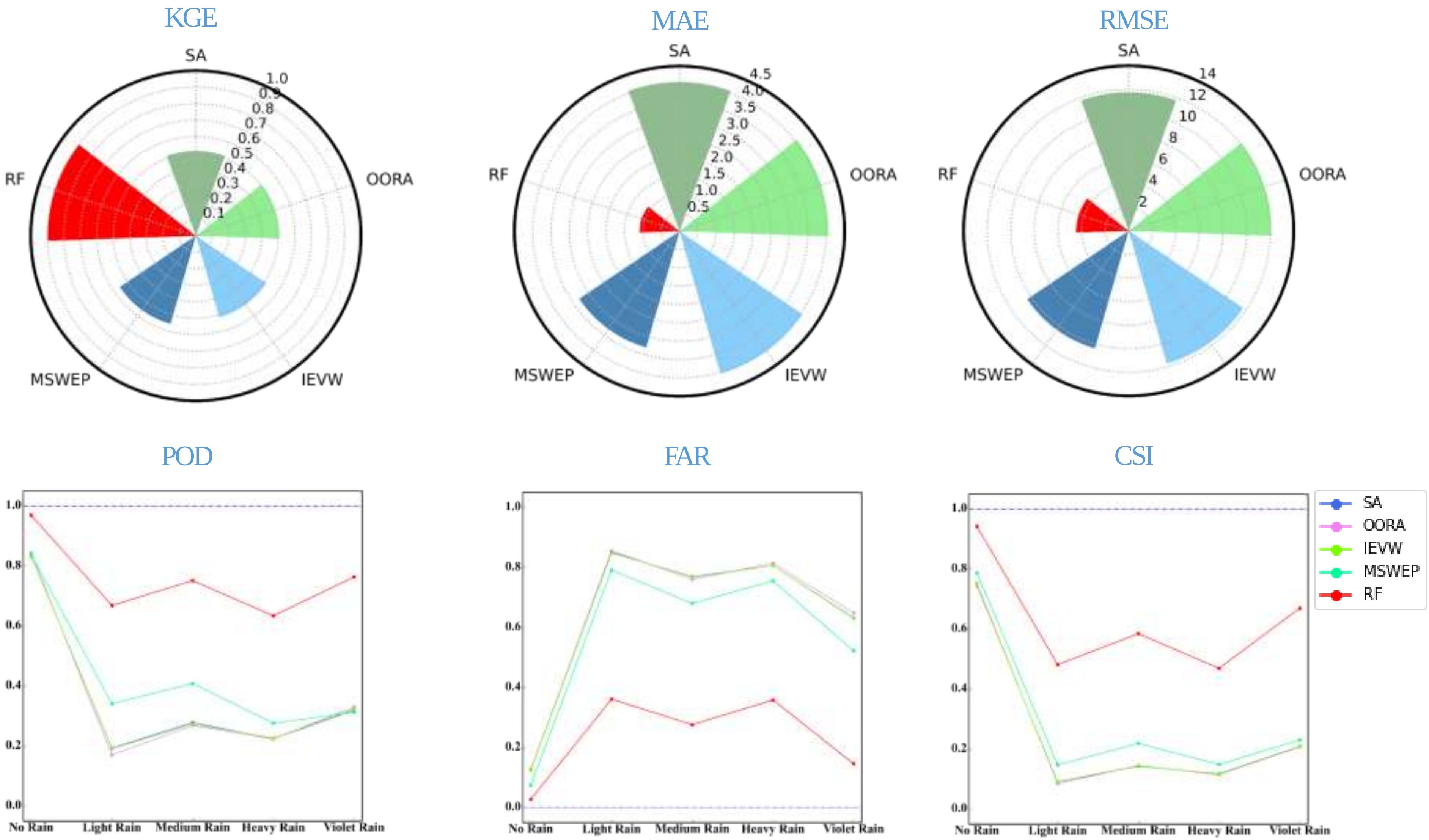


Fig 4. Performance of different merging procedures using the continuous and categorical indices.

Daily Analysis

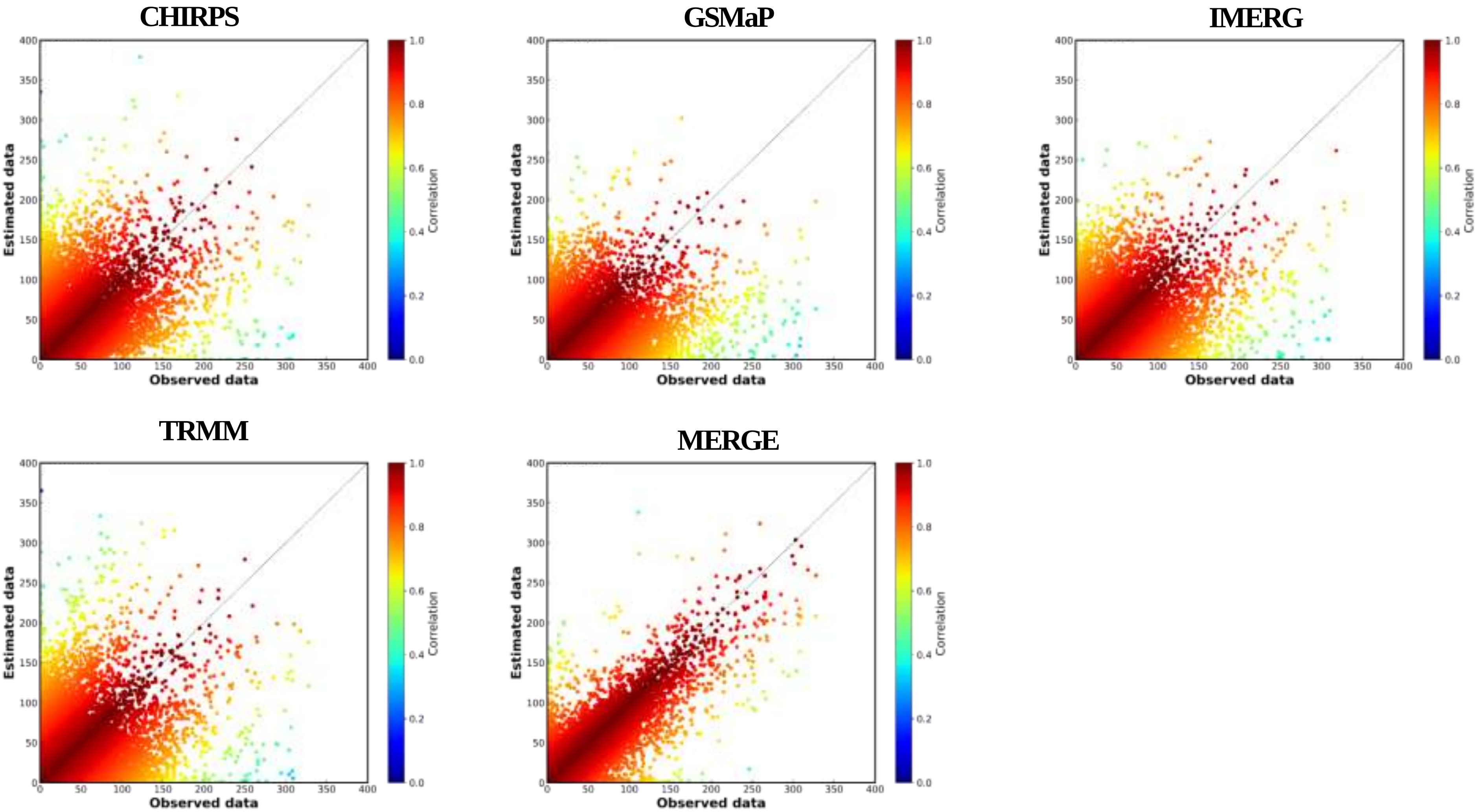
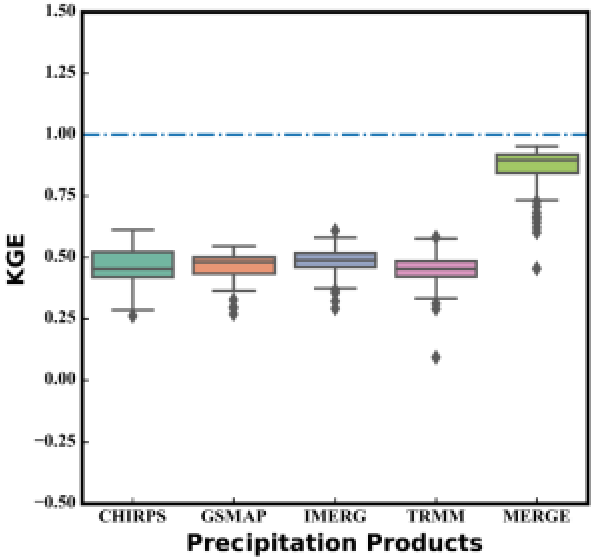


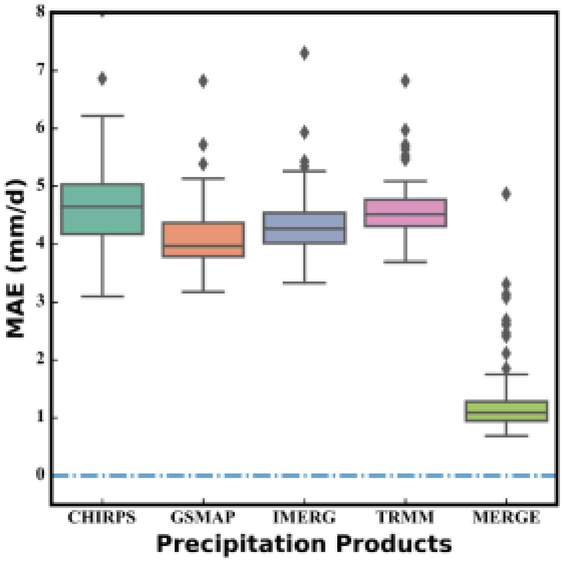
Fig 5. Correlation between P products and observed data for whole period of time from 2003-2017 at the daily scale.

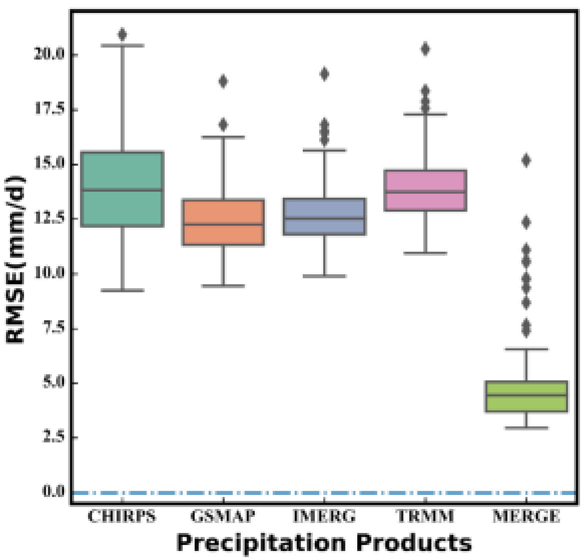
Daily Analysis

Table 2. Median errors of all P products with ground-based observations datasets from 2003-2017

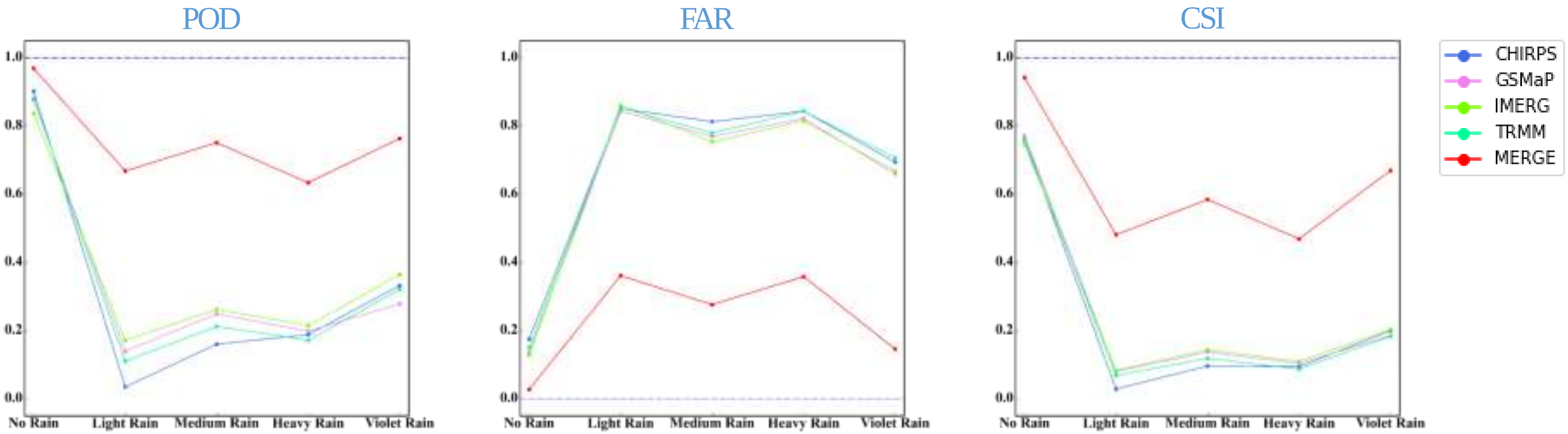
Precipitation products	CHIRPS	GSMaP	IMERG	TRMM	MERGE
r	0.464	0.504	0.531	0.470	0.984
β	1.031	0.920	1.134	1.051	0.961
γ	0.968	0.887	0.867	0.988	0.956
KGE	0.454	0.481	0.489	0.454	0.897
MAE(mm/d)	4.647	3.965	4.266	4.513	1.087
RMSE(mm/d)	13.828	12.248	12.523	13.730	4.445







Daily Analysis


POD

FAR

CSI

Fig 7. Median values of the categorical indices of performance at the five P intensity classes.

- ❖ The no –rain events were well captured by all products.
- ❖ The CSI presents the best performance for no-rain events followed by extreme events (>= 40 mm.d^{-1})
- ❖ FAR values were consistently the worst for the light rain intensities ([1, 5) mm.d^{-1})
- ❖ The POD and CSI of merging product has highest value while FAR is lowest when merge product was compared with other products.

Conclusions

❖ Random Forest (RF) was applied in order to obtain a suitable representation of P patterns in the whole region of interest.

❖ Among different testing methods, the merging method was carried out with multiple satellite products as a more suitable way than the single one.

❖ The performance of merged product **significantly increased** when more rain-gauge stations were used to train the model.

Further investigations

❖ Need more analysis about the satellite-based precipitation products at difference spatial resolution and consider the influence of topographic factor.

Thank for your attention !