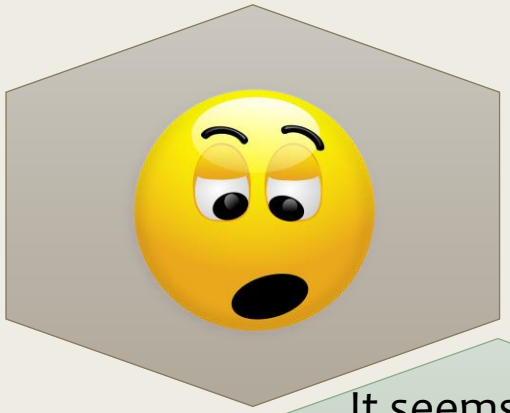
VERIFICATION PRACTICES IN ARPAE

Maria Stefania Tesini

WG5 – COSMO GENERAL MEETING 2020

OUTLINE

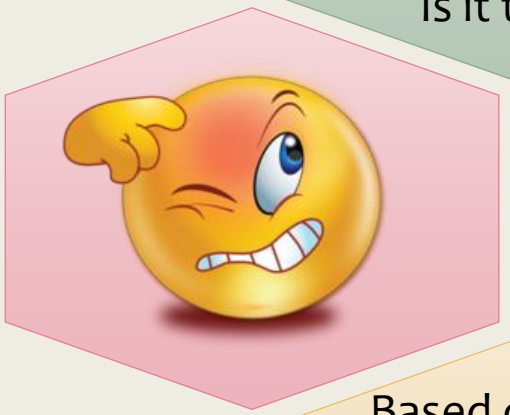


Our Verifications techniques are more or less the same since many years

- Versus for T2m,TD2m, WS, MSLP…
- DIST methodology applied to catchment areas for QPF

It seems to me that also the results are the same since many years too… Is it true?

- Let's look to the trend over time of some scores of T2m and Precipitation forecast for some models in use at Arpae



It can be difficult to highlight the quality of a model using a single score

- Especially for QPF, errors can be due to different sources (e.g. misses or false alarm) and scores are very dependent on the chosen thresholds

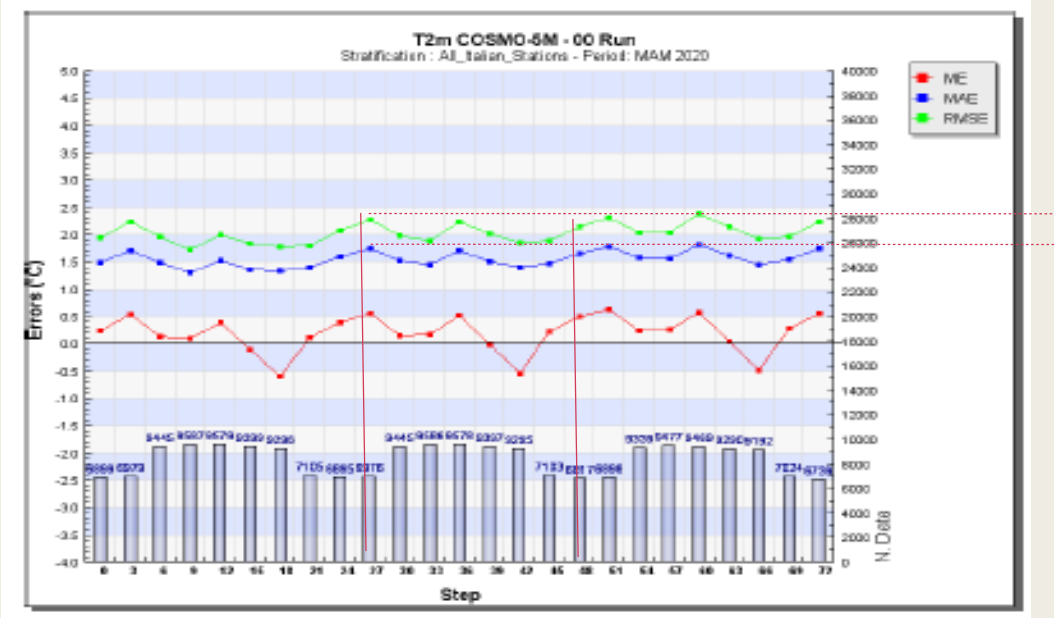
Based on user oriented verification of QPF, a different approach has been tested

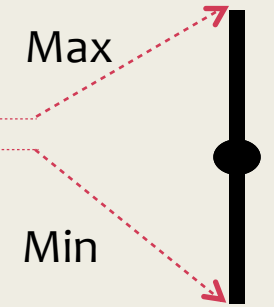
- Separate QPF in categories
- Visual representation using "bubbles plot"
- Build a multi-category contingency table
- Use of Gerrity score to quantify the results in a single score

T2m trend

- **Period:** DJF2013-14→MAM2020 (seasonal verification)
- **Verification system:** VERSUS (Nearest Point 3D optimized)
- **Observational dataset:** Italian Synop stations
- **Models:** COSMO-I7 → COSMO-5M, COSMO-I2 → COSMO-2I (with overlapping period during 2018-2019), IFS-ECMWF
- **Performance metrics:** mean of RMSE from +27h and +48h - step 3 hours with variability in that time range

Italian Synop stations

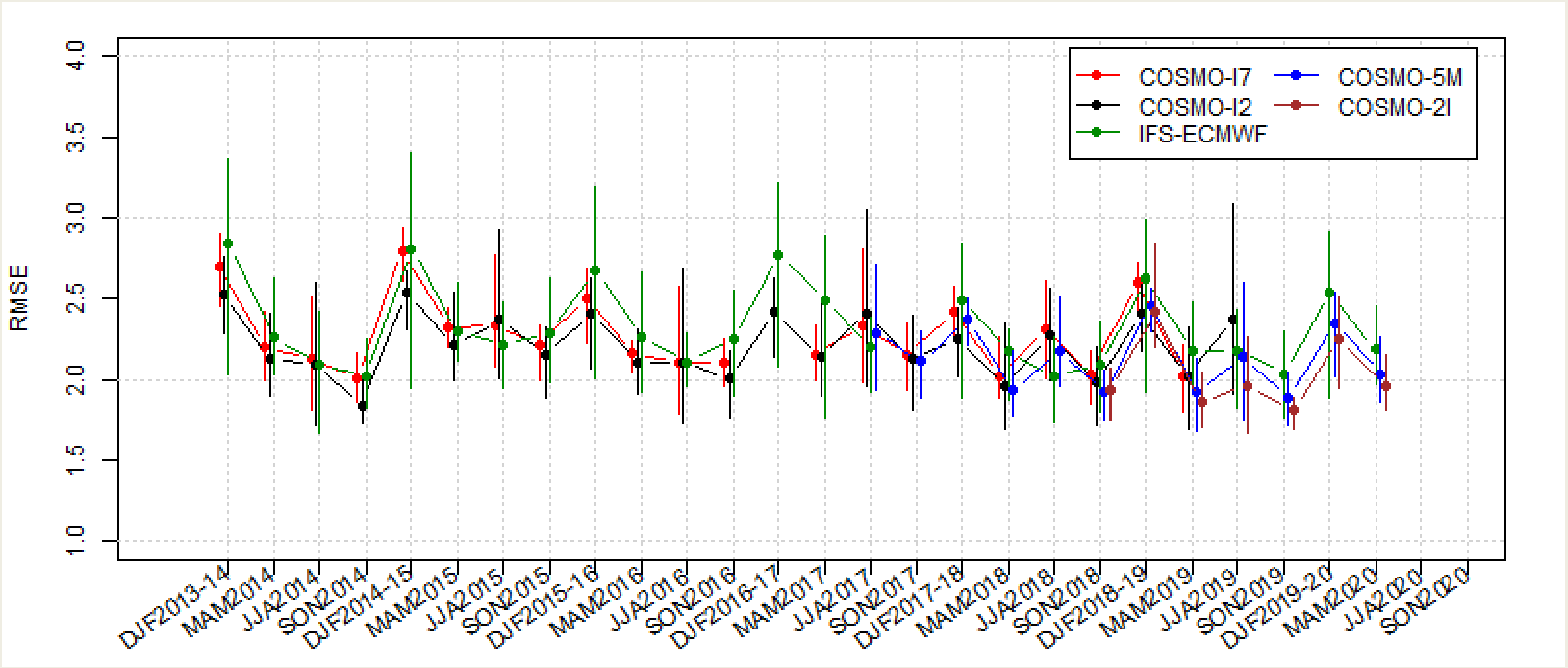




- The symbol represent the mean value of RMSE between +27h e +48h
- The bar represent the variability of RMSE in that interval

N.B. it is not an error bar!!

T2m: RMSE trend

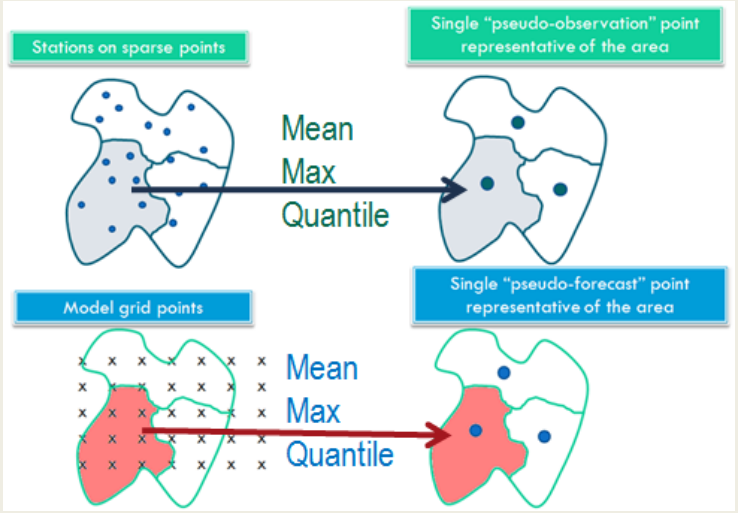


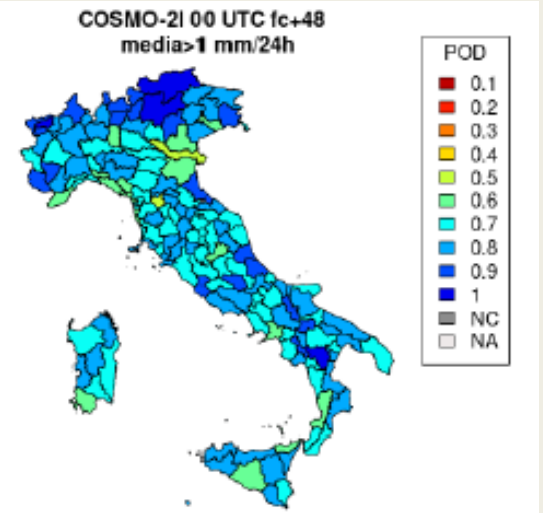
Symbols represent the mean value of RMSE between +27h e +48h
Bars represent the variability of RMSE in that interval
N.B. they are not error bars!!

QPF: TS trend

- **Period:** JJA2017→MAM2020 (seasonal verification)
  (before JJA2017 the verification domain was limited to Northern Italy)
- **Verification system :** DIST applied to catchment areas
- **Observational dataset:** National Civic Protection Department high-resolution rain-gauges network
- **Models:** COSMO-I7 → COSMO-5M, COSMO-I2 → COSMO-2I (with overlapping period during 2018-2019), IFS-ECMWF
- **Performance metrics:** TS evaluated for average and maximum of precipitation accumulated from +24h to +48h exceeding some thresholds:
  - *1 - 5 mm/24h for mean*
  - *10 – 20 mm/24h for max*

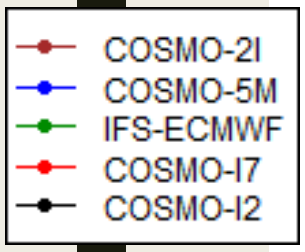
DPCN rain-gauges network

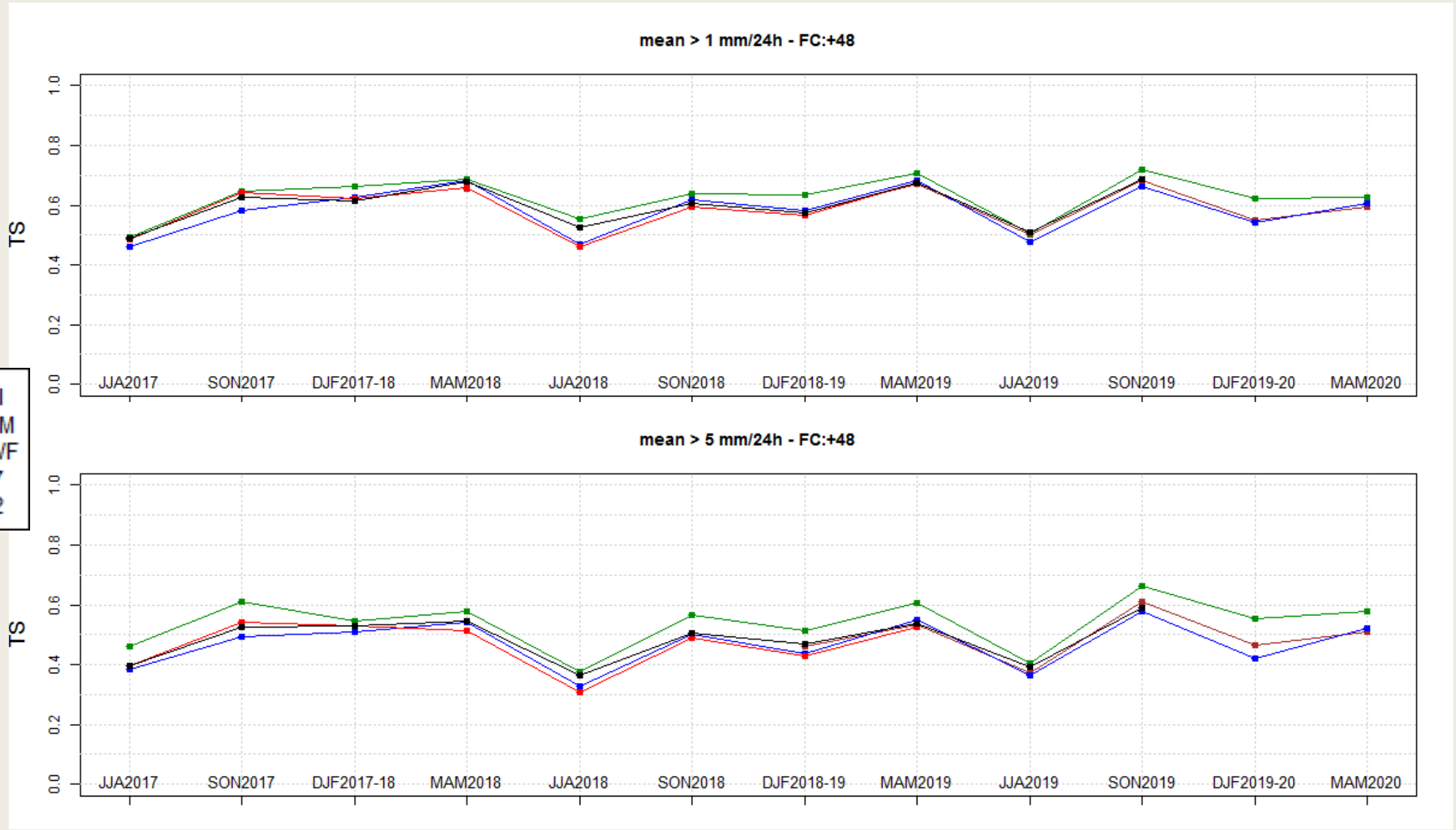




QPF: TS trend

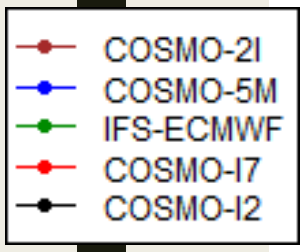
MEAN

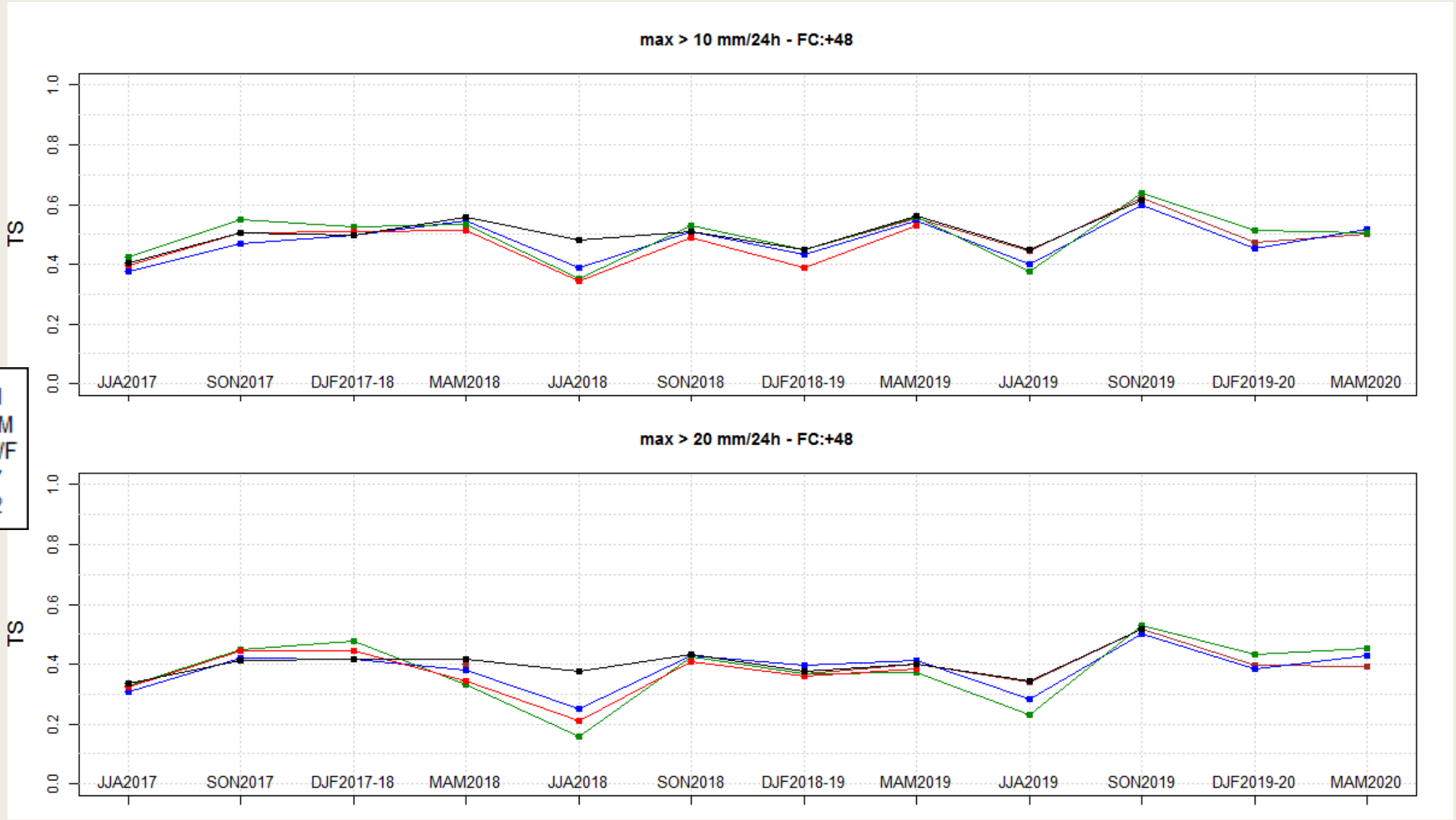




QPF: TS trend

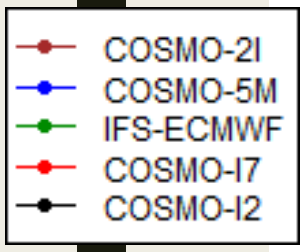
MAX

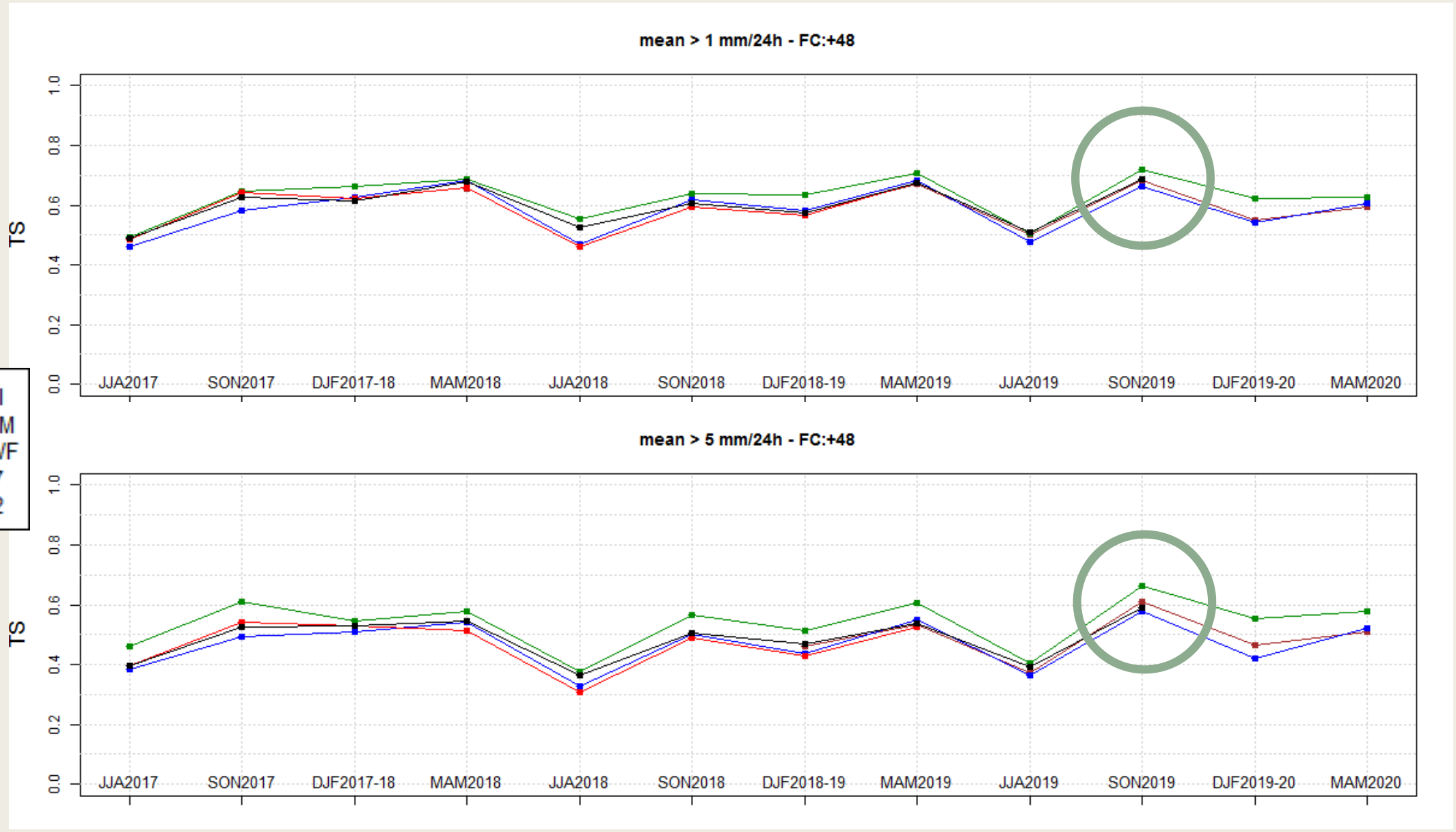




QPF: TS trend

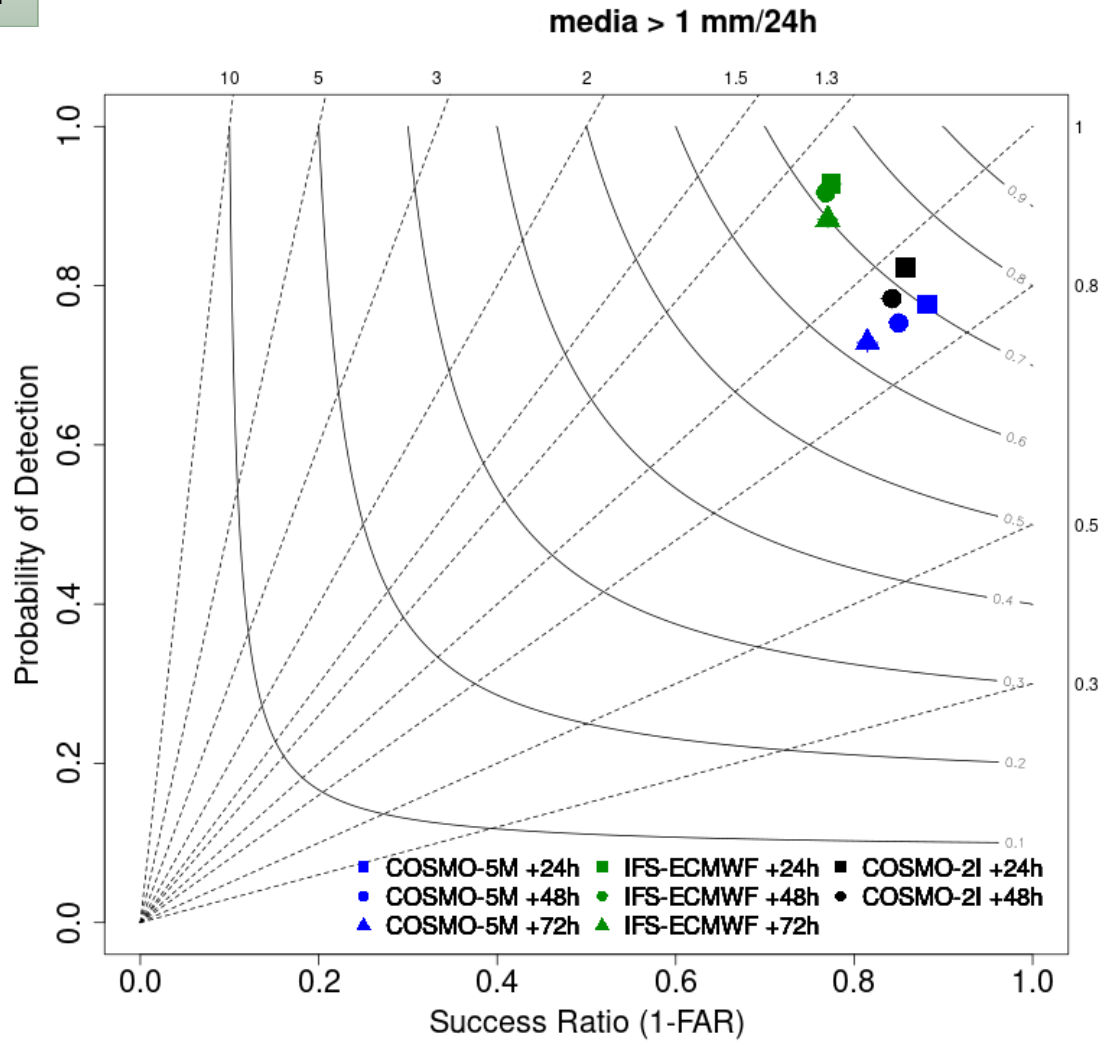
MEAN

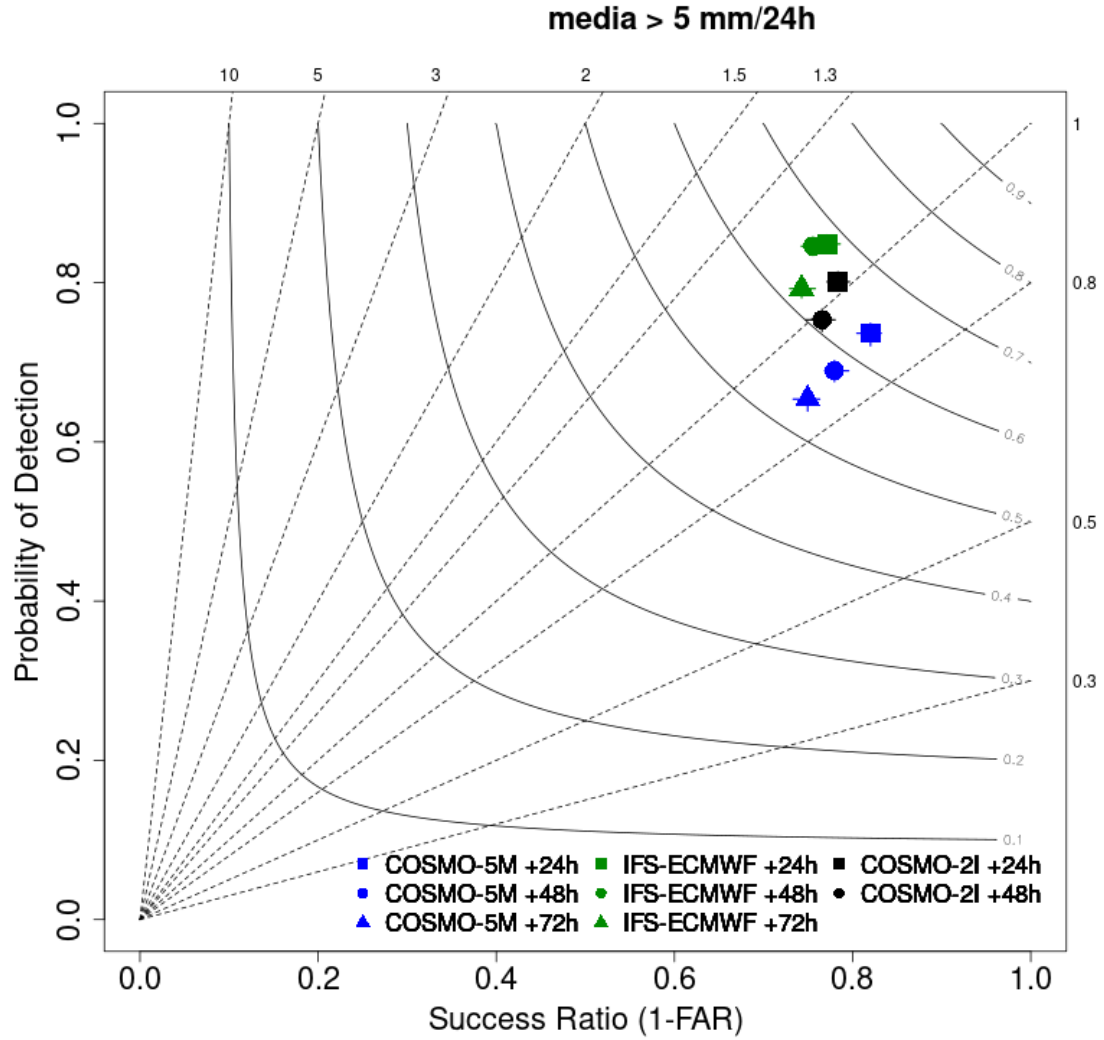




SON2019

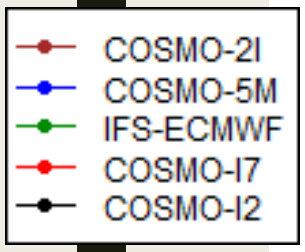
MEAN

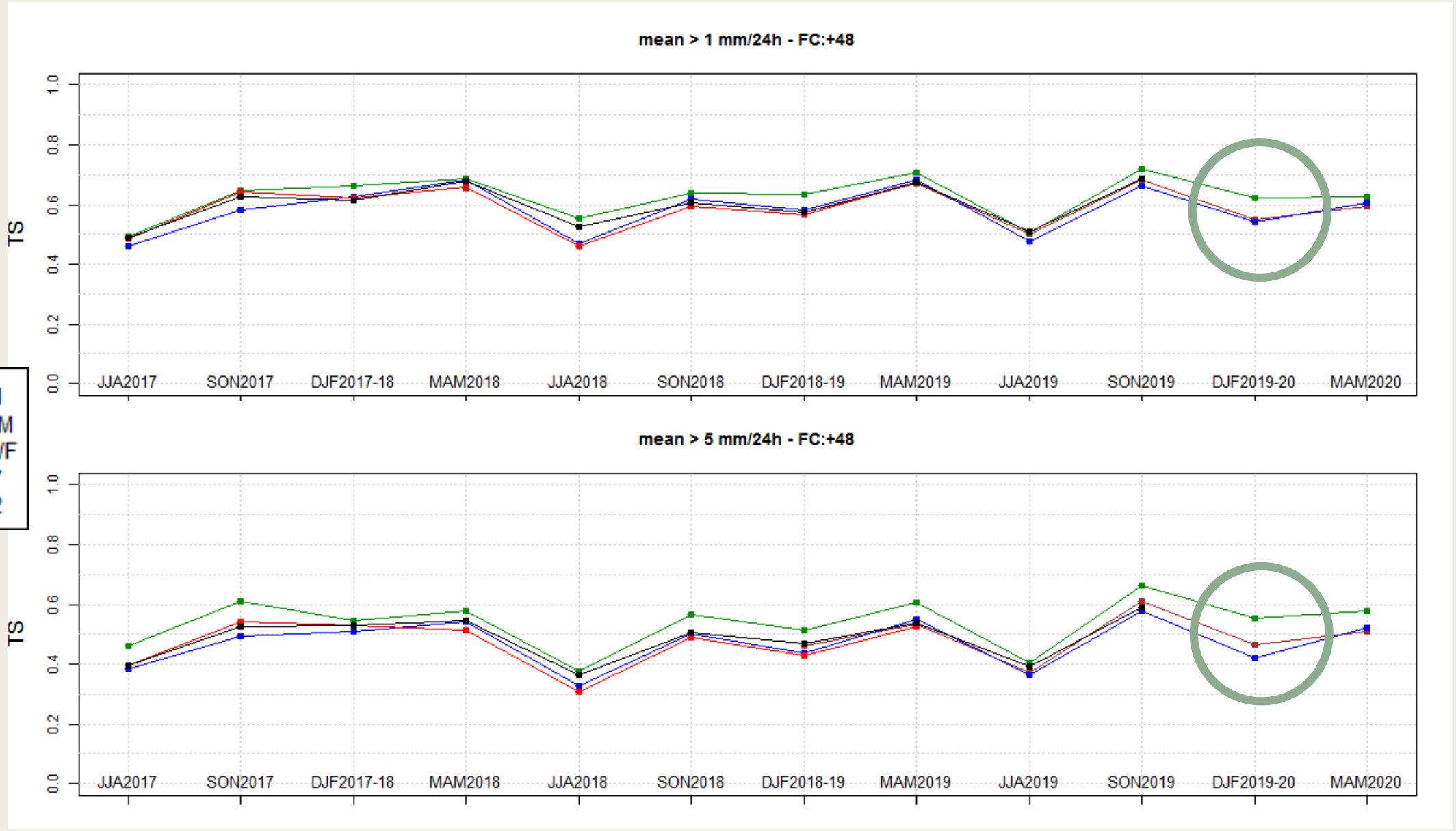




QPF: TS trend

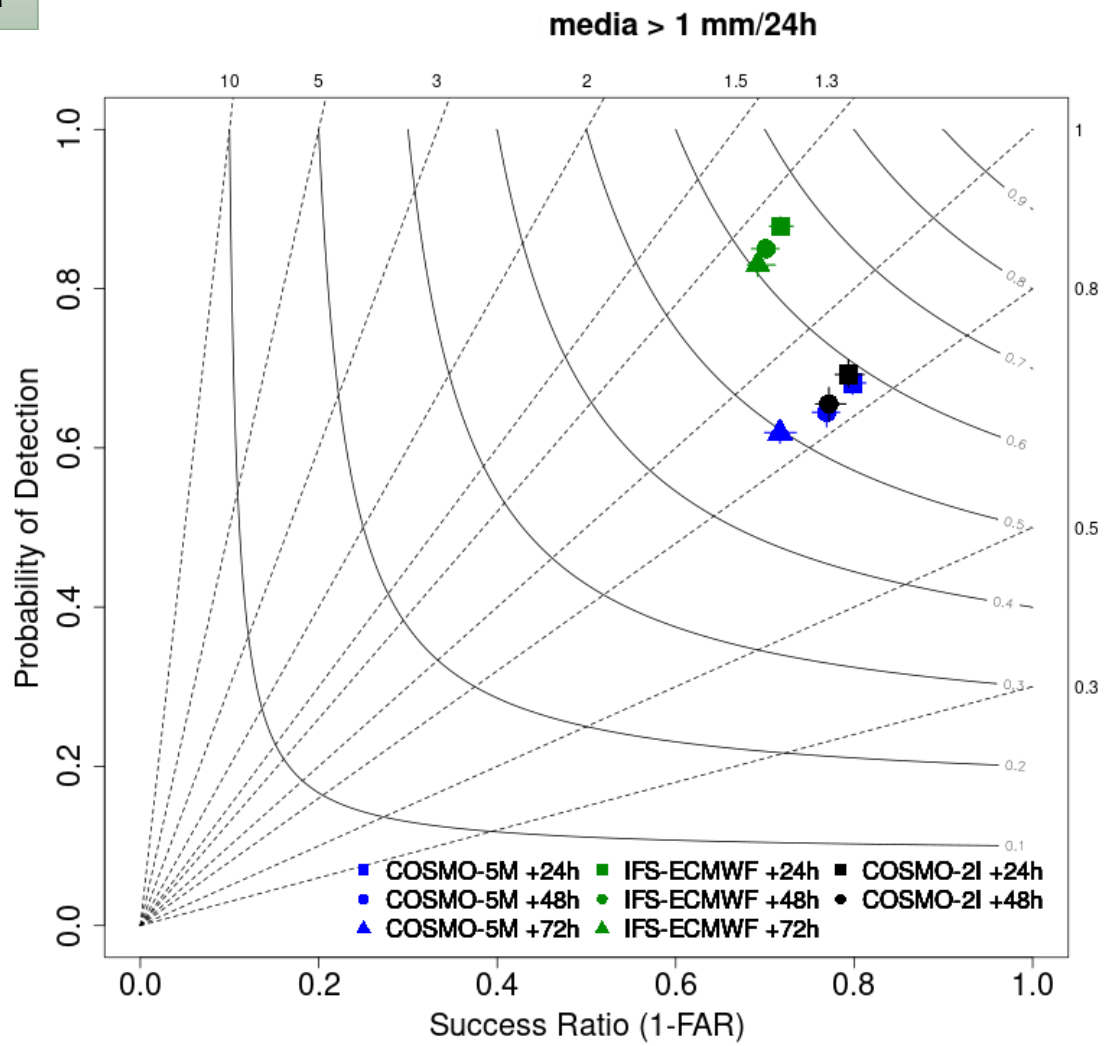
MEAN

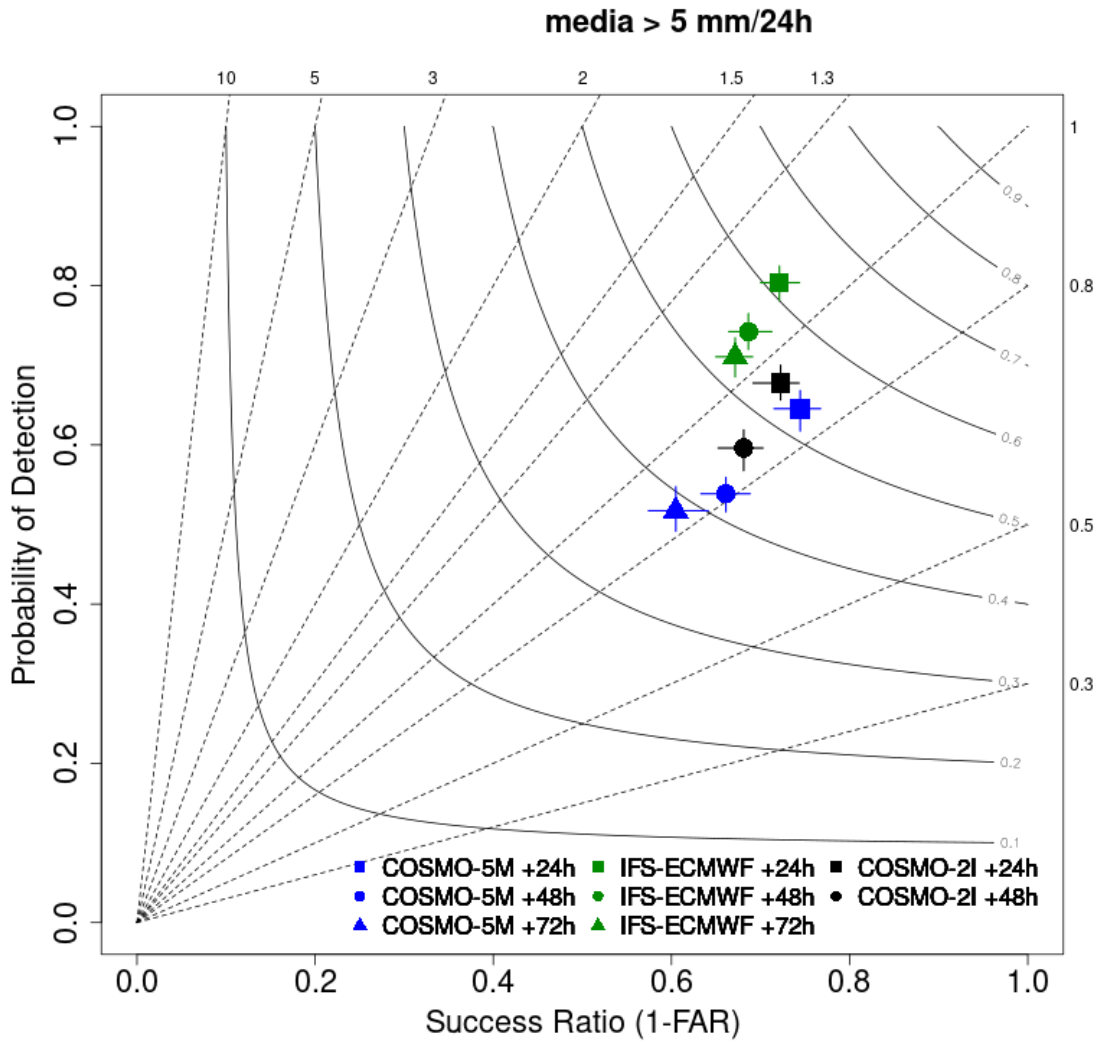




DJF2019-20

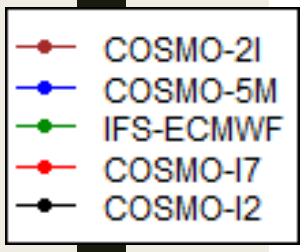
MEAN

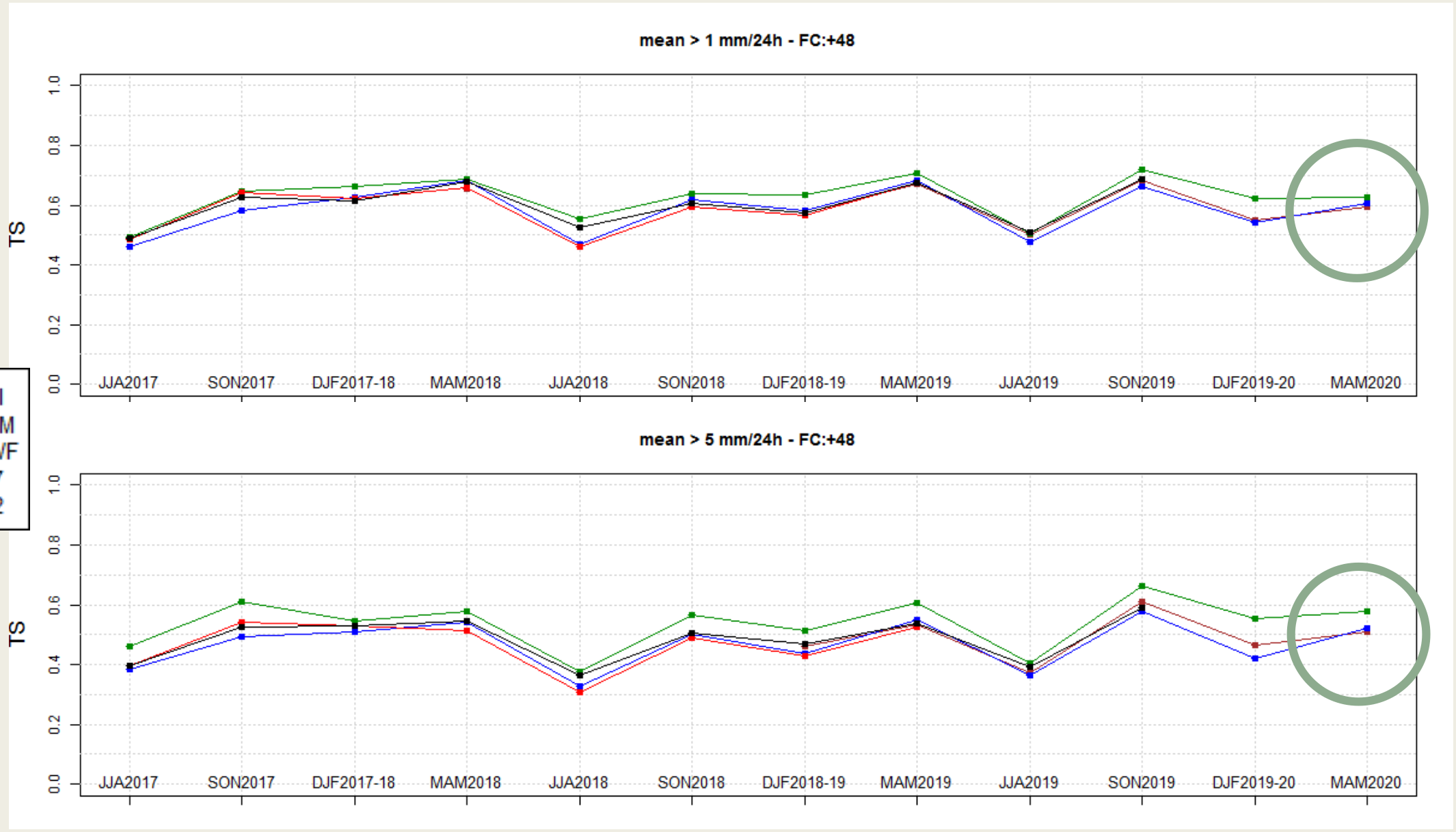




QPF: TS trend

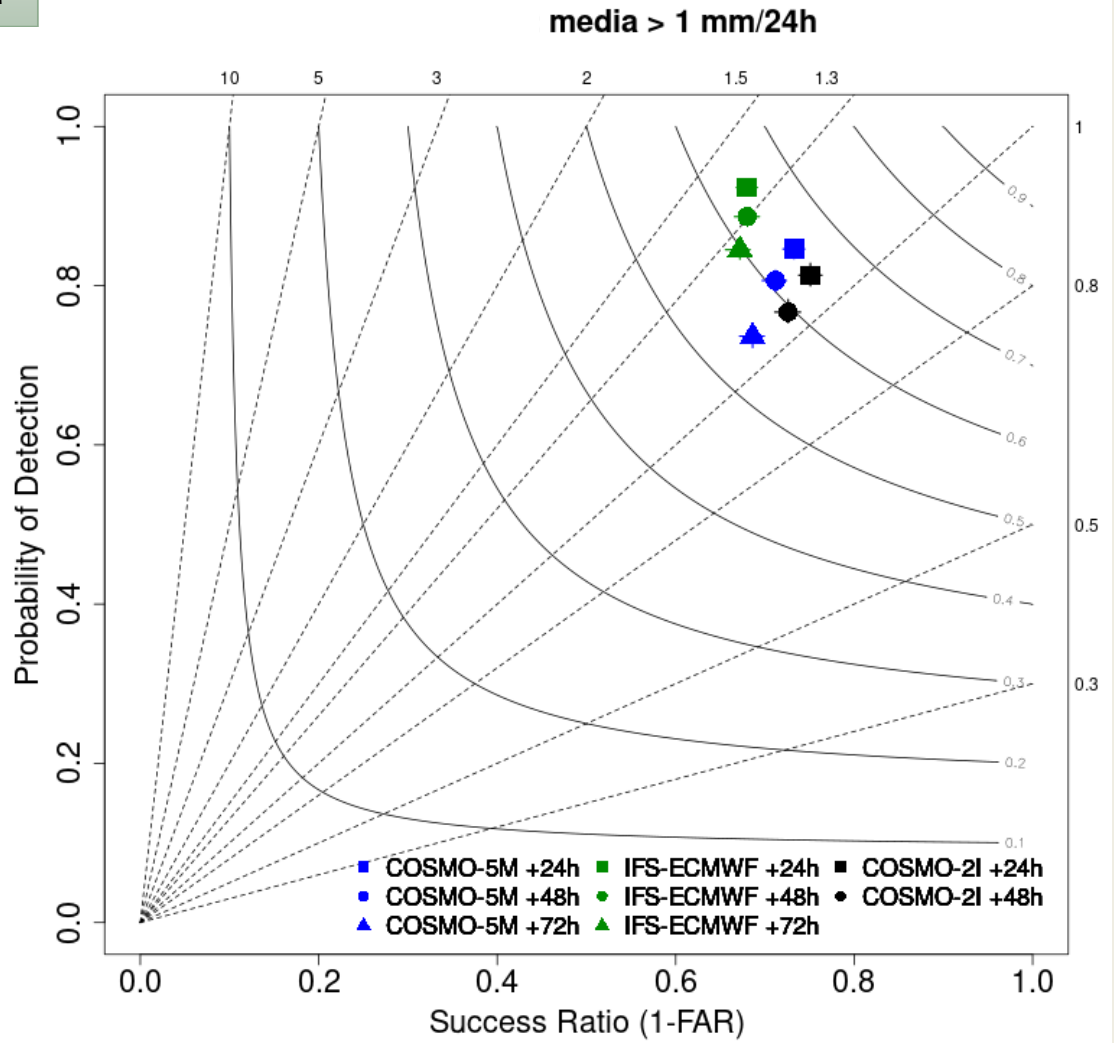
MEAN

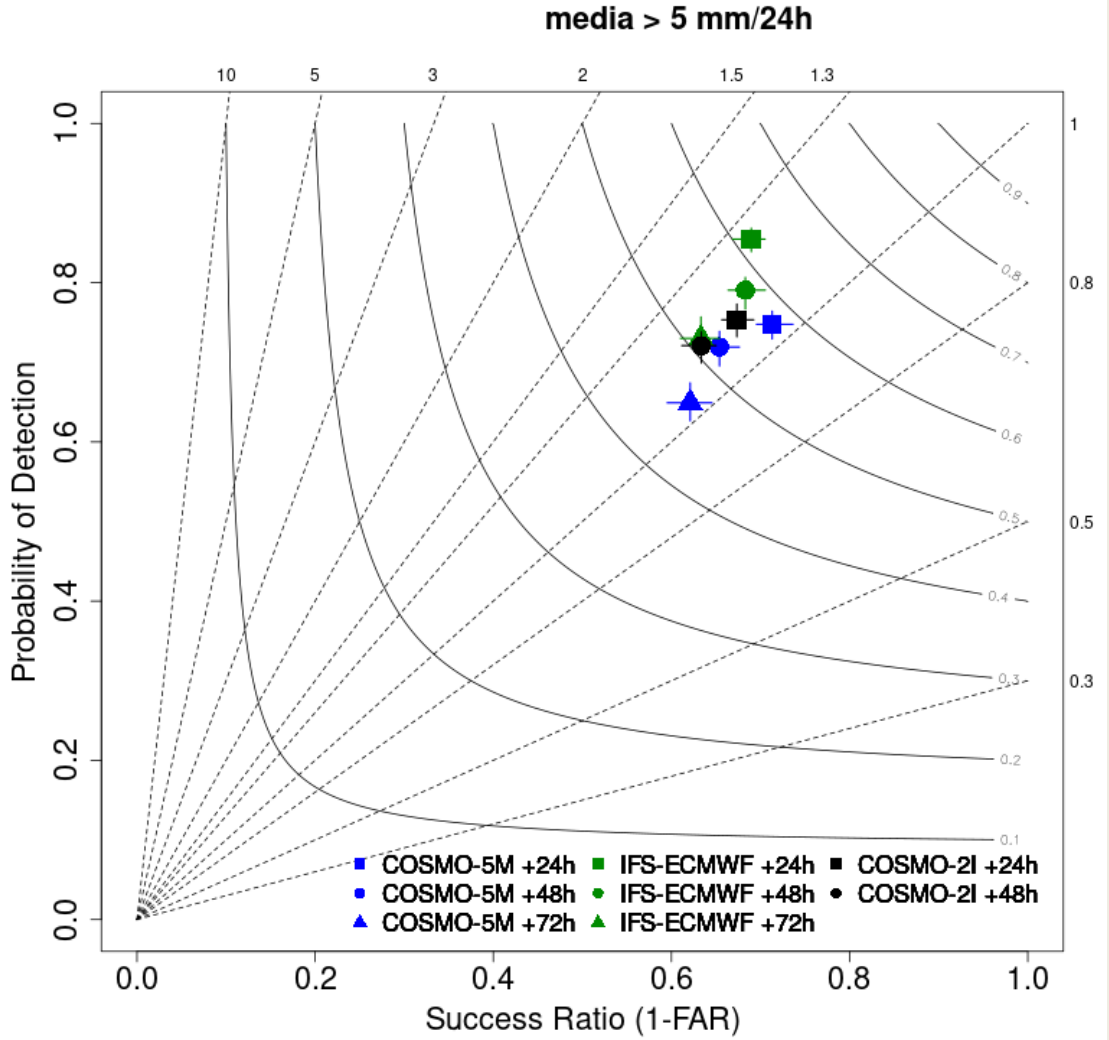




MAM2020

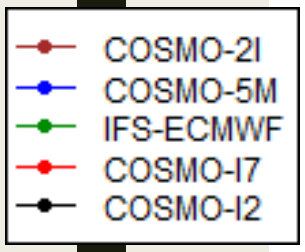
MEAN

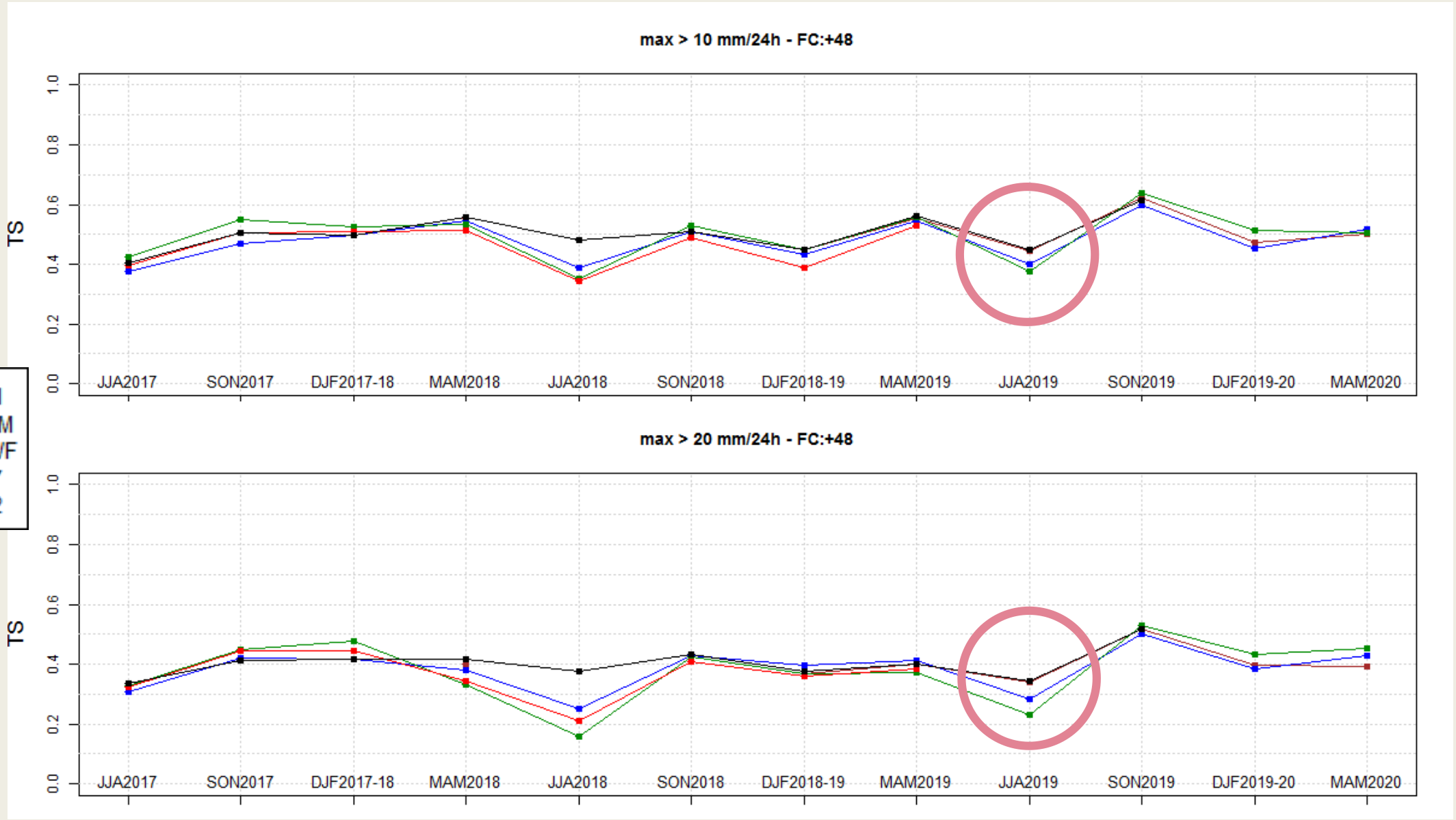




QPF: TS trend

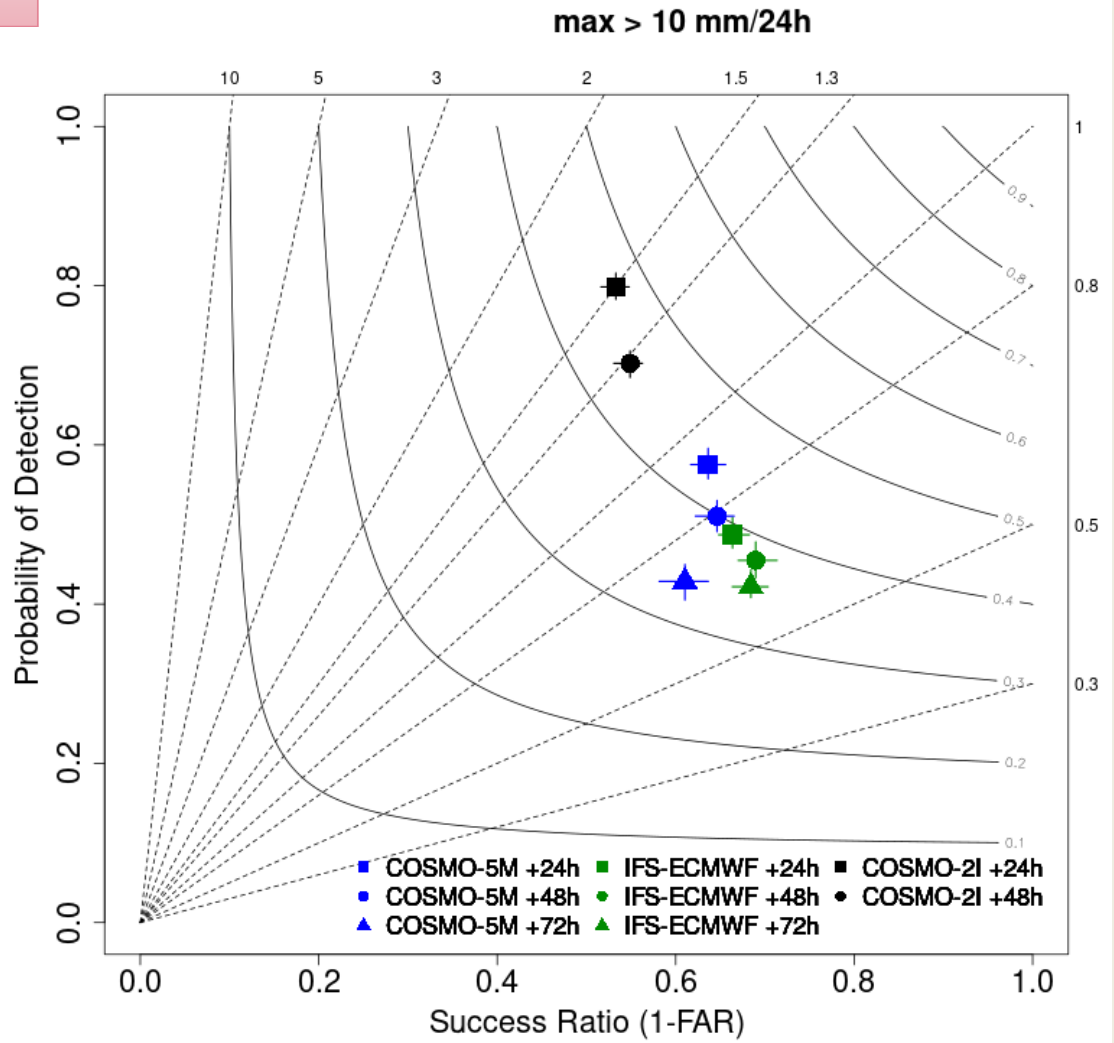
MAX

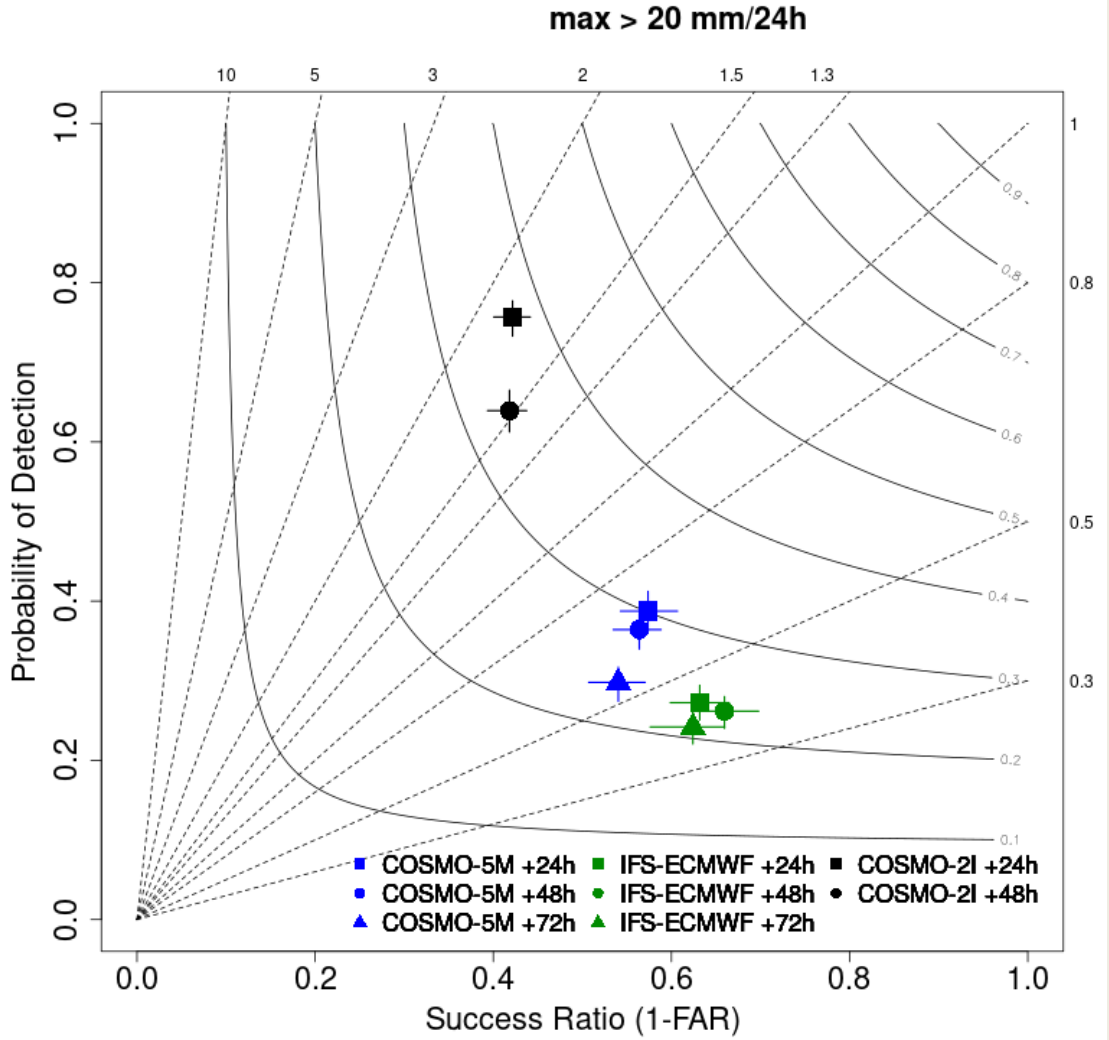




JJA2019

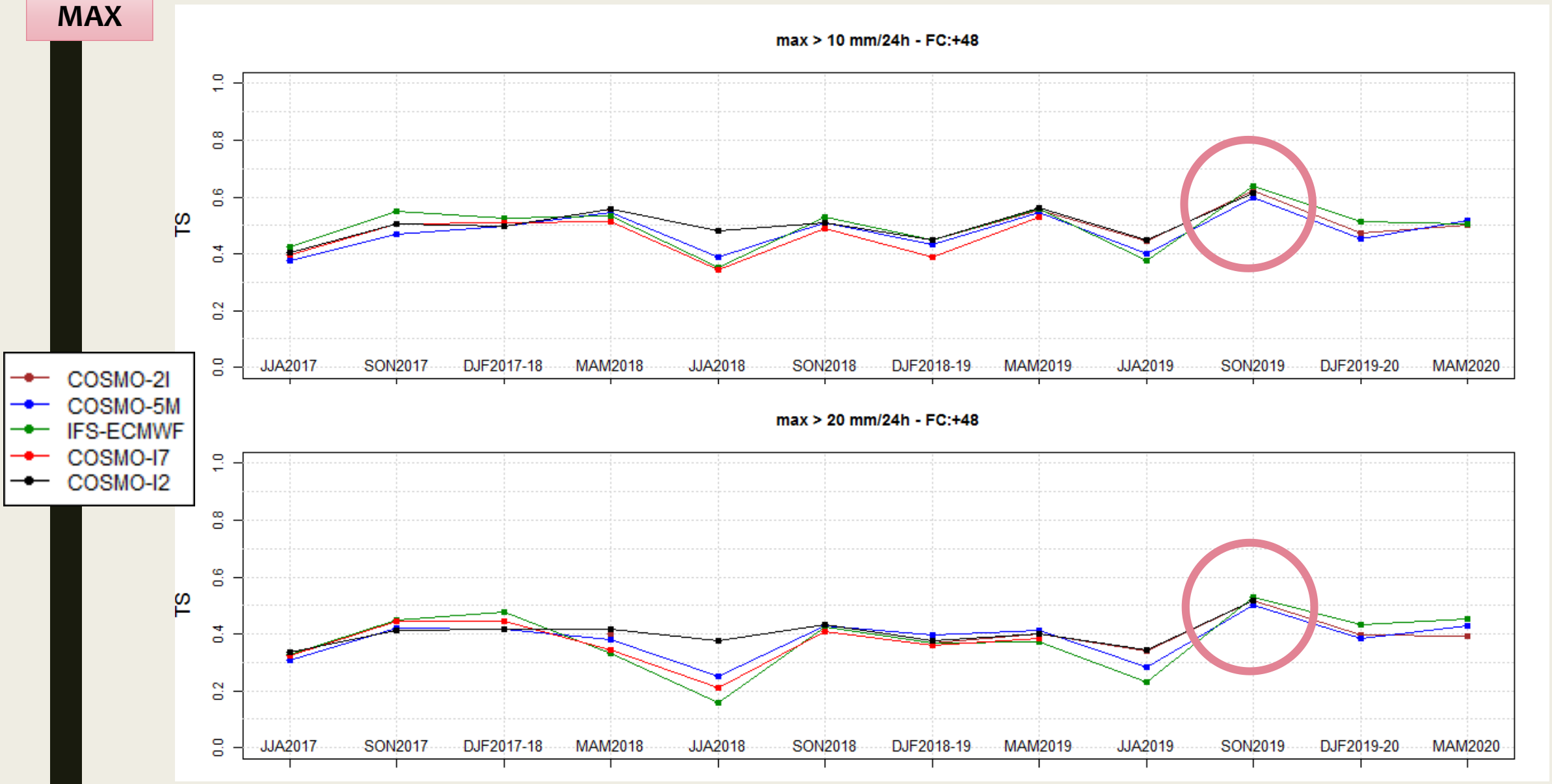
MAX


max > 10 mm/24h
Probability of Detection
Success Ratio (1-FAR)
COSMO-5M +24h
IFS-ECMWF +24h
COSMO-2I +24h
COSMO-5M +48h
IFS-ECMWF +48h
COSMO-2I +48h
COSMO-5M +72h
IFS-ECMWF +72h


max > 20 mm/24h
Probability of Detection
Success Ratio (1-FAR)
COSMO-5M +24h
IFS-ECMWF +24h
COSMO-2I +24h
COSMO-5M +48h
IFS-ECMWF +48h
COSMO-2I +48h
COSMO-5M +72h
IFS-ECMWF +72h

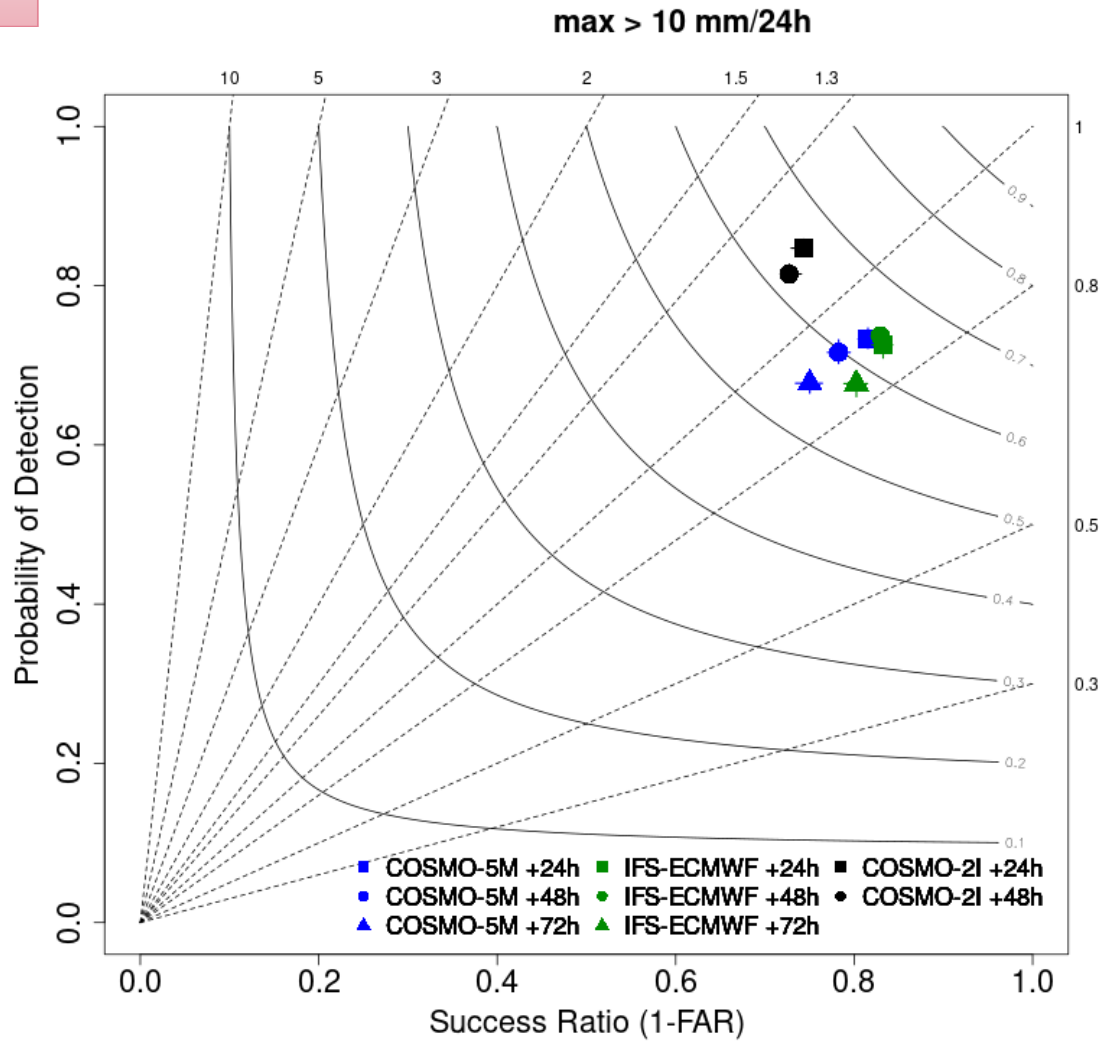
QPF: TS trend

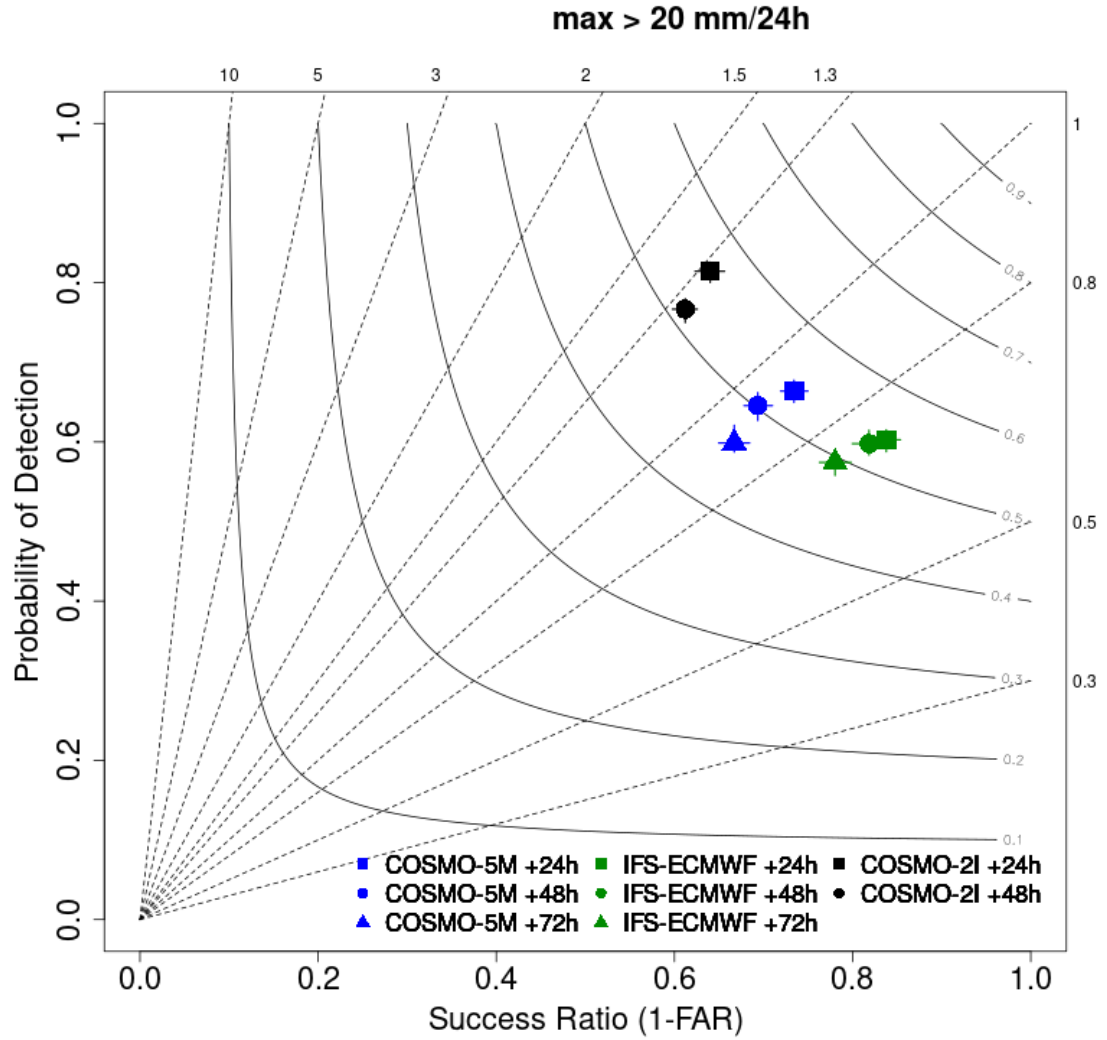
MAX



SON2019

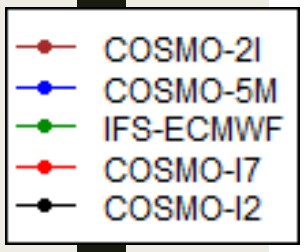
MAX

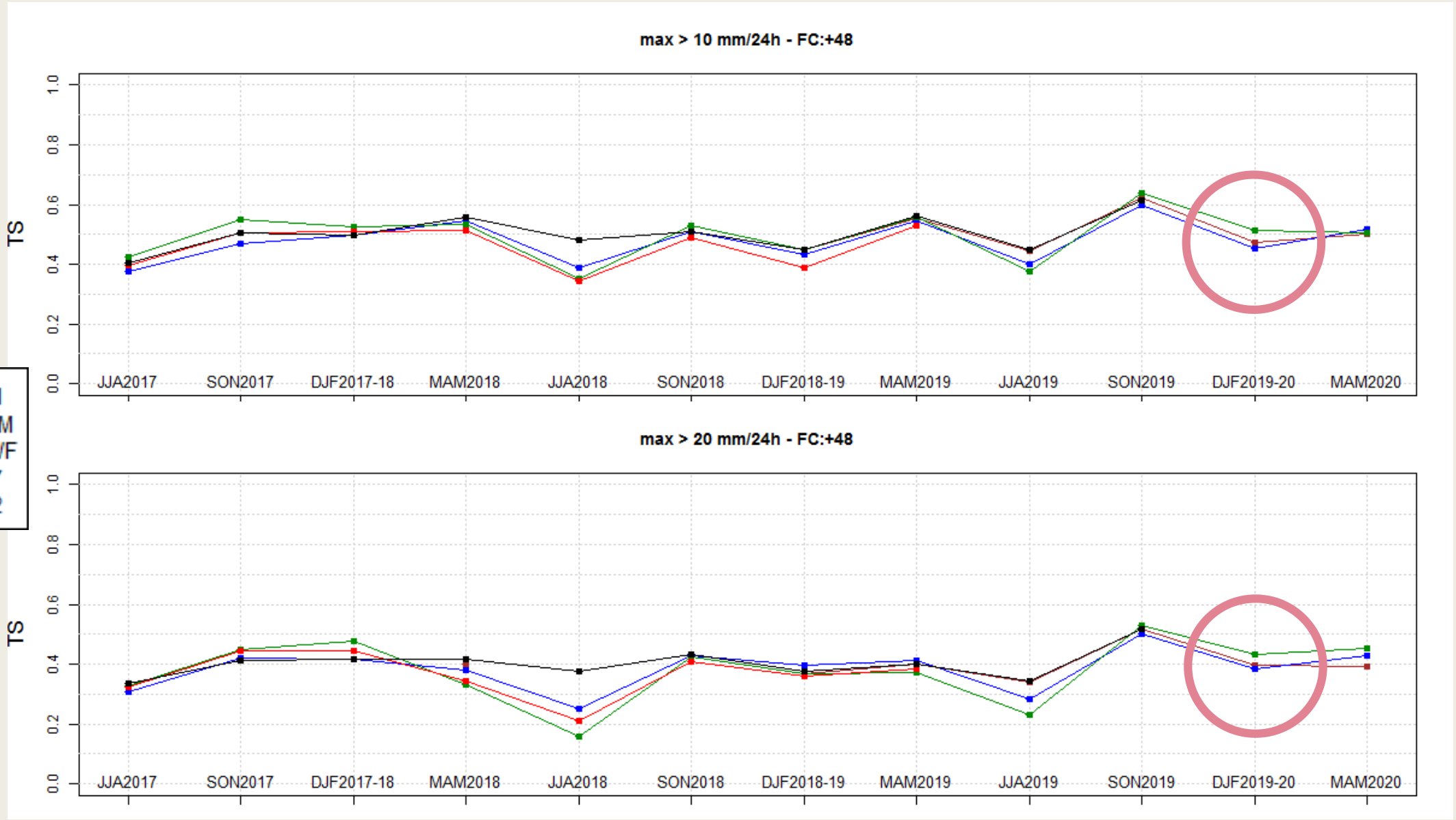

max > 10 mm/24h
Probability of Detection
Success Ratio (1-FAR)
COSMO-5M +24h
IFS-ECMWF +24h
COSMO-2I +24h
COSMO-5M +48h
IFS-ECMWF +48h
COSMO-2I +48h
COSMO-5M +72h
IFS-ECMWF +72h


max > 20 mm/24h
Probability of Detection
Success Ratio (1-FAR)
COSMO-5M +24h
IFS-ECMWF +24h
COSMO-2I +24h
COSMO-5M +48h
IFS-ECMWF +48h
COSMO-2I +48h
COSMO-5M +72h
IFS-ECMWF +72h

QPF: TS trend

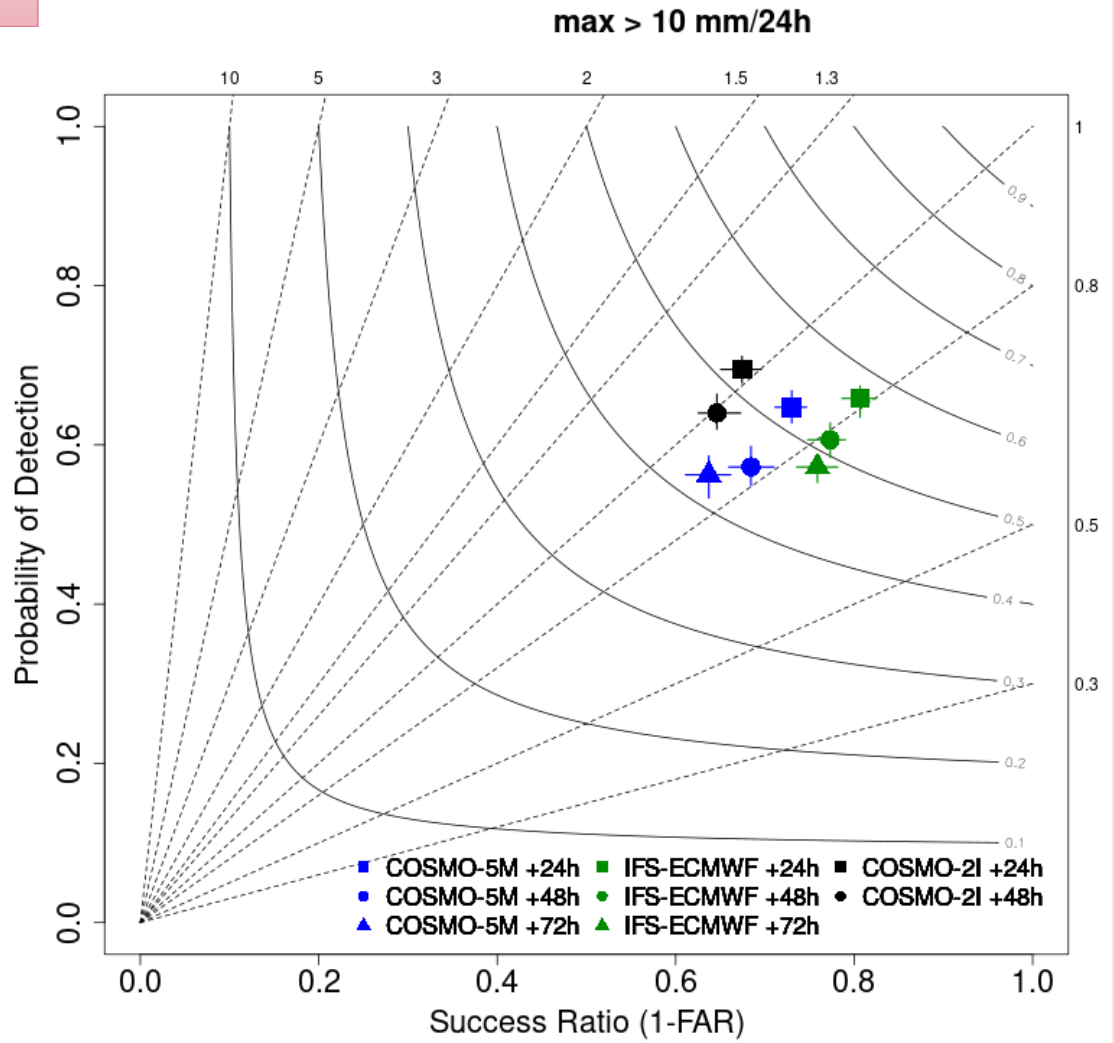
MAX

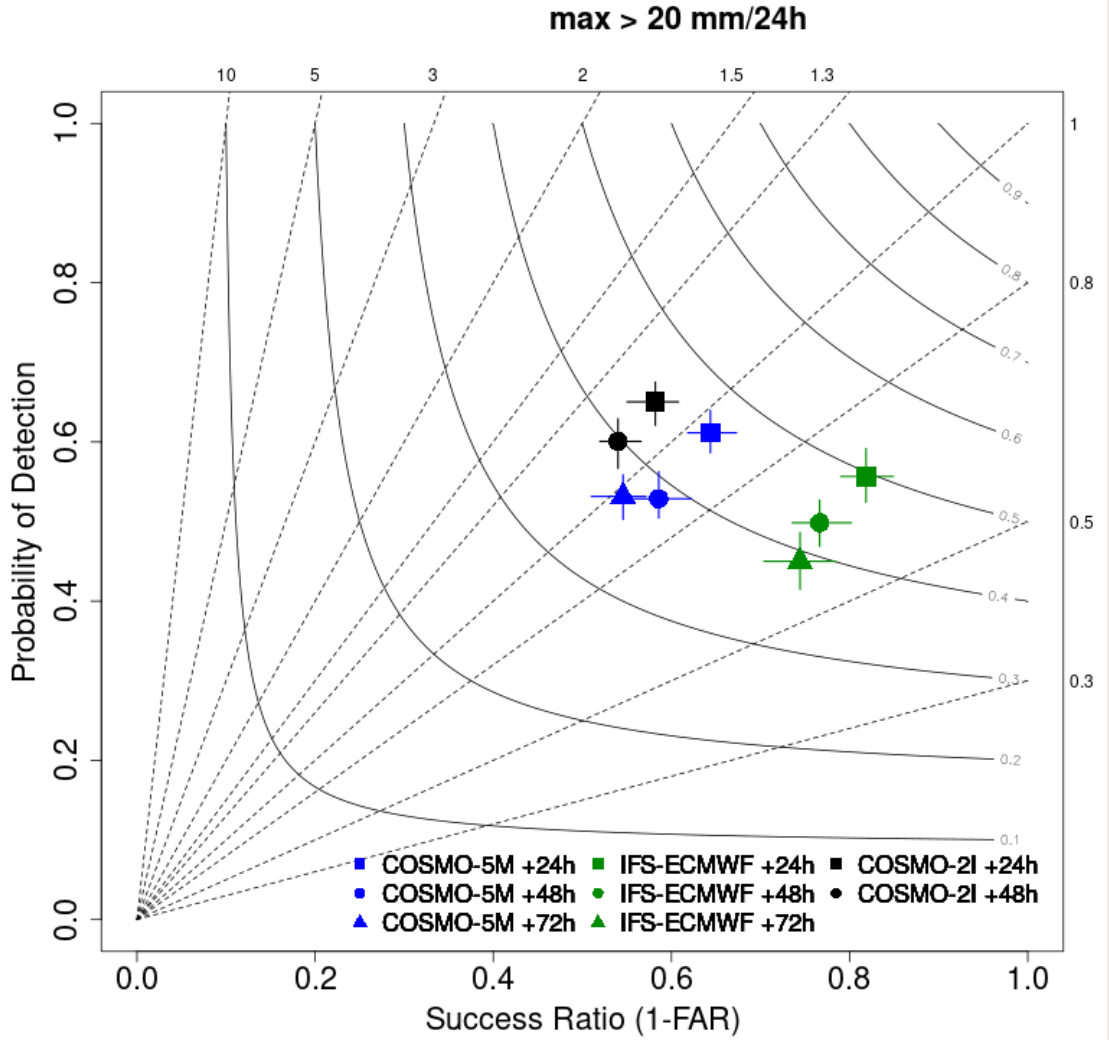




DJF2019-20

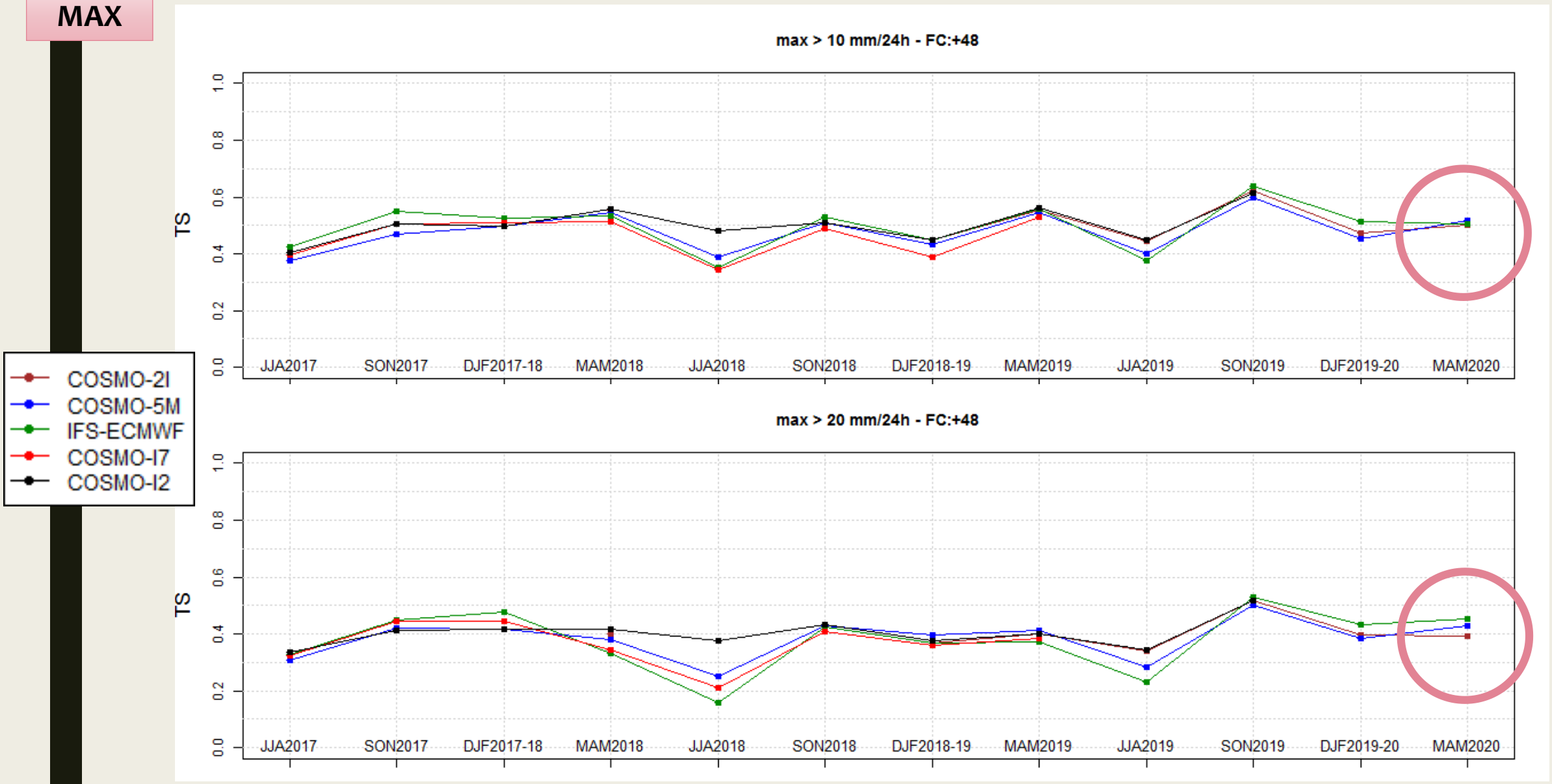
MAX





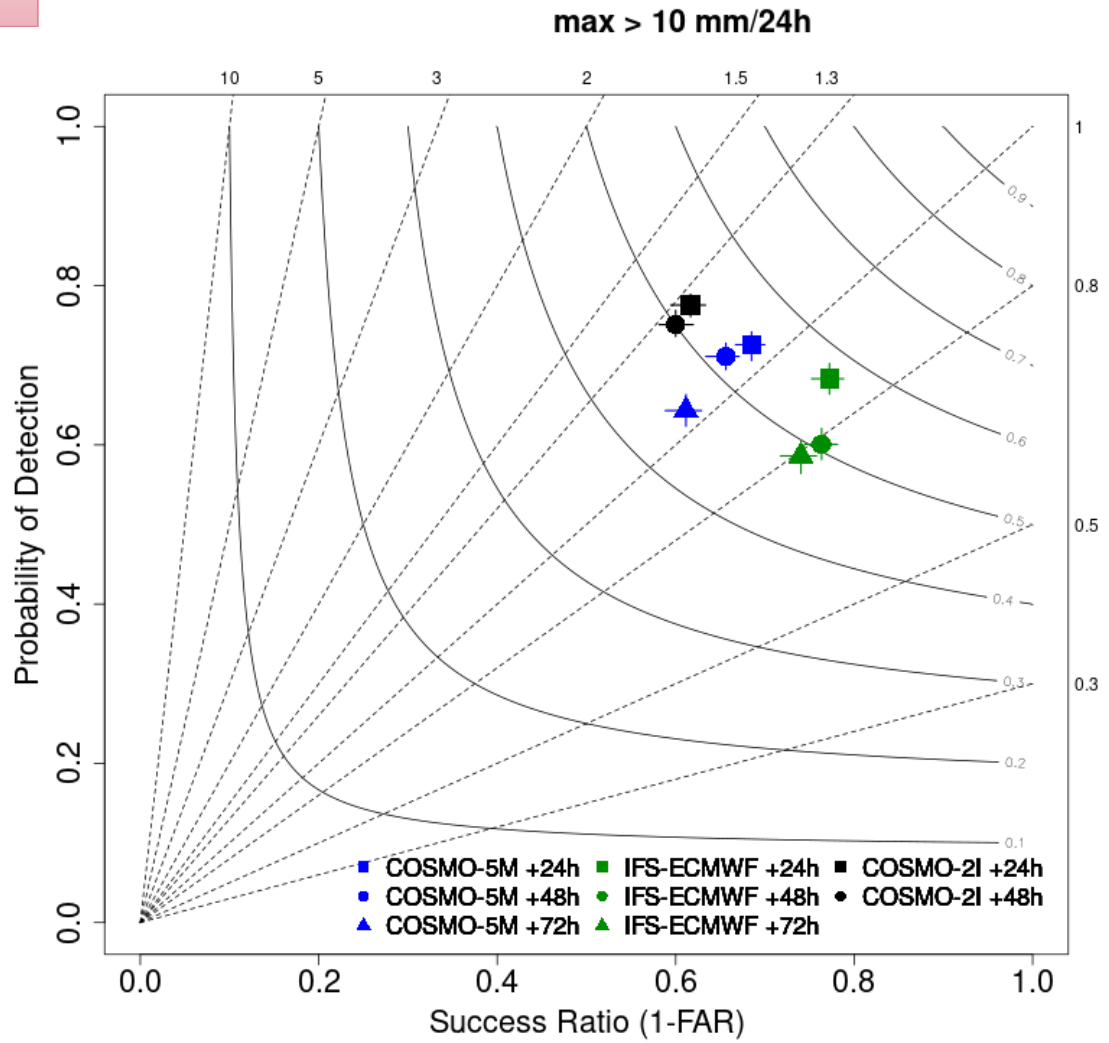
QPF: TS trend

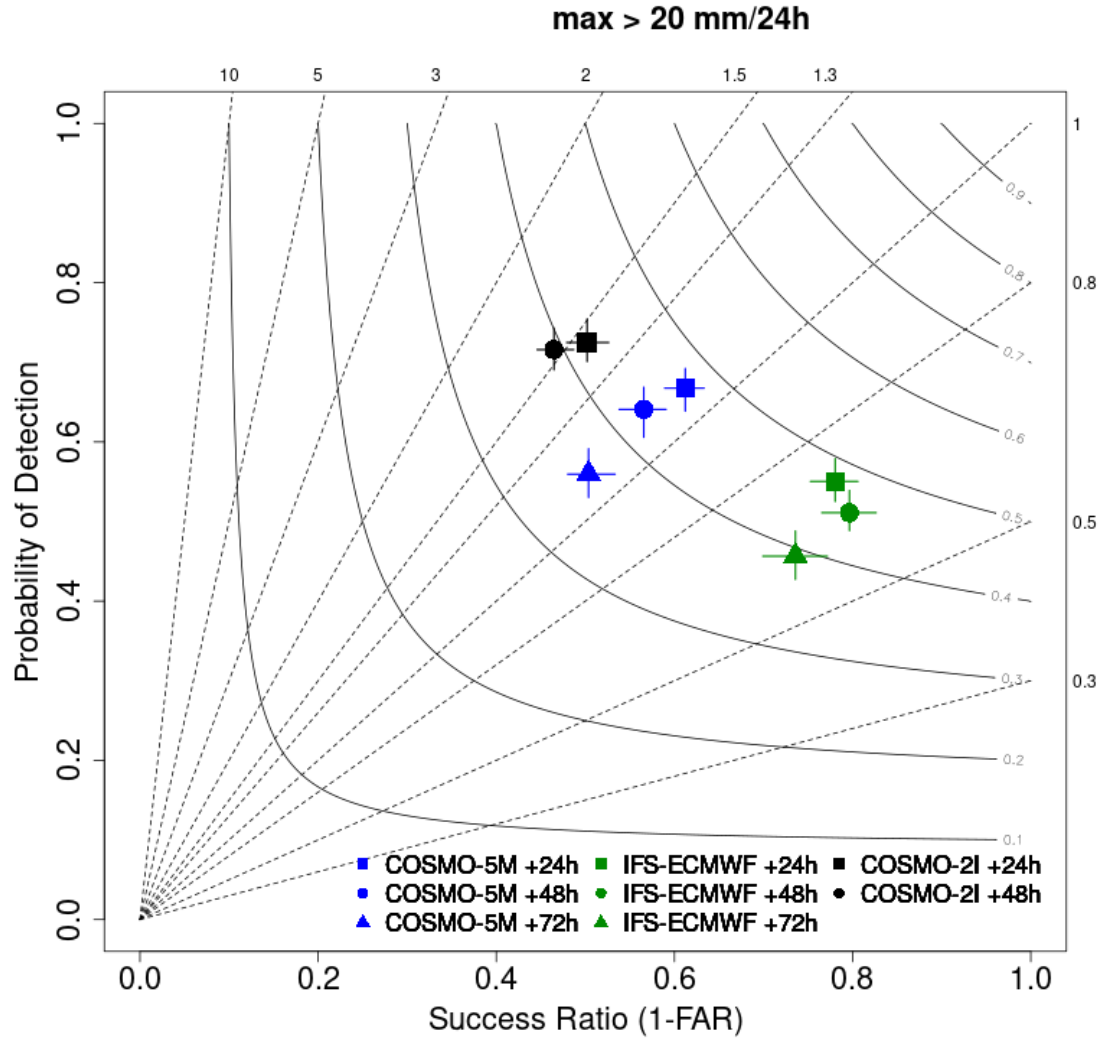
MAX



MAM2020

MAX

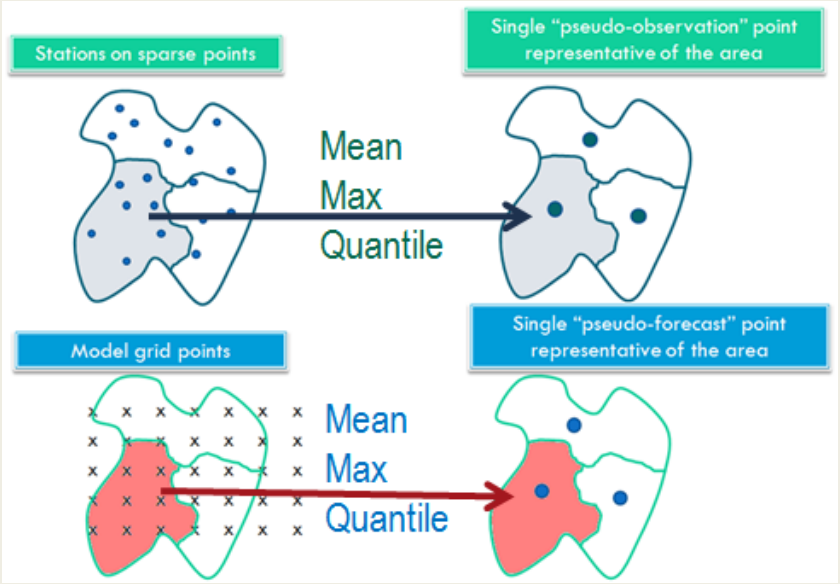

max > 10 mm/24h
Probability of Detection
Success Ratio (1-FAR)
COSMO-5M +24h
IFS-ECMWF +24h
COSMO-2I +24h
COSMO-5M +48h
IFS-ECMWF +48h
COSMO-2I +48h
COSMO-5M +72h
IFS-ECMWF +72h


max > 20 mm/24h
Probability of Detection
Success Ratio (1-FAR)
COSMO-5M +24h
IFS-ECMWF +24h
COSMO-2I +24h
COSMO-5M +48h
IFS-ECMWF +48h
COSMO-2I +48h
COSMO-5M +72h
IFS-ECMWF +72h

User oriented verification

- Observed and forecast precipitation, aggregated on the catchment areas, have been divided into classes

CLASSES FOR MEAN PRECIPITATION	MEAN AMOUNT IN 24h (mm)
NO PRECIPITATION	<0.2
NON SIGNIFICANT	0.2 – 5
LIGHT	5-20
MODERATE	20-45
HEAVY	>45

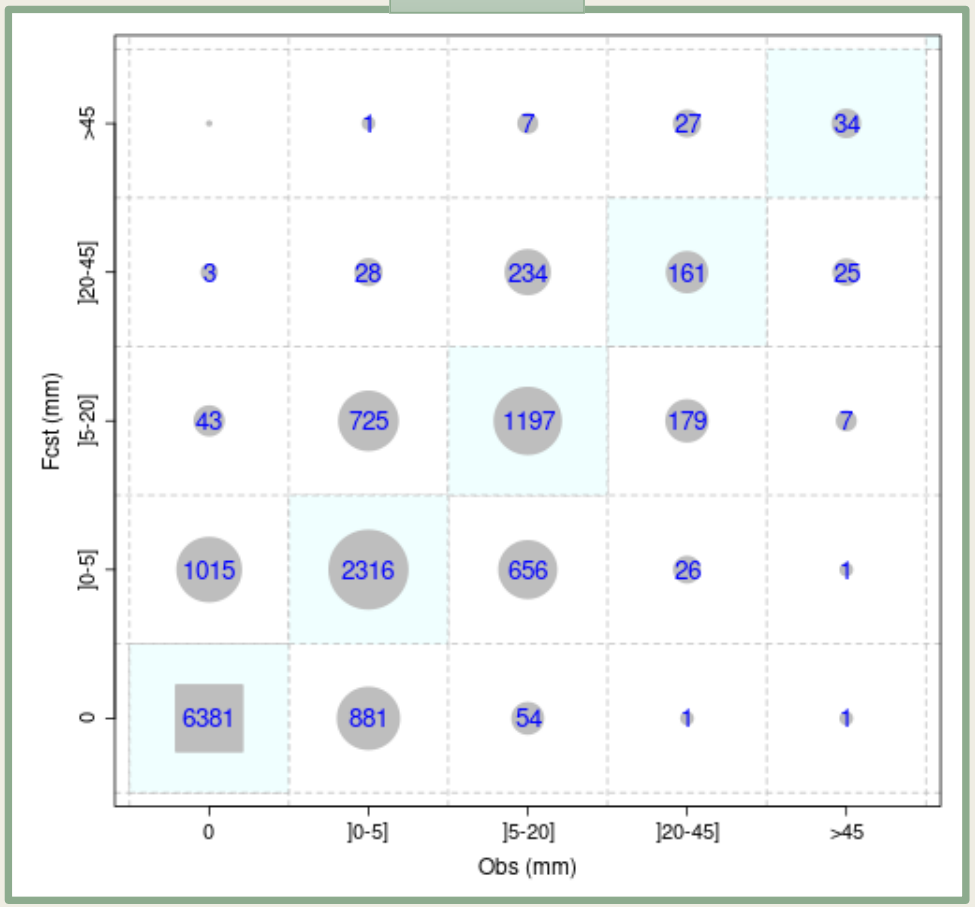


CLASSES FOR MAX PRECIPITATION							
MAX AMOUNT IN 24h (mm)	0.2 -5	5-25	25-50	50-75	75-100	100-150	>150

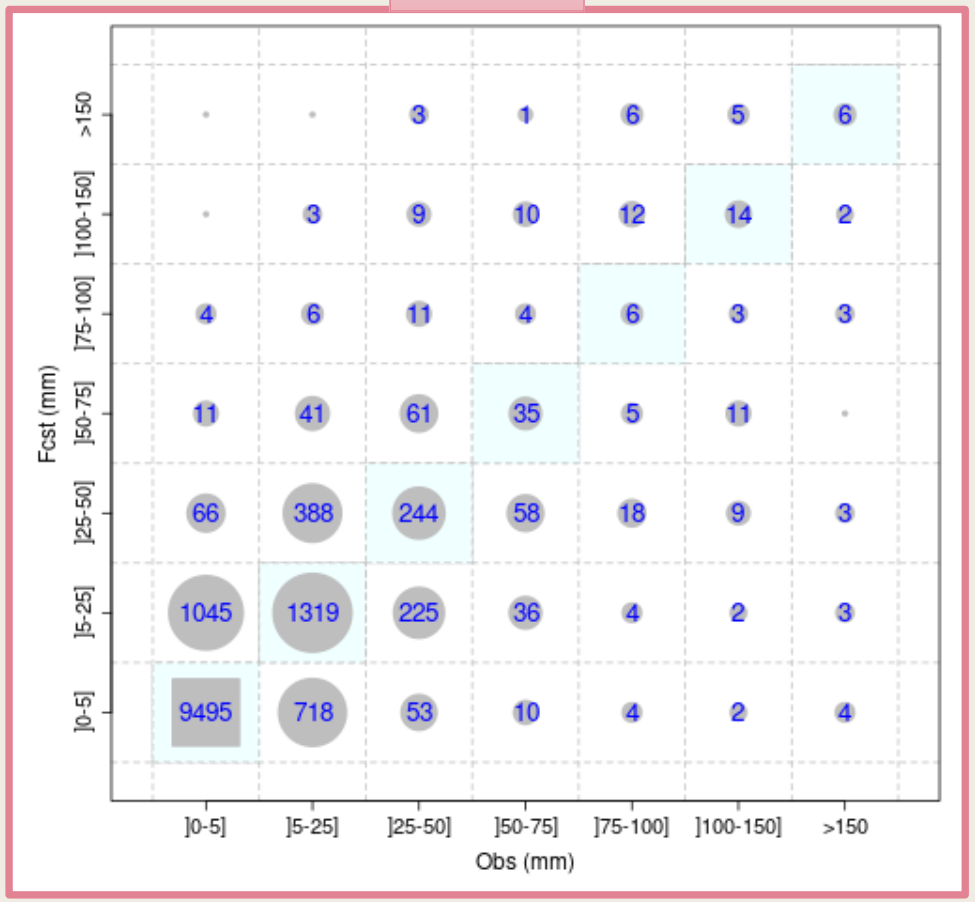
Visual verification with “bubble plots”

- Bubble plot is a sort of the scatter plot, in which the data points are replaced with bubbles. The sizes of the bubbles are determined by the number of events.
  (The square symbol is used for the most populated category to preserve the proportions of the other bubbles)

MEAN



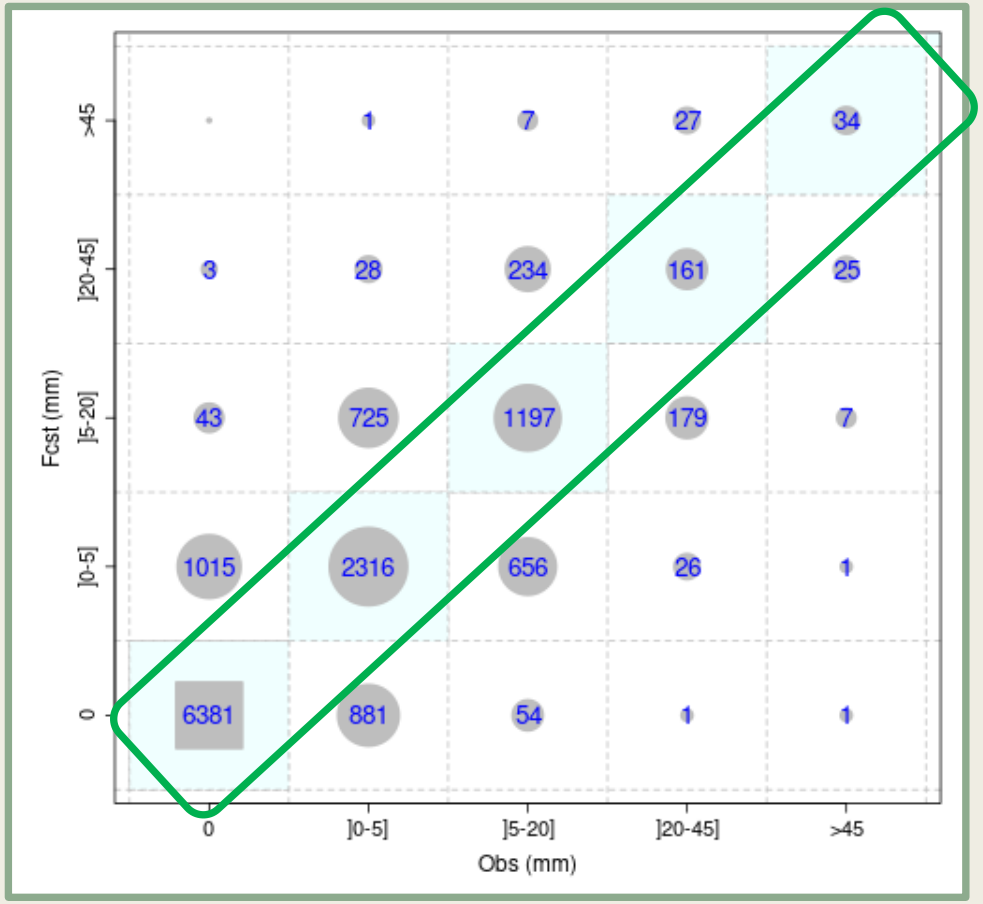
MAX

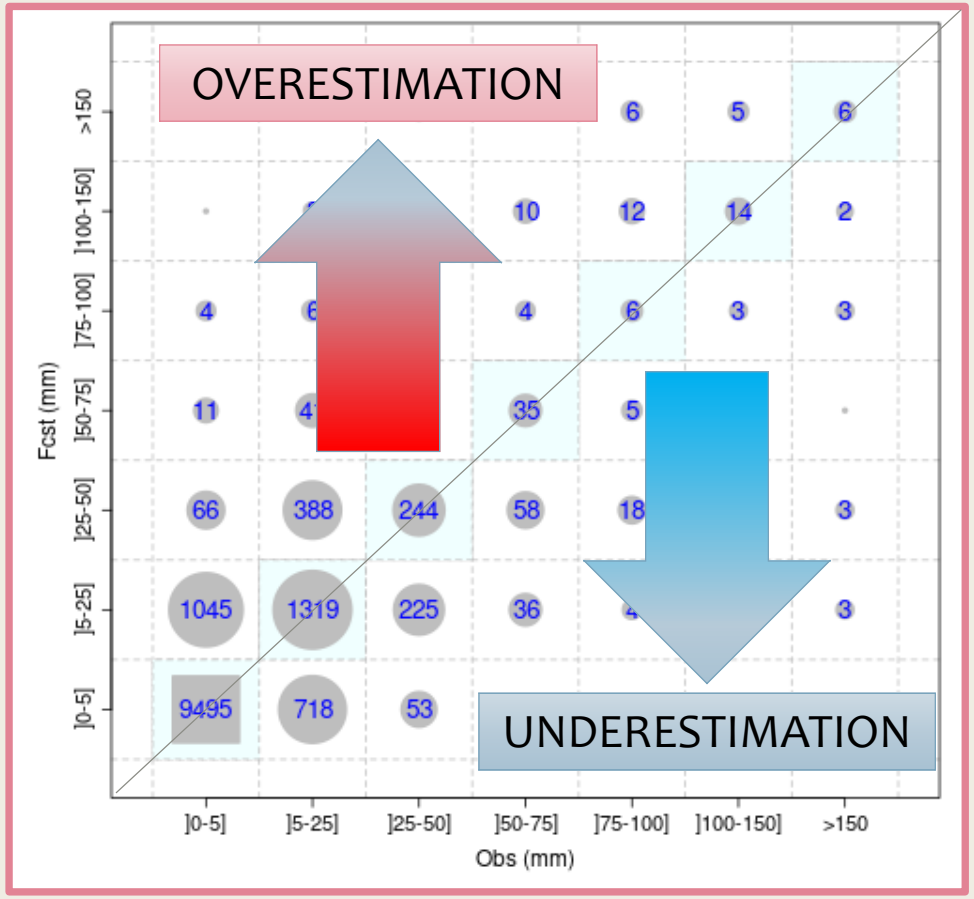


Visual verification with “bubble plots”

- The advantage of this approach is that the nature of the forecast errors can more easily be diagnosed

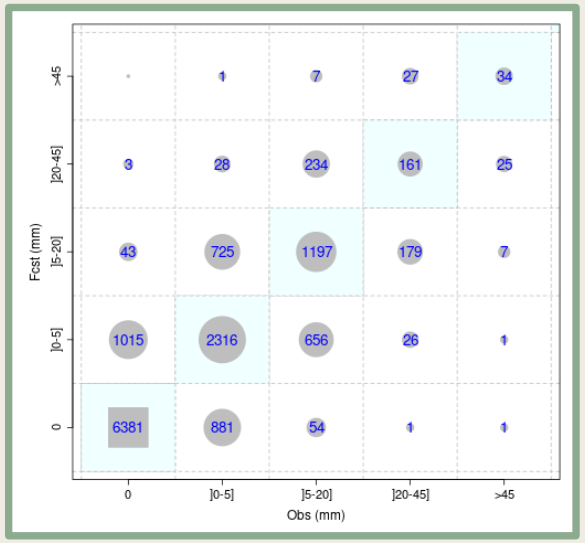
✓ CORRECT FORECASTS LIE ON THE DIAGONAL





Multi-category forecast verification

- Bubbles blot can be viewed as a multi-category contingency table
- Even if there are fewer statistics that summarize the performance of multi-category forecasts than for dichotomous (yes/no) forecasts, is possible to condense the results into a single number:
  - The choice fell on the **Gerrity Score**



Multi-category Contingency Table

		Observed Category				Total
	i,j	1	2	...	K	
Forecast Category	1	$n(F_1,O_1)$	$n(F_1,O_2)$	...	$n(F_1,O_K)$	$N(F_1)$
	2	$n(F_2,O_1)$	$n(F_2,O_2)$	...	$n(F_2,O_K)$	$N(F_2)$
	...	...	...	...	...	...
	K	$n(F_K,O_1)$	$n(F_K,O_2)$	...	$n(F_K,O_K)$	$N(F_K)$
Total		$N(O_1)$	$N(O_2)$	...	$N(O_K)$	N

In this table $n(F_i,O_j)$ denotes the number of forecasts in category i that had observations in category j, $N(F_i)$ denotes the total number of forecasts in category i, $N(O_j)$ denotes the total number of observations in category j, and N is the total number of forecasts.

Gerrity Score

- **Answers the question:** *What was the accuracy of the forecast in predicting the correct category, relative to that of random chance?*
- **Range:** -1 to 1, 0 indicates no skill. Perfect score: 1
- **Characteristics:** Uses all entries in the contingency table, does not depend on the forecast distribution, and is equitable (i.e., random and constant forecasts score a value of 0).
  GS does not reward conservative forecasting like HSS and HK, but rather rewards forecasts for correctly predicting the less likely categories.
  Smaller errors are penalized less than larger forecast errors. This is achieved through the use of the scoring matrix

Gerrity score - $GS = \frac{1}{N}\sum_{i=1}^{K}\sum_{j=1}^{K} n(F_i, O_j)\, s_{ij}$

where s_{ij} are elements of a scoring matrix given by

$$s_{ij} = \frac{1}{K-1}\left(\sum_{r=1}^{i-1} a_r^{-1} + \sum_{r=i}^{K-1} a_r\right) \quad (i = j, \text{ diagonal}),$$

$$s_{ij} = s_{ji} = \frac{1}{K-1}\left(\sum_{r=1}^{i-1} a_r^{-1} - (j-i) + \sum_{r=i}^{K-1} a_r\right) \quad (i \neq j, \text{ off-diagonal})$$

$$a_i = \left(1 - \sum_{r=1}^{i} p_r\right) \Big/ \sum_{r=1}^{i} p_r$$

with the sample probabilities (observed frequencies) given by $p_i = N(O_i) / N$

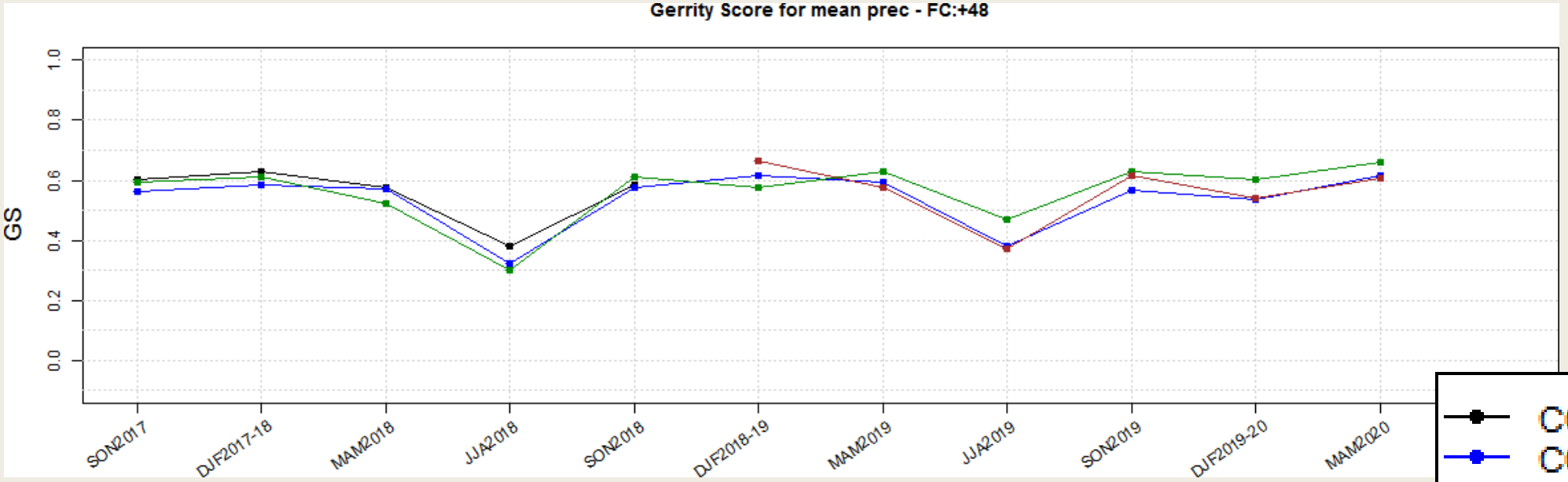
JWGFVR WWRP WCRP World Climate Research Programme

WWRP/WGNE Joint Working Group on Forecast Verification Research

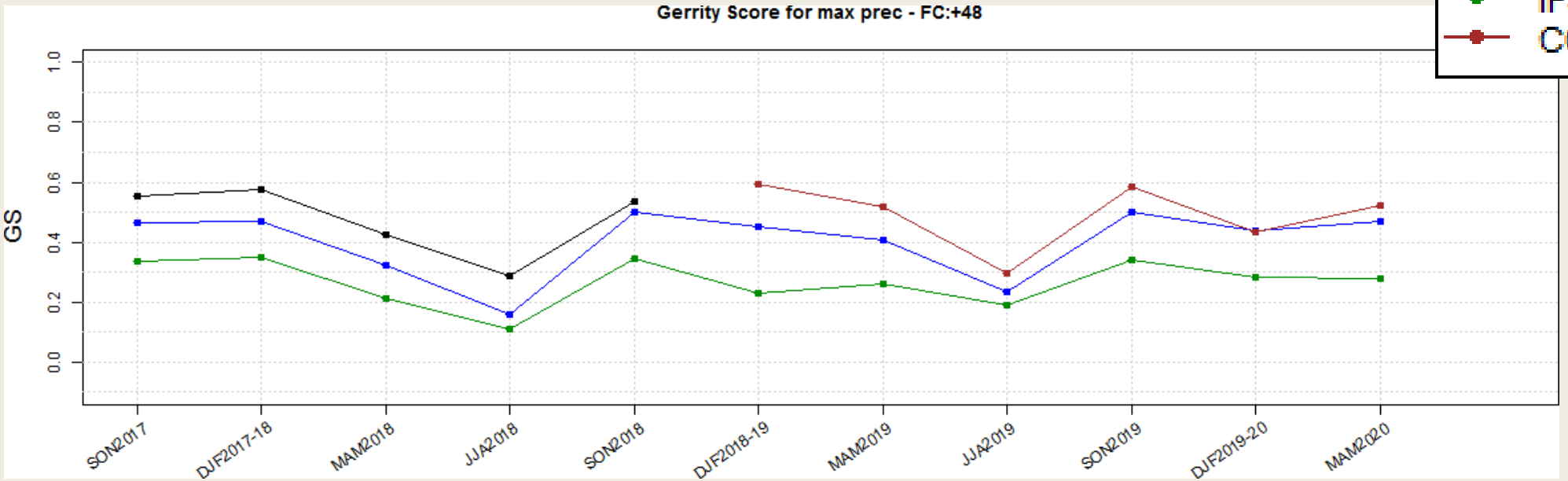
https://www.cawcr.gov.au/projects/verification/

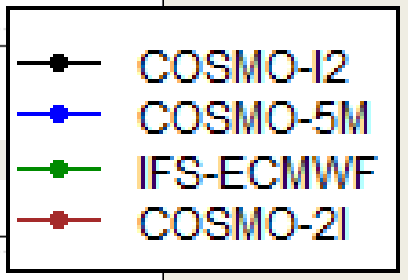
QPF: GS trend

MEAN



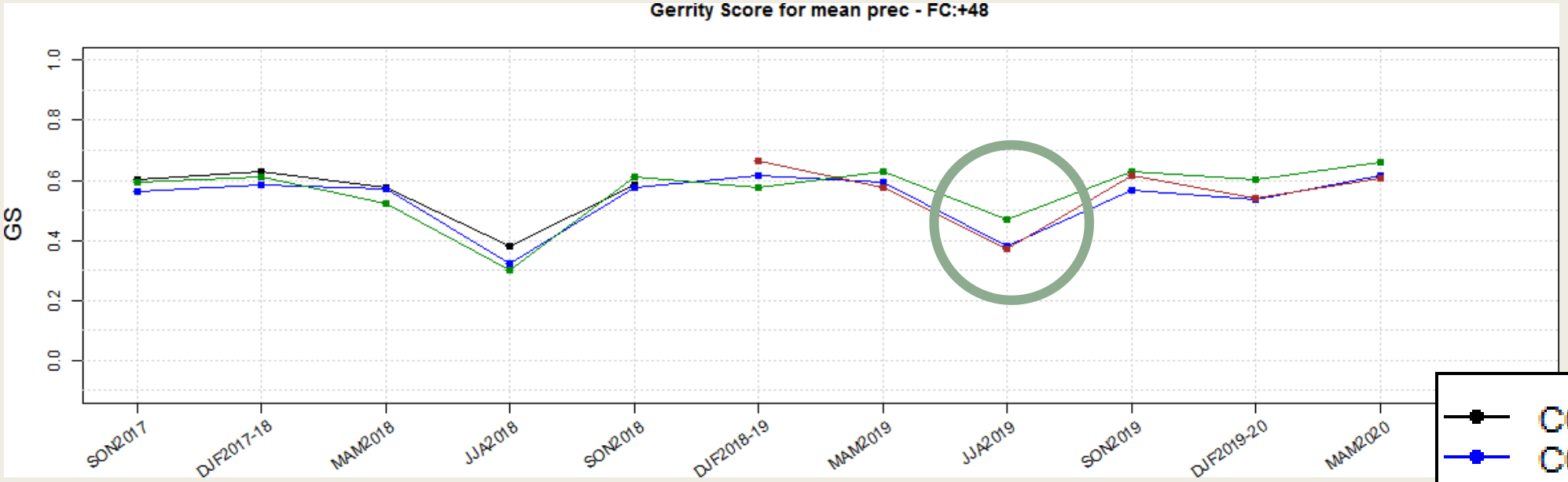
MAX



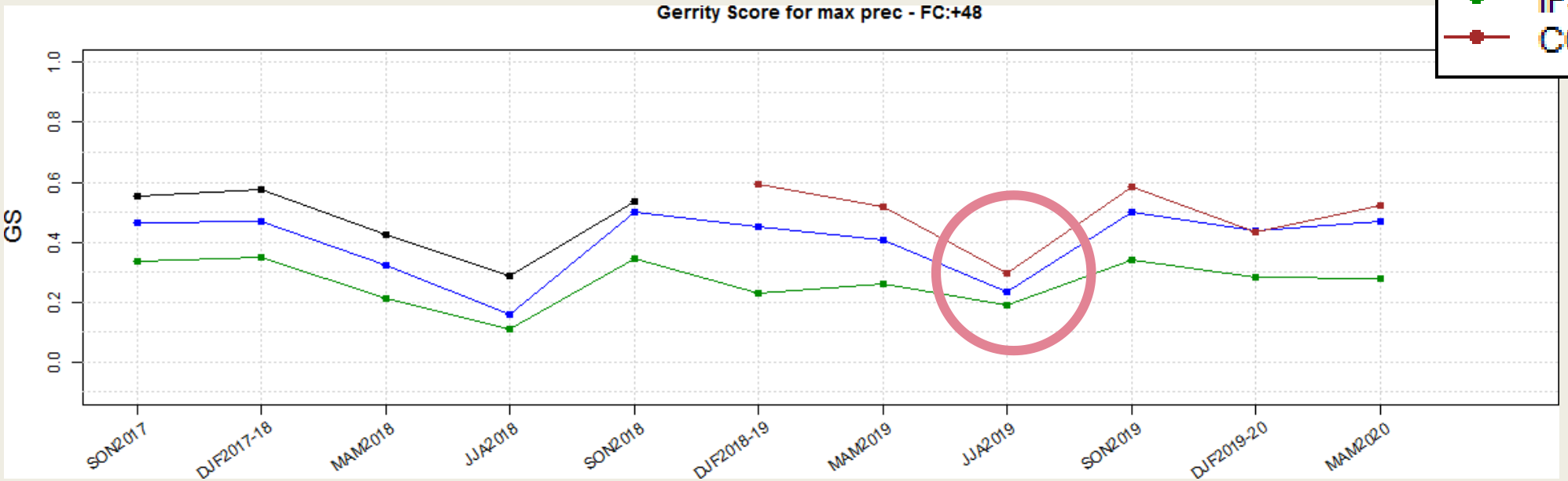


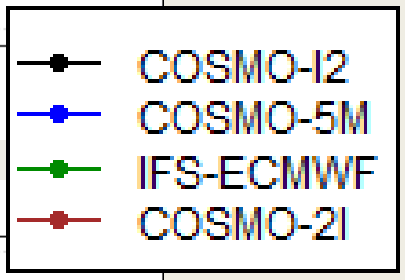
QPF: GS trend

MEAN



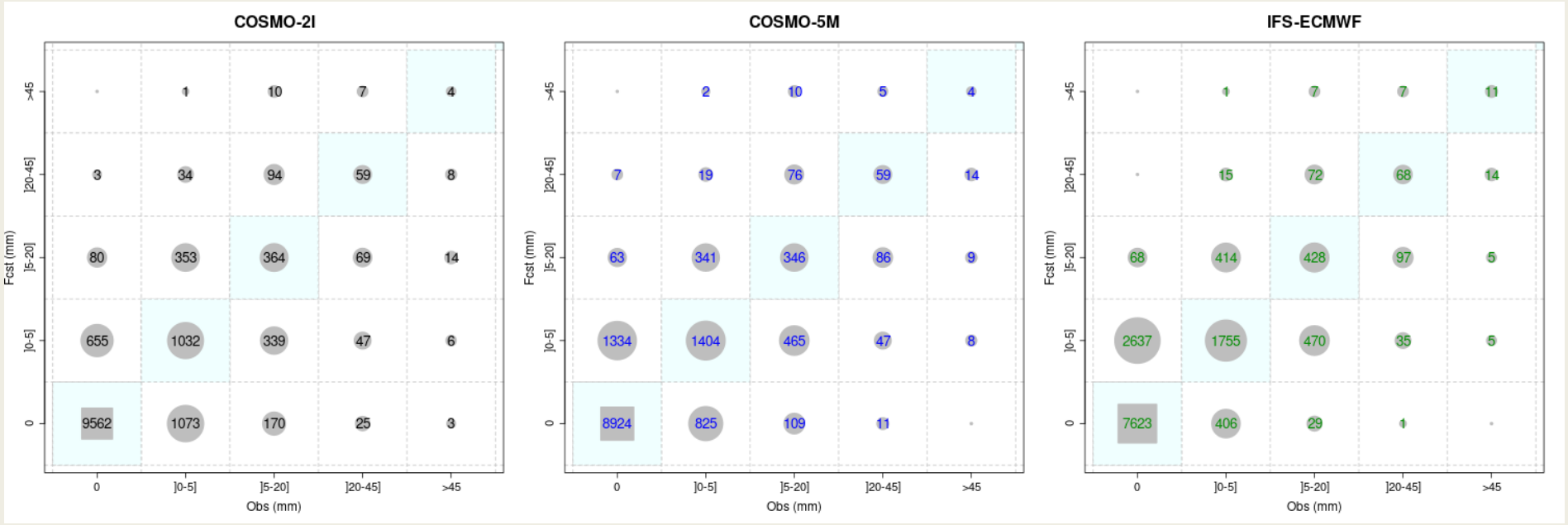
MAX



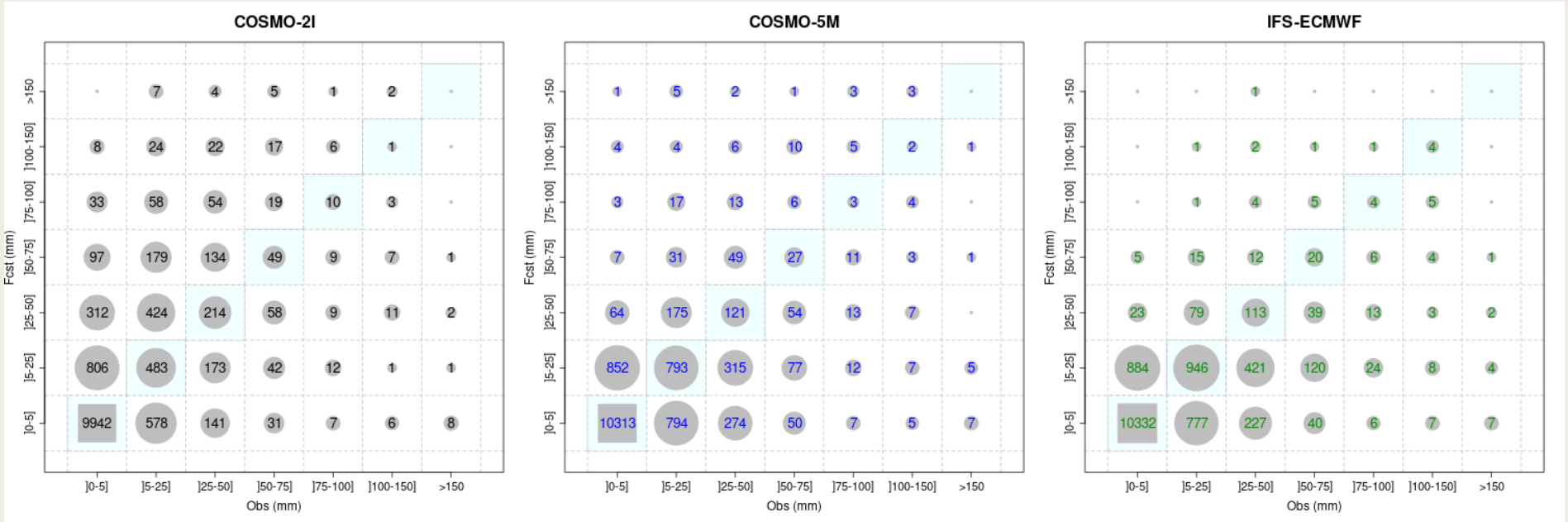


JJA2019

MEAN

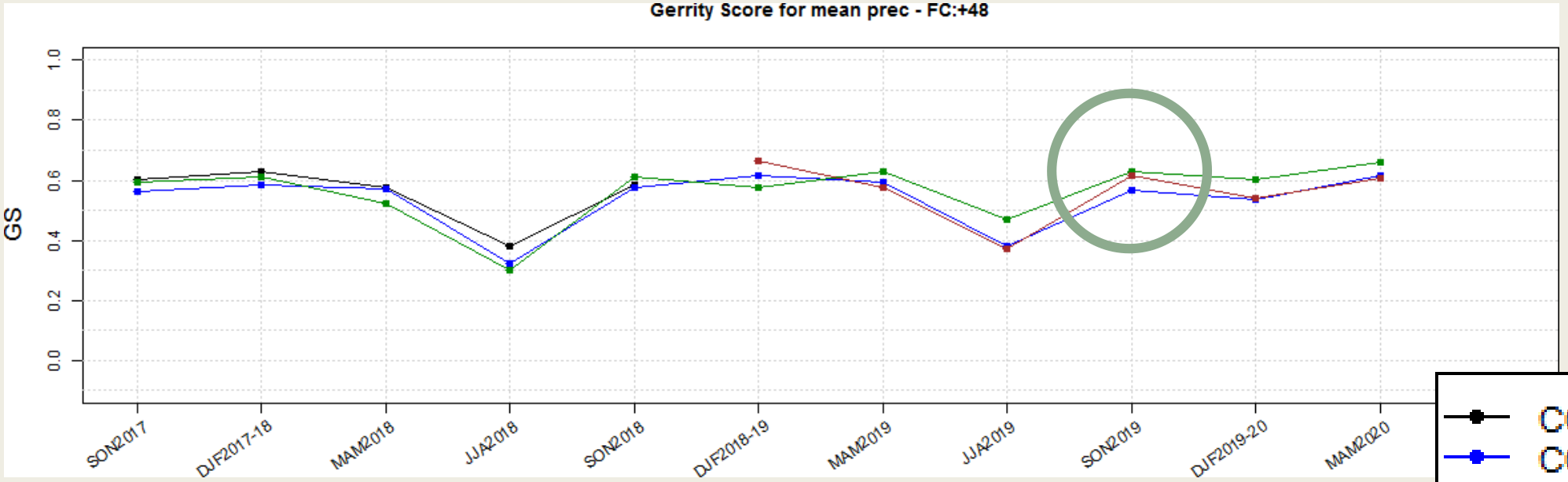


MAX

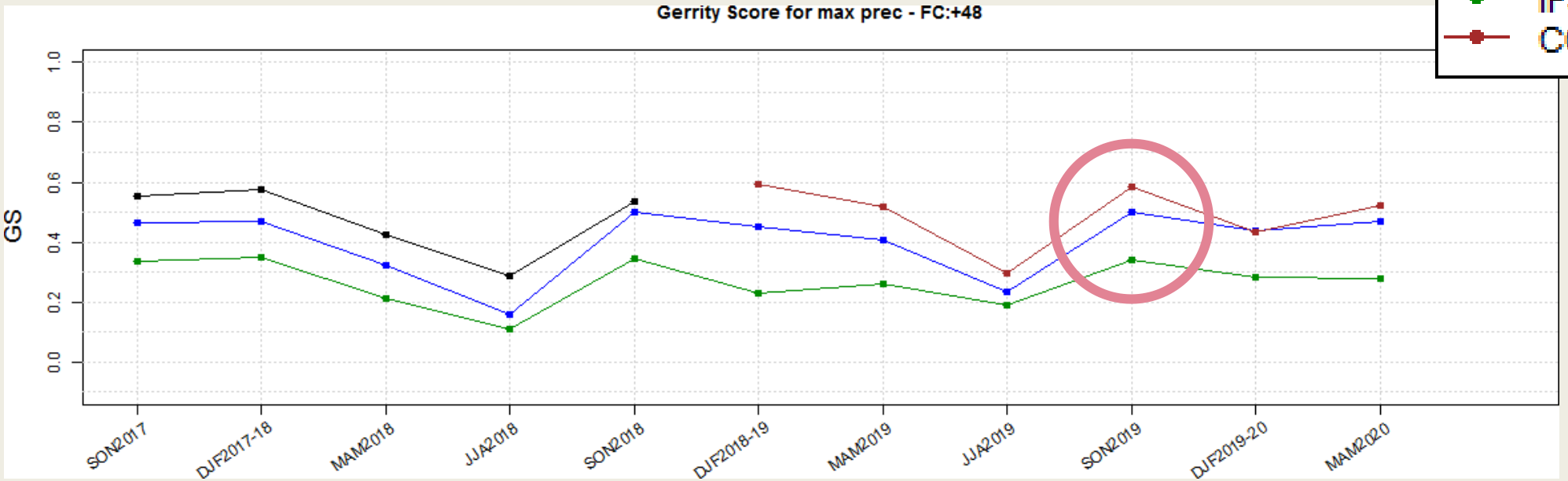


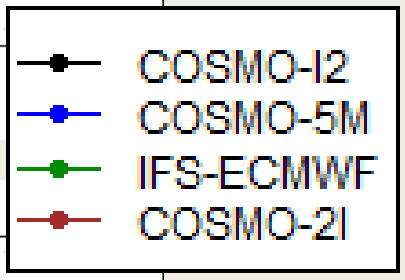
QPF: GS trend

MEAN



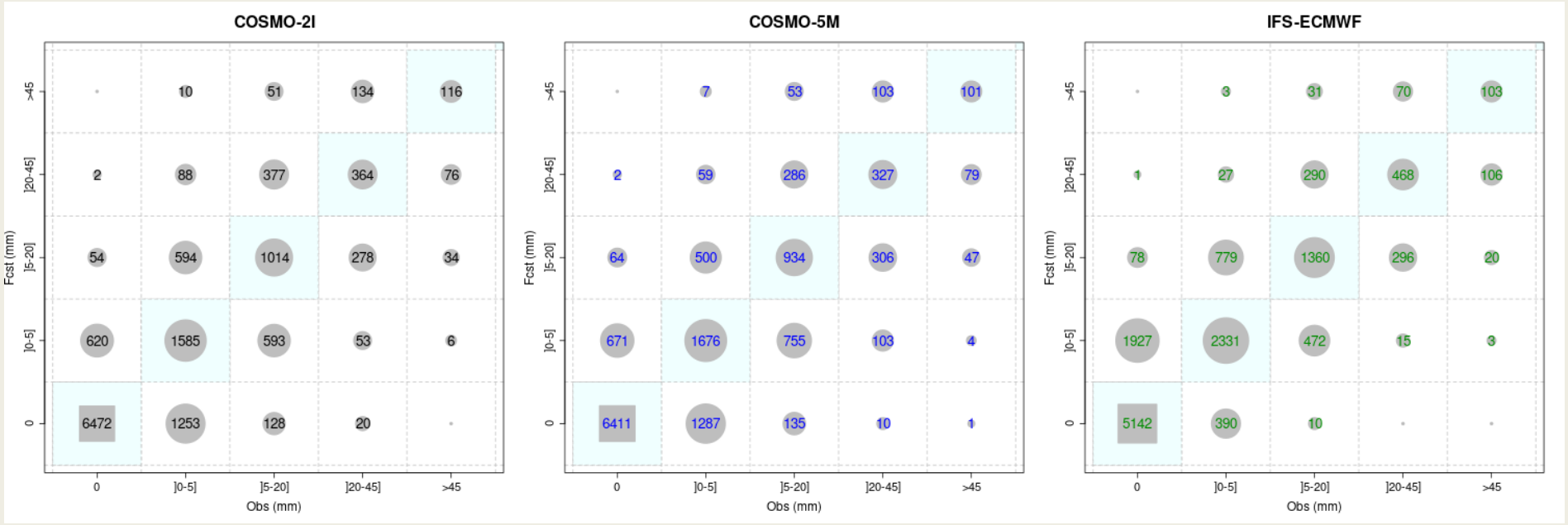
MAX



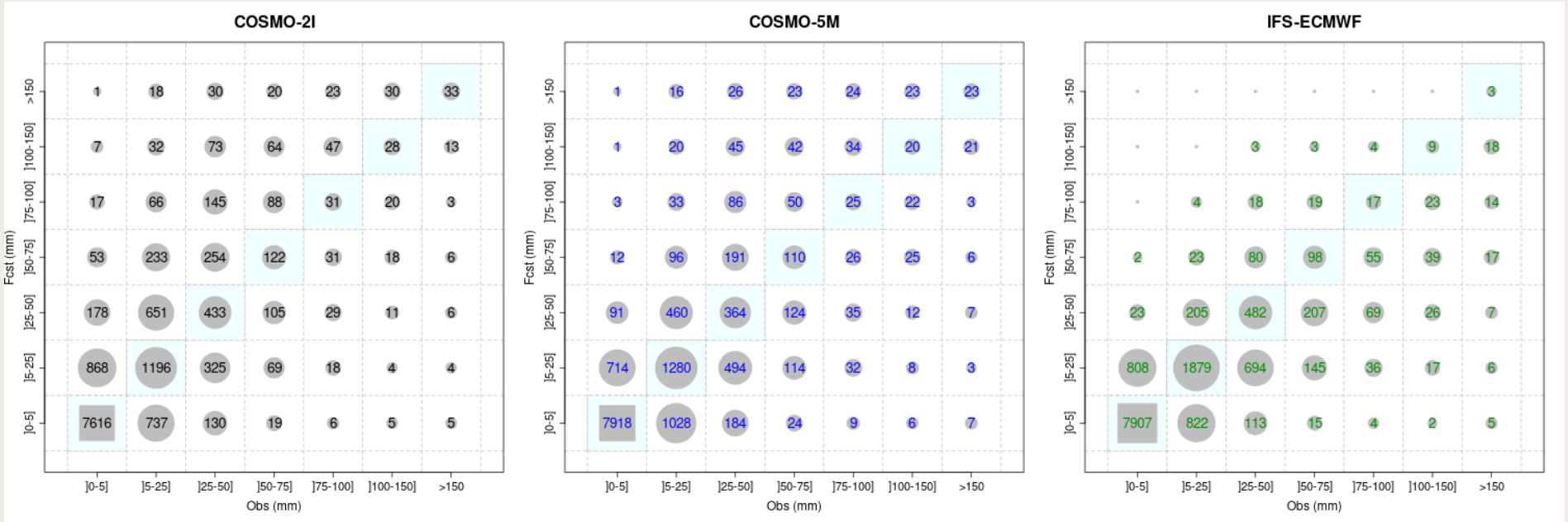


SON2019

MEAN

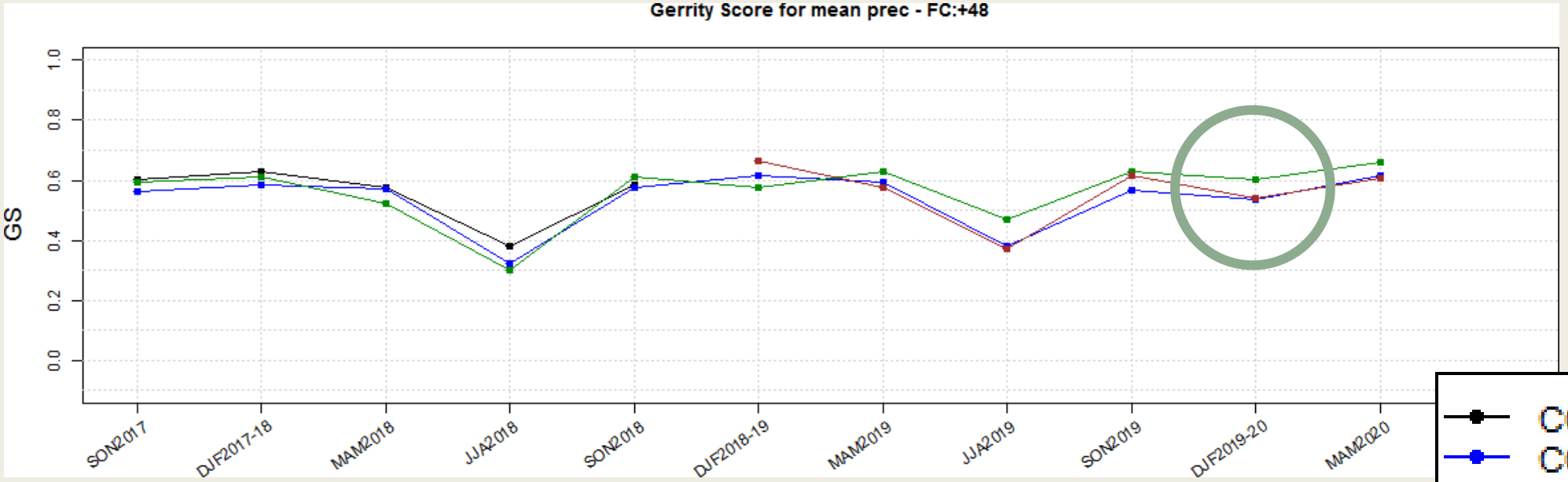


MAX

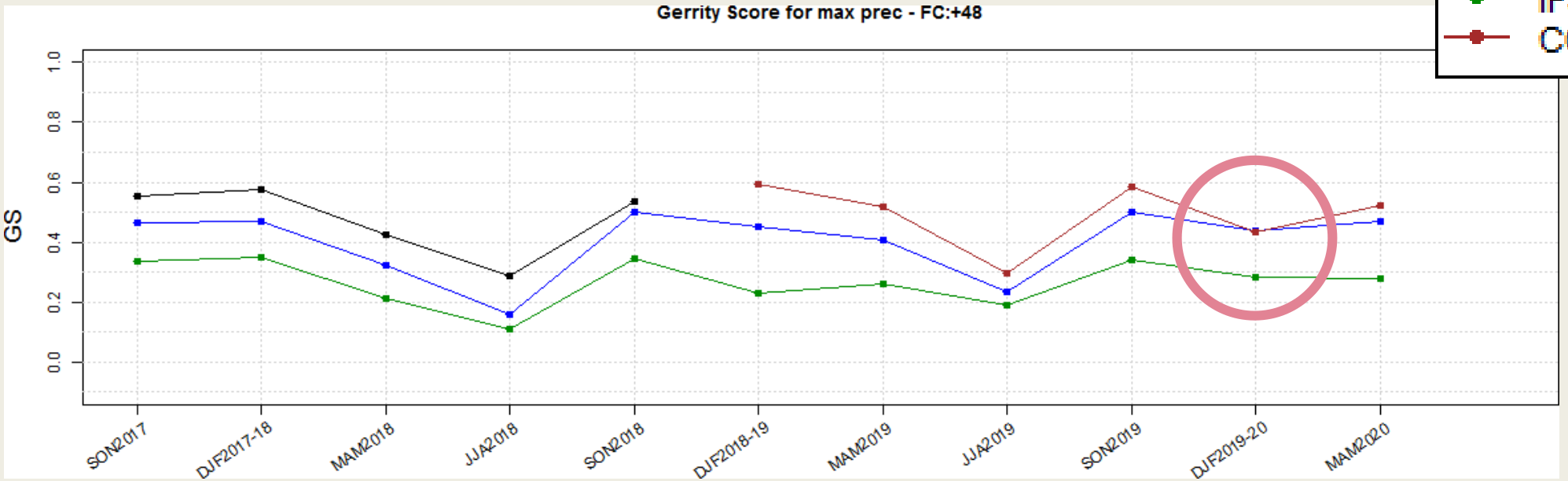


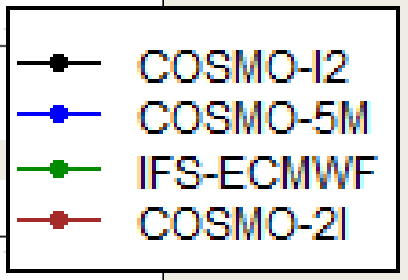
QPF: GS trend

MEAN



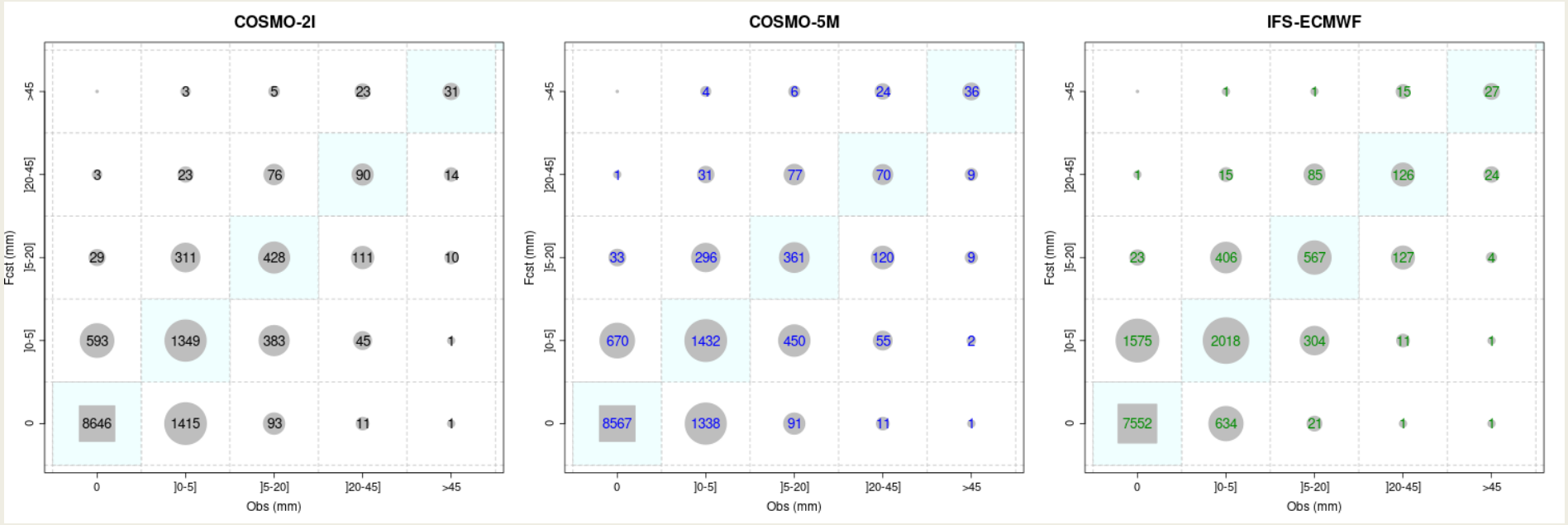
MAX



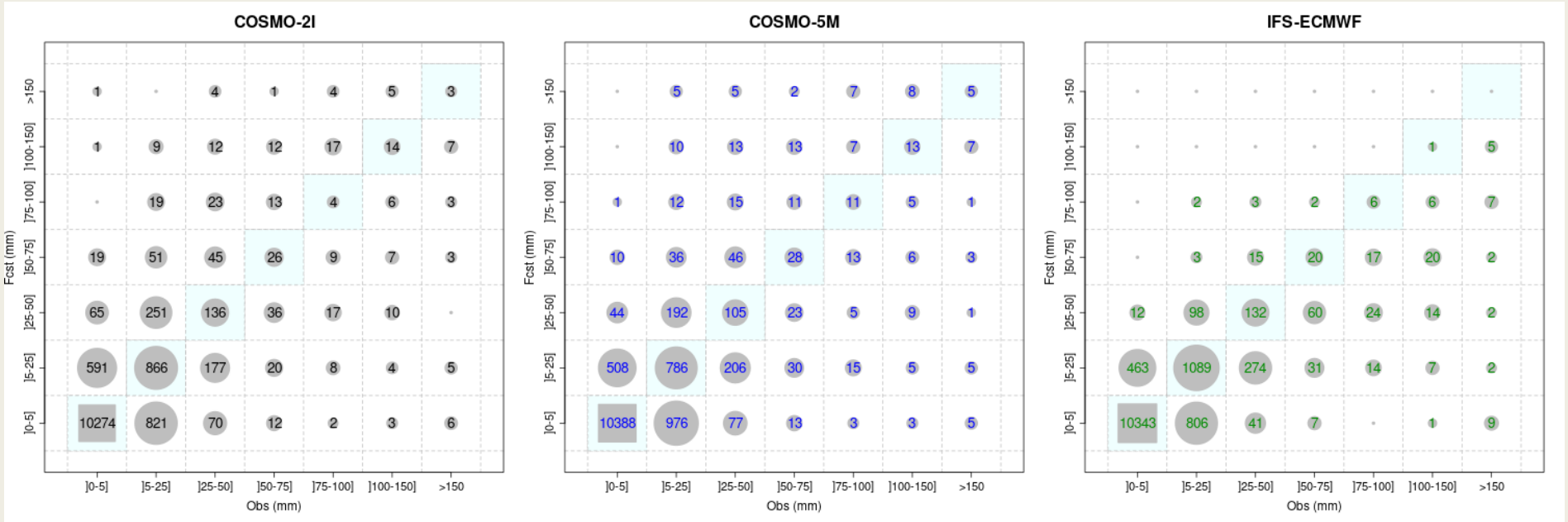


DJF2019-20

MEAN

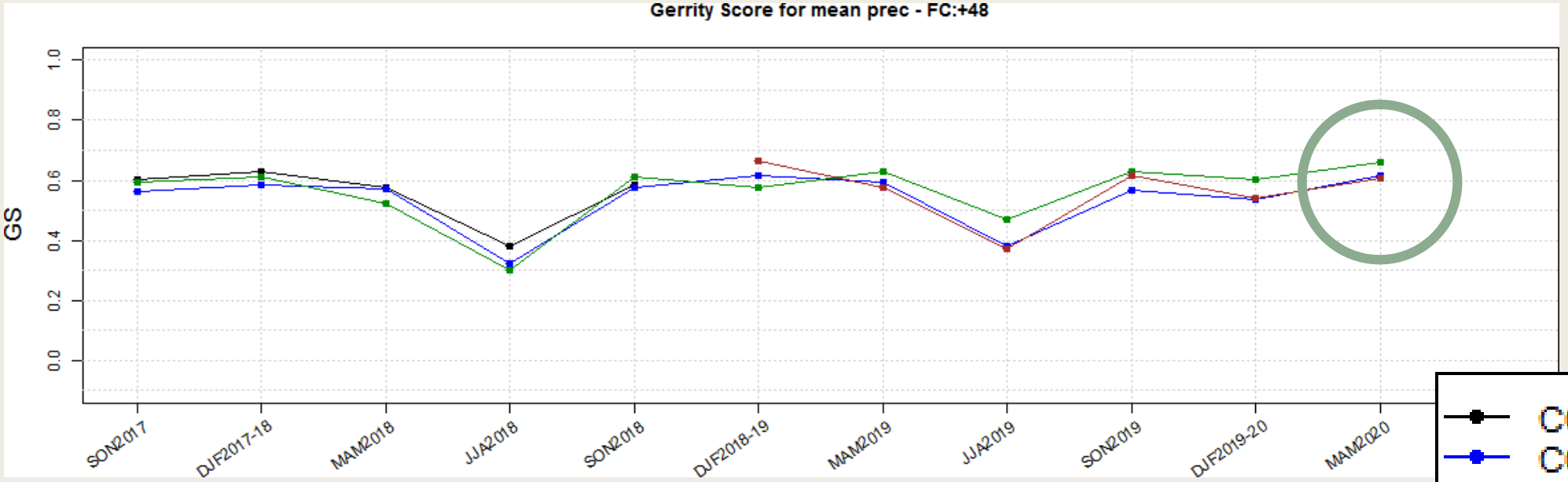


MAX

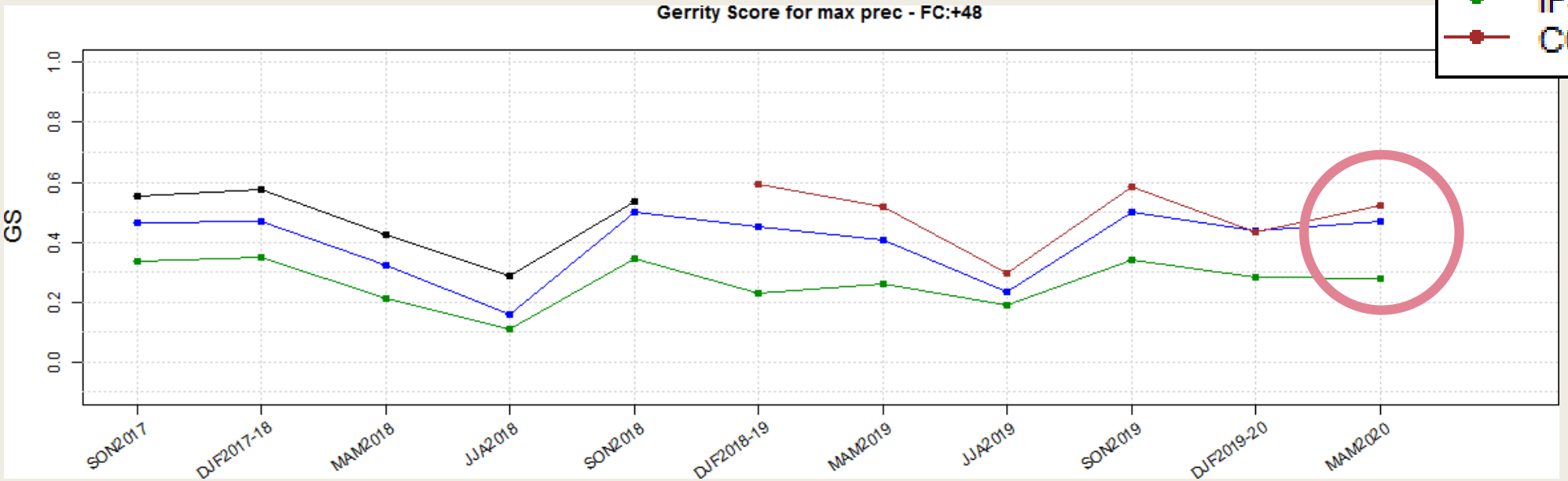


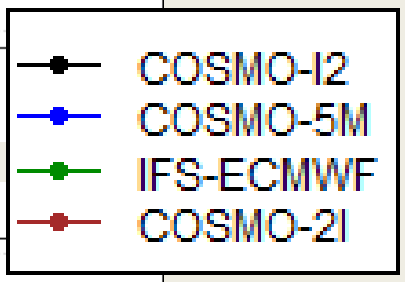
QPF: GS trend

MEAN



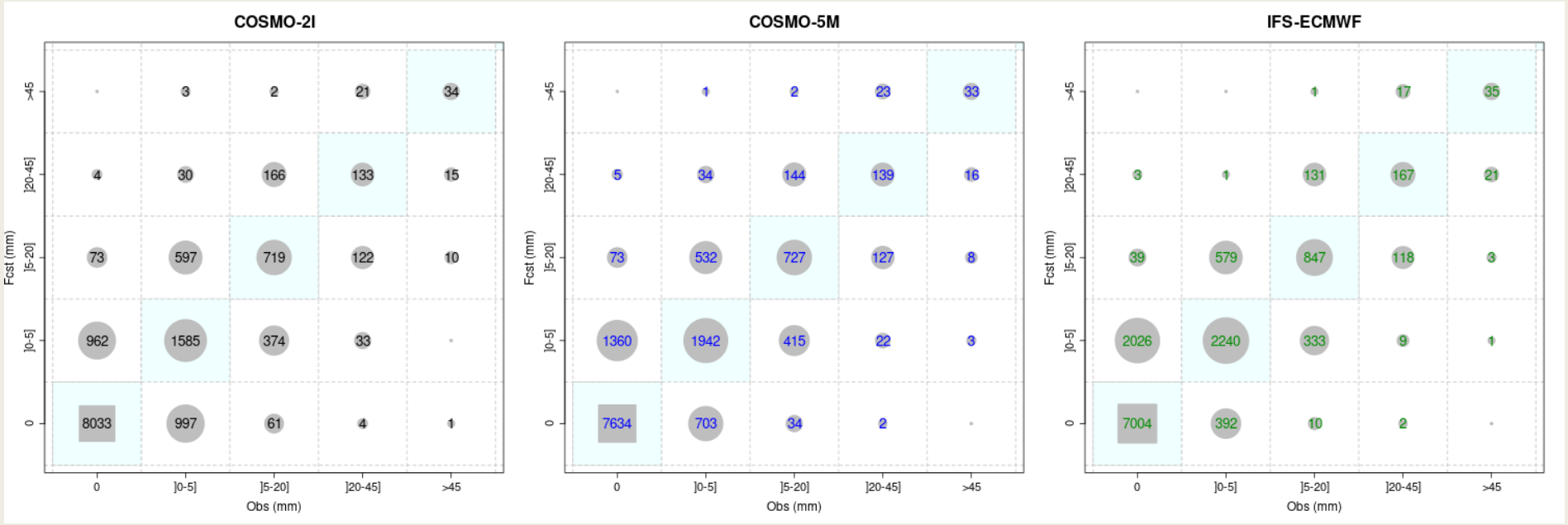
MAX



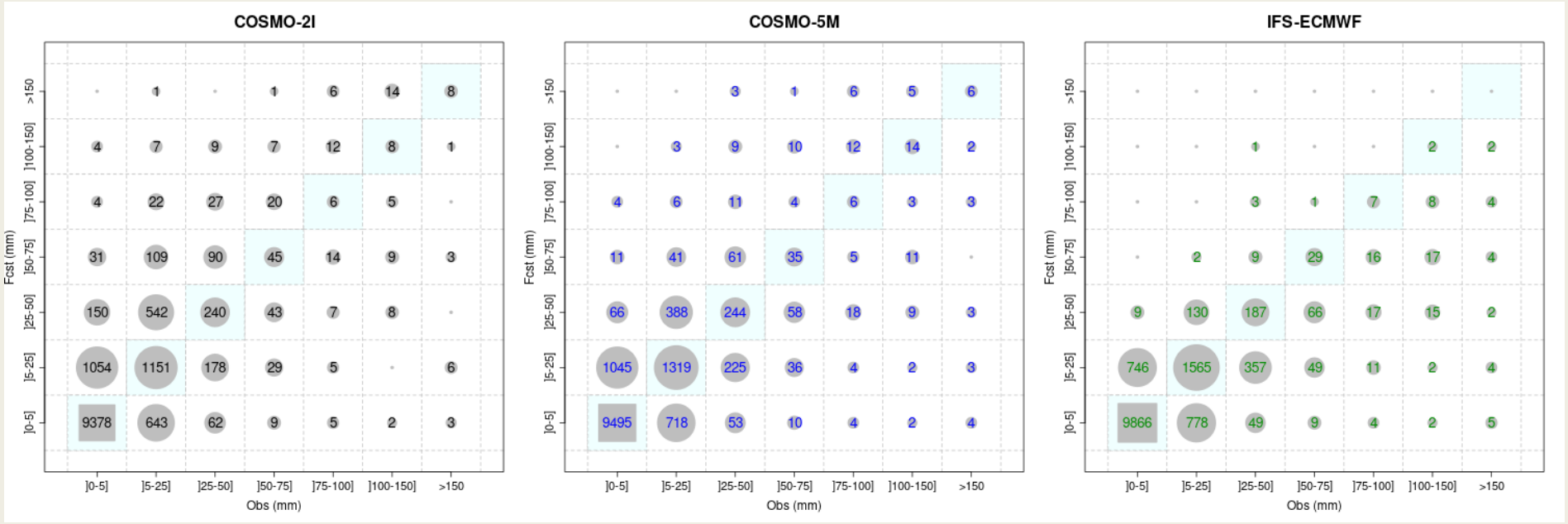


MAM2020

MEAN



MAX



CONCLUSION(1)

Trend over time for RMSE of T2m shows small improvement (also comparing corresponding seasons)

- The new model versions (Cosmo-5M that replaced Cosmo-I7 and Cosmo-2I that replaced Cosmo-I2) are slightly better than the previous ones
- Cosmo models perform a bit better than IFS-ECMWF
- Unfortunately errors are still large and during the operational forecast we have to face with overestimations of more than 3/4 degrees during clear night (in every seasons) or with too high maxima temperature in summer

Also trends over time of TS of QPF (mean/max for different threshold) do not show significant improvement (a little increase seems a bit more evident for maximum in the last year)

- IFS-ECMWF seems to perform better than Cosmo models if mean value of precipitation is considered and vice versa for maximum
- the TS does not give much credit to subjective impression that forecasters have using models operationally as it penalizes false and missed alarms in the same way

The use of Performance Diagram for different indicator (mean/max) and several thresholds helps to better highlight the behavior of models

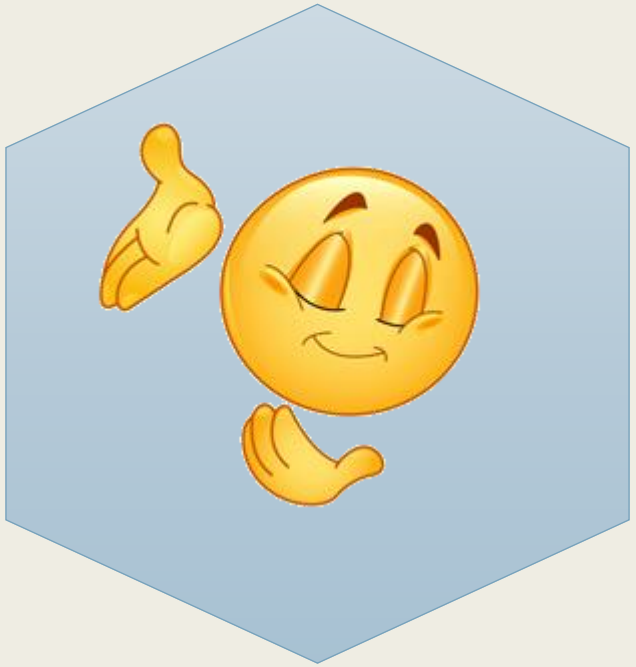
- For "mean" IFS-ECMWF tends to overestimate the number of events (with high POD) while Cosmo models have less overestimations (but with lower POD) and in some seasons they underestimate the events
- For the "max" COSMO-2I has better POD but with large overestimation of the events and higher number of false alarms. IFS-ECMWF (especially for higher thresholds) underestimate the number of events with many misses, but low false alarms.

CONCLUSION(2)

The “user oriented verification” based on the classification in classes of precipitation (always considering mean and max over catchment areas) allows to better represent the overall quality of models

- The use of “bubble plots” as a visual representations of a multi-category contingency table allows to understand the type of forecast errors
- The Gerrity score allow to quantify in a single indicator the accuracy of the forecast in predicting the correct category
- The trend over time of the GS shows that Cosmo models are better than IFS-ECMWF if the “max” is considered but vice versa for “mean”, even if the differences between models are smaller

Verification methodologies in Arpae have not shown significant innovations over the years, however we have tried to improve the way of presenting the results so that they could provide useful information especially to forecasts users



Thanks for your attention!