Norwegian
Meteorological
Institute

METreport

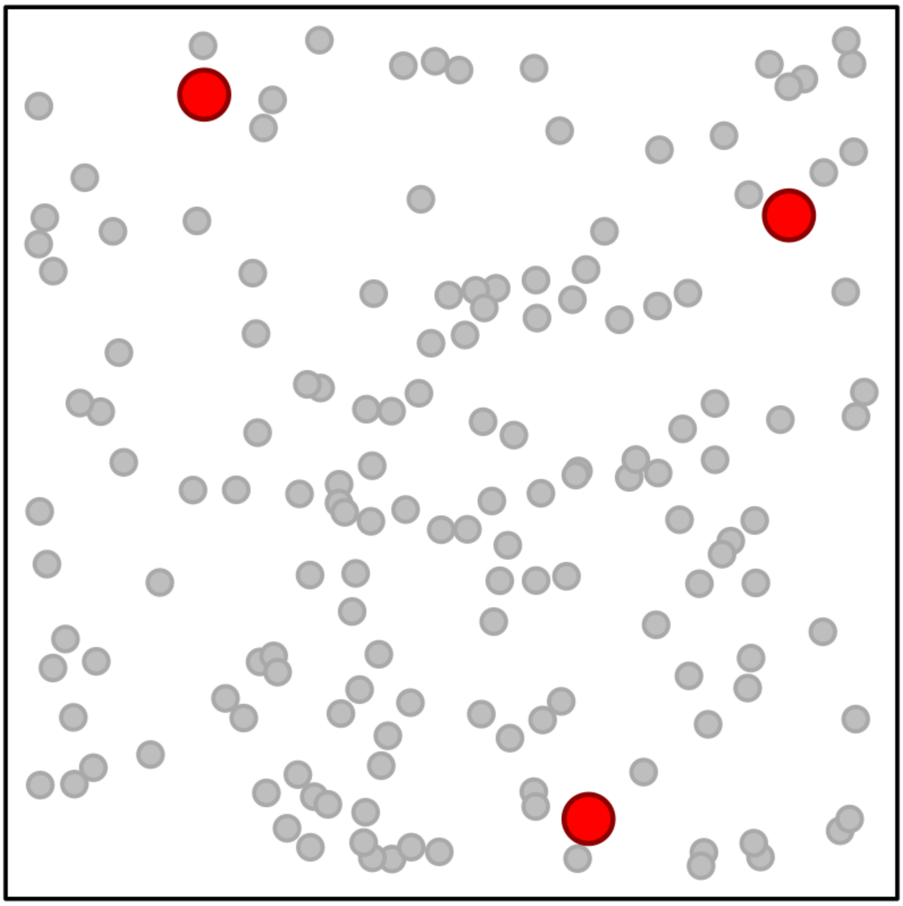
No. 13/2021
ISSN 2387-4201
Free

A spatial consistency test for the quality control of meteorological observations

Part III: Experiments on real-world data

Line Båserud and Cristian Lussana

The Norwegian Meteorological Institute, Oslo, Norway



Norwegian
Meteorological
Institute

METreport

Title A spatial consistency test for the quality control of meteorological observations. Part III: Experiments on real-world data	**Date** November 22, 2021
Section Division for Climate Services	**Report no.** 13/2021
Author(s) Line Båserud and Cristian Lussana	**Classification** ● Free ○ Restricted

Abstract

This report is the third of three reports describing the work that has been done at the Norwegian Meteorological institute (MET Norway) to develop and test spatial quality control methods for temperature and precipitation. The first report describes the methodology applied for spatial quality control. The second report describes a set of idealized experiments that have been made on temperature and precipitation data. This report describes the application of spatial quality control procedures to data collected by the observational network available at MET Norway.

Keywords

data quality control, in-situ observations, spatial statistics, automatic procedures

Disciplinary signature Hans Olav Hygen	Responsible signature Cecilie Stenersen

Abstract

This report is the third of three reports describing the work that has been done at the Norwegian Meteorological institute (MET Norway) to develop and test spatial quality control methods for temperature and precipitation. The first report describes the methodology applied for spatial quality control. The second report describes a set of idealized experiments that have been made on temperature and precipitation data. This report describes the application of spatial quality control procedures to data collected by the observational network available at MET Norway.

Contents

1 Introduction

This document describes the work that has been done at the Norwegian Meteorological institute (MET Norway) to develop and test spatial quality control methods for temperature and precipitation. This report describes the application of spatial quality control procedures to data collected by the observational network available at MET Norway.

The experiments evaluate the behaviour of the Spatial Consistency Test (SCT) on real-world data. The version of the SCT used in this report is the sct_resistant from titanlib (https://github.com/metno/titanlib/wiki/Spatial-consistency-test-resistant) described in detail in *Lussana and Båserud* (2021). The strategy builds around comparing the results of the SCT to the results from the official quality control at MET Norway for a subset of stations with assumed high quality and "close supervision". Synthetic errors are randomly introduced to the measurements in a set of experiments to investigate the typical error size that the SCT will detect. The SCT is run using all available in situ observations (stations owned/maintained by MET Norway, partners, and private citizens), however the statistical evaluation is done only for the high quality stations.

2 Temperature experiments on real-world data

2.1 Statistical comparison against a reference truth

Original and corrected observational values have been collected from the database of MET Norway for the Regional Basic Climatological Network (RBCN) stations listed in Table 1 for temperature and precipitation with a 1 h temporal resolution for 2016–2020, to serve as a test bed for the spatial quality control experiments with the SCT on real-world data. Based on the increase in number of citizen observations over the last years the period of 2020 was chosen for the analysis.

The RBCN data contain stationid (id number of the station), obstime (time of observation in UTC), paramid (code for weather element), typeid (code for message format when sending the data), original values (raw values), corrected values (corrected values by the official quality control at MET Norway), and a 16-digit controlinfo (coded information on which quality control routines have been performed on the data and their corresponding results). The RBCN data had initially one file per station (including several weather elements and message formats). These files have been transformed into having instead all stations in one file per timestep, as the spatial quality control is run on all available observations, one time step at a time.

Some of the RBCN stations changed message format during 2020, which resulted in periods with duplicate time steps included in the files during overlaps between the old and new message formats. The time of the switch was chosen from manual inspection of logs of message format priorities within the database for station information, and the duplicate time steps were removed.

For each run of the spatial quality control, the RBCN stations were blended with observations from the Yr-production chain. These include other MET stations and also citizen observations from Netatmo (https://www.netatmo.com/no-no/weather/weatherstation). The blending was done by removing the stations from the Yr-production files that match the exact latitude and longitude in the RBCN files.

The stations Svalbard lufthavn and Vardø Radio have not been used for the evaluation as they are too remote (i.e. do not have enough neighbours) for spatial QC (marked gray in Tab 1).

stationid	lat	lon	height	name	element
4780	60.2065	11.0802	202	GARDERMOEN	TA, RR
16560	62.07165	9.1147	638	DOMBÅS - NORDIGARD	TA, RR
16610	62.1133	9.2862	973	FOKSTUGU	TA
18700	59.9423	10.72	94	OSLO - BLINDERN	TA, RR
24890	60.567	9.1323	166	NESBYEN - TODOKK	TA, RR
36200	58.3988	8.7893	12	TORUNGEN FYR	TA, RR
44560	58.8843	5.637	7	SOLA	TA, RR
47300	59.3065	4.8723	55	UTSIRA FYR	TA, RR
50540	60.383	5.3327	12	BERGEN - FLORIDA	TA, RR
62480	62.8585	6.5378	20	ONA II	TA, RR
69100	63.4597	10.9305	12	VÆRNES	TA, RR
71550	63.7045	9.6105	10	ØRLAND III	TA
80740	66.9035	13.646	9	REIPÅ	TA, RR
82290	67.267	14.3637	11	BODØ VI	TA
90450	69.6537	18.9368	100	TROMSØ	TA, RR
90490	69.6767	18.9133	8	TROMSØ - LANGNES	TA
97251	69.4635	25.5023	131	KARASJOK - MARKANNJARGA	TA, RR
98550	70.3707	31.0962	10	VARDØ RADIO	TA, RR
99840	78.2453	15.5015	28	SVALBARD LUFTHAVN	TA, RR

Table 1: Norwegian stations within the Regional Basic Climatological Network (RBCN) considered for this report. The stations marked in gray were found to be isolated for the current SCT settings, and have not been used for the evaluation. The stations Utsira Fyr, Ona II, and Karasjok - Markannjarga were isolated only for precipitation.

For the RBCN stations (Tab. 1, and red stations in Fig. 1) the SCT is run on the original observations. The observations from the other MET stations use the values retrieved during the Yr-production and these are expected to have gone through at least the first part of the official quality controls. The citizen observations, that are the main bulk of the background observations, are raw data. The evaluation of the results is done only at the RBCN stations by comparing the good/bad flag given by the SCT against the reference good/bad flag given by the official QC. The reference flag is deemed as bad if the original observation differs from the corrected value after the official QC. The reference flag is good if no change has been made. If the original value is missing then that observation is omitted for the current time step.

The SCT threshold is the deciding factor when flagging an observation as bad. The SOD score (the deviation of the SCT score from its areal average, normalized by the dispersion) of each observation is compared to this threshold (*Lussana and Båserud*, 2021). For the statistical runs on temperature observations the SCT was set up with thresholds that do not distinguish between positive and negative cv-increments, we have instead used a unique value for T (i.e. $T = T_{+} = T_{-}$). For these tests the threshold T was varied between 2, 3, 4, and 6. The SCT was run over the domain seen in Fig. 1 using the settings seen in Tab. 2.

parameter	value
ε^2	0.5
r_{in}	20 km
r_{out}	40 km
k	3
$p_{\text{out,mn}}$	5
$p_{\text{out,mx}}$	50
D_{z}	400 m
D_{mn}	1 km
D_{mx}	20 km
$\boldsymbol{v}^{\text{mn}}$	$\boldsymbol{y}^{\text{o}} - 1°\text{C}$
$\boldsymbol{v}^{\text{mx}}$	$\boldsymbol{y}^{\text{o}} + 1°\text{C}$
$\boldsymbol{a}^{\text{mn}}$	$\boldsymbol{y}^{\text{o}} - 20°\text{C}$
$\boldsymbol{a}^{\text{mx}}$	$\boldsymbol{y}^{\text{o}} + 20°\text{C}$
$\boldsymbol{y}^{\text{b}}$	Theil-Sen linear regression

Table 2: Parameters used for temperature. For a description of the mathematical notation see Tab. 1 in *Lussana and Båserud* (2021).

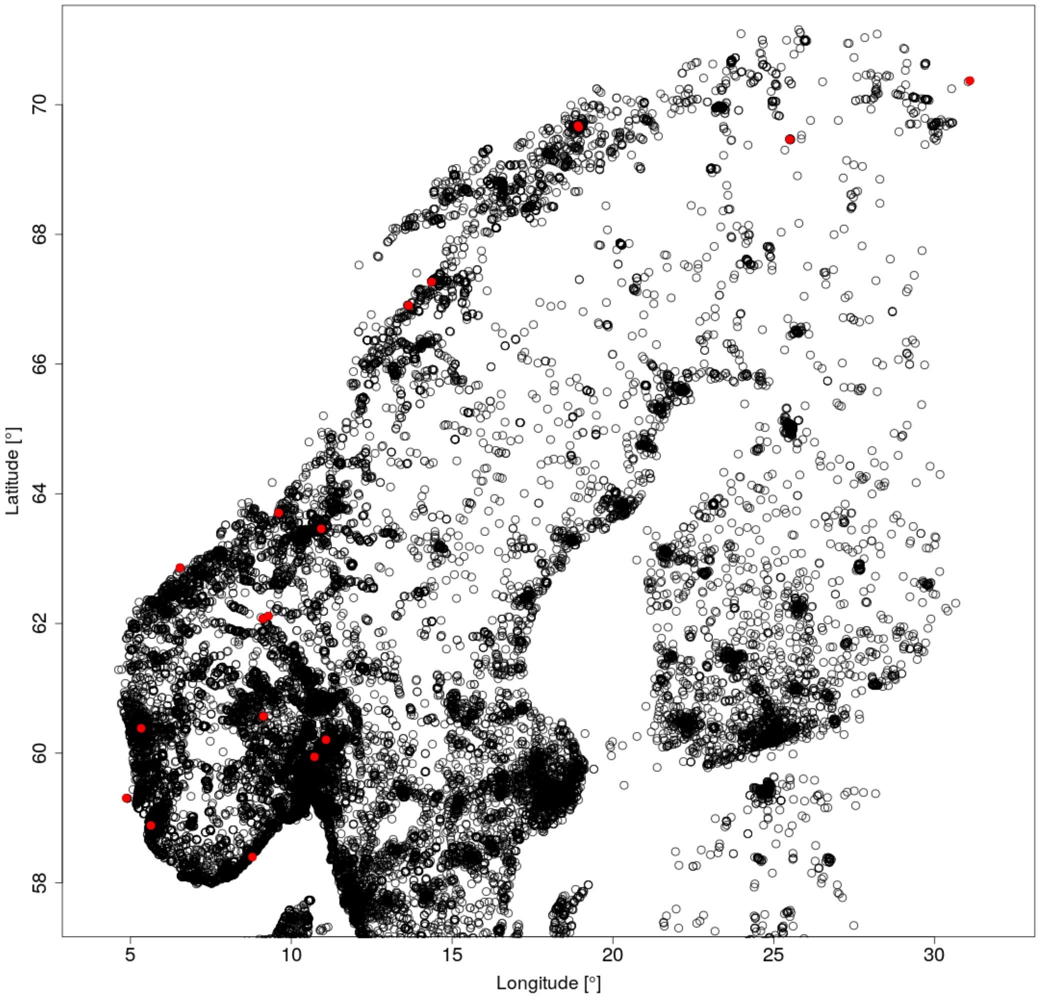


Figure 1: Location of stations used to run the SCT. Regional Basic Climatological Network (RBCN) stations in red.

Figure 2 shows the number of temperature observations from 2020 that the SCT flags as hits (upper left panel), false alarms (upper right panel), misses (lower left panel), and correct negatives (lower right panel), in reference to the official QC for the different SCT thresholds 2, 3, 4, and 6. The official QC flags 75 observations from the RBCN stations as bad for the year 2020. The SCT catches about 50 of these, and misses about 25. A few more of the bad observations are caught for the lowest threshold compared to the others. However, for the lowest threshold the amount of false alarms increases and the amount of correct negatives decreases significantly. When the SCT threshold increases it gets harder for the routine to flag observations as bad, and hence the amount of false alarms decreases, and the amount of correct negatives increases. The rate of change decreases as the thresholds increase.

Figure 3 shows the equitable threat score (ETS, see description in Appendix A) for the SCT for the different thresholds. As the number of hits and misses are relatively constant between the thresholds, while the amount of false alarms and correct negatives changes a lot more, the ETS increases for the higher thresholds.

The evaluation was also done on a monthly basis. The results can be seen in Figs. 13–24 in the supplementary material in Appendix B. The hits are detected in January, August, September, November, and December. The misses are detected in January, July, September, October, November, and December. There are no bad observations within the RBCN stations in February, March, April, May, and June.

To investigate further why the SCT does not catch 25 of the 75 bad observations for 2020, the bad observations are investigated in more detail. The station at Sola has several unphysical values (-99.9) for some of the time steps over three consecutive days in November, while the station at Ørland has the same (-99.8) for two consecutive days in January. The values from Sola have been identified as special values, meaning that the original value is missing, by a range check within the official QC. The corrected values were interpolated using model values within a prognostic numerical spatial control. The original values from Ørland were identified and interpolated in the same way, but in addition a later timeseries check deemed the corrections as uncertain. The station at Tromsø-Langnes has during six different occasions observations corrected from the value 0 (found in August, September, October and December). One is adjusted just 0.7 °C while the rest are adjusted between 3.3 to 12.9 °C. The bad observations were detected by either a formal consistency control or a step/dip/freeze control. The corrections were done manually by human quality control inspection. Nesbyen has one short period (4 consecutive time steps) over the transition from November to December with values slightly corrected by the human quality control, from the value of 0.6 to 0.7 °C. Værnes has two bad observations (found in August and December). These are corrected by the human quality control from the value 1 to 16.7 °C and from 1.7 to 1 °C, respectively. The station Bergen-Florida has one short period in January and one in September with minor corrections. The temperature values were marked as bad by the step/dip/freeze test (frozen values), and were adjusted between 0.1-0.3 °C by the human quality control. Utsira has 4 bad observations in July. These were also marked as bad by the step/dip/freeze test (frozen values), and adjusted by 0.1-0.3 °C by the human quality control.

Examples of small and large corrections to the temperature values can be seen in Fig 4.

The left panel shows one of the cases from Bergen-Florida (2300 UTC 28 September 2020 to 0400 UTC 29 September 2020), while the right panel shows one of the cases from Tromsø-Langnes (1900 UTC 18 December 2020). The SCT does not flag the small errors from Bergen. For the example with the larger error the SCT flags the observation for thresholds 2, 3, and 4.

The chosen settings for the range of valid values, $\boldsymbol{v}^{\mathrm{mn}} = \boldsymbol{y}^{\mathrm{o}} - 1^{\circ}\mathrm{C}$ and $\boldsymbol{v}^{\mathrm{mx}} = \boldsymbol{y}^{\mathrm{o}} + 1^{\circ}\mathrm{C}$, made it harder to flag the smallest errors. The range of valid values specify the tolerable size of the deviations between good observations and the leave-one-out estimated values at the same locations by means of neighbouring observations. If all the leave-one-out estimated values at neighbouring locations are close to the corresponding observations, that is within the range of valid values, then all the observations are set to good. For synthetic errors of the size of 1 °C, we should expect most of the perturbed observations to be flagged as good, because of the chosen settings for $\boldsymbol{v}^{\mathrm{mn}}$ and $\boldsymbol{v}^{\mathrm{mx}}$. However, when an observation has a significant representativeness error, for instance, just adding or removing 1 °C can be enough to change its judgement to bad.

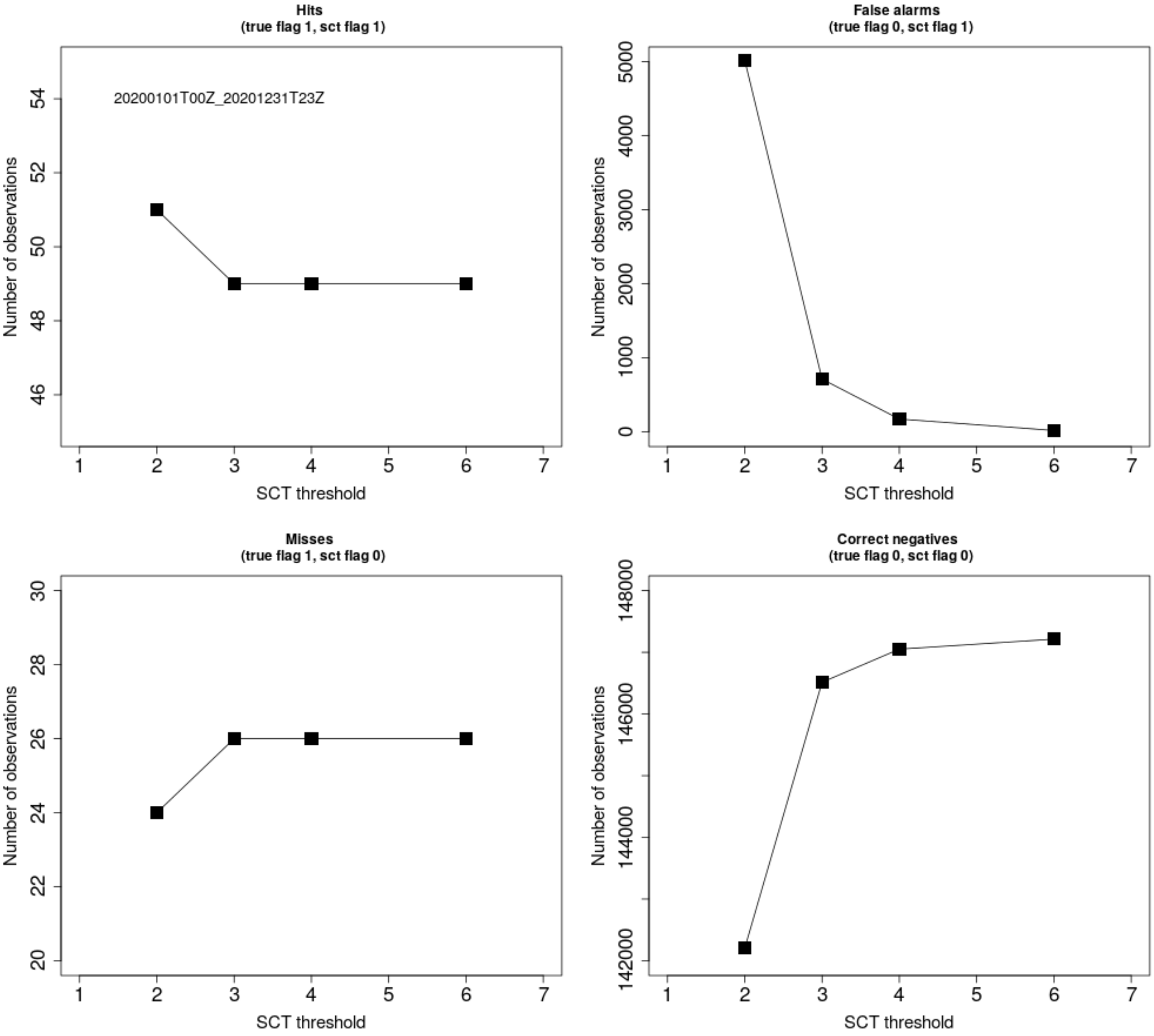


Figure 2: Number of temperature observations that are found by the SCT to be hits (upper left panel), false alarms (upper right panel), misses (lower left panel), and correct negatives (lower right panel), for the different SCT thresholds 2, 3, 4, and 6.

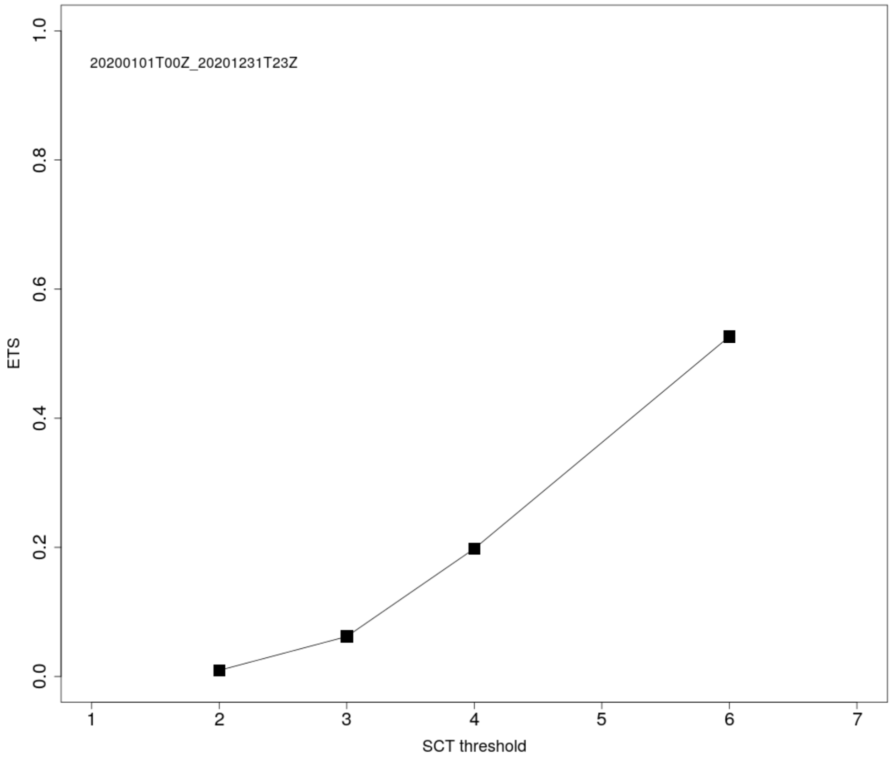


Figure 3: Equitable threat score (ETS) for temperature for the SCT against the different SCT thresholds 2, 3, 4, and 6.

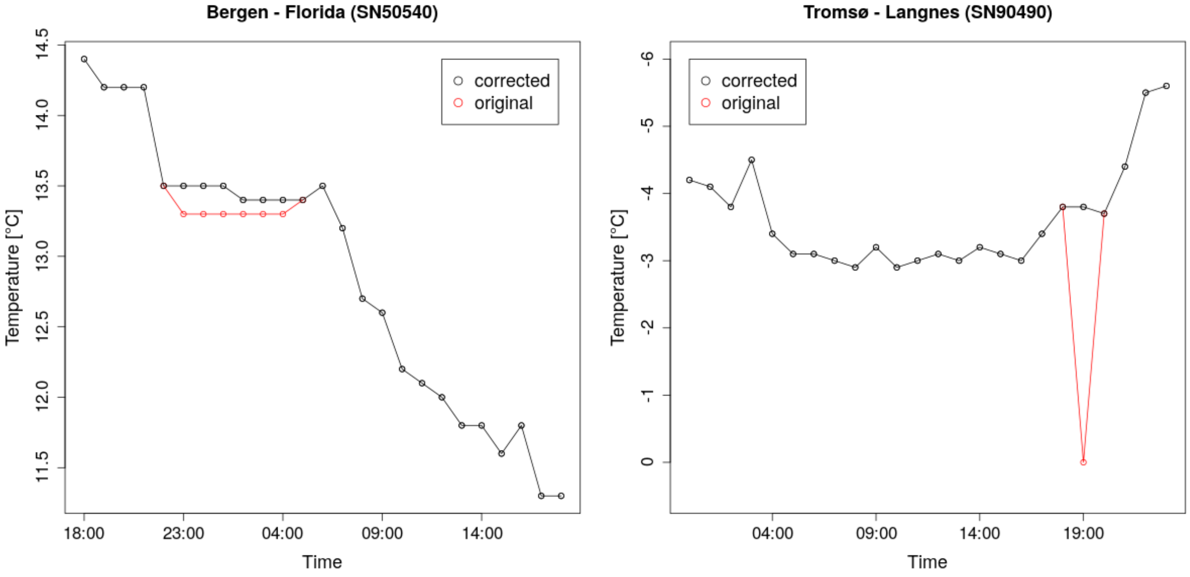


Figure 4: Examples of small (left panel, Bergen-Florida 2300 UTC 28 September 2020 to 0400 UTC 29 September 2020) and large (right panel, Tromsø-Langnes 1900 UTC 18 December 2020) corrections to the temperature values from the RBCN stations.

2.2 Synthetic error perturbations

In order to find the typical size of the errors that the SCT with these settings (Tab. 2) detects, the SCT was rerun for February while perturbing the values with synthetic errors. This month is without any bad observations in the RBCN data initially, which makes it suited for this test.

For each time step (hourly) 10% of the RBCN stations (typically 2 stations per time step) were chosen randomly and perturbed with positive and negative errors of 0.1, 0.5, 1, 2, 3, and 5 °C. This was repeated four times to get more robust statistics. The variation between the four simulations was below 3.3% for the negative synthetic errors, and below 3% for the positive.

Figure 5 shows the hits, false alarms, misses, and correct negatives for the range of synthetic error perturbations for the different SCT thresholds. In general we get more hits and fewer misses as the error increases. The smallest errors (0.1 and 0.5) get almost no hits. This is linked to the range of valid values discussed in the previous section. The smaller the threshold the more hits and also fewer misses. The difference between the thresholds is largest for the errors of medium size (e.g. 2 and 3 °C) and decreases again for 5. Errors around 5 °C are more clearly detected no matter how strict the threshold is.

The rate of change of the curves of hits and misses is largest between errors of size 1 and 2 °C for the lowest threshold, and changes towards being largest between 3 and 5 °C for the highest threshold. The false alarms and the correct negatives are independent of the error size for all thresholds. However, there are differences between the thresholds. The lowest thresholds have clearly more false alarms and less correct negatives. That means the SCT flags too many observations. There are slight differences between negative and positive synthetic errors (circles vs triangles in Fig. 5).

The combined results are summarized in Fig. 6. Here we have the probability of detection (POD), the probability of false detection (POFD), and the equitable threat score (ETS), for the different synthetic error perturbations and SCT thresholds. The POD increases as the errors increase. It increases fastest for the lowest thresholds. The POD for the lowest threshold is almost equal to 1 for errors of 5 °C. The POFD is below 1% for all thresholds, except for $T = 2$ where it is 3%. The ETS reaches about 0.7 for T2 and T4. The highest is T3 with 0.8. This is because T3 has the second most hits, but also a lower false alarm rate than T2. Again it is clear that the biggest difference between thresholds can be found for the medium error sizes.

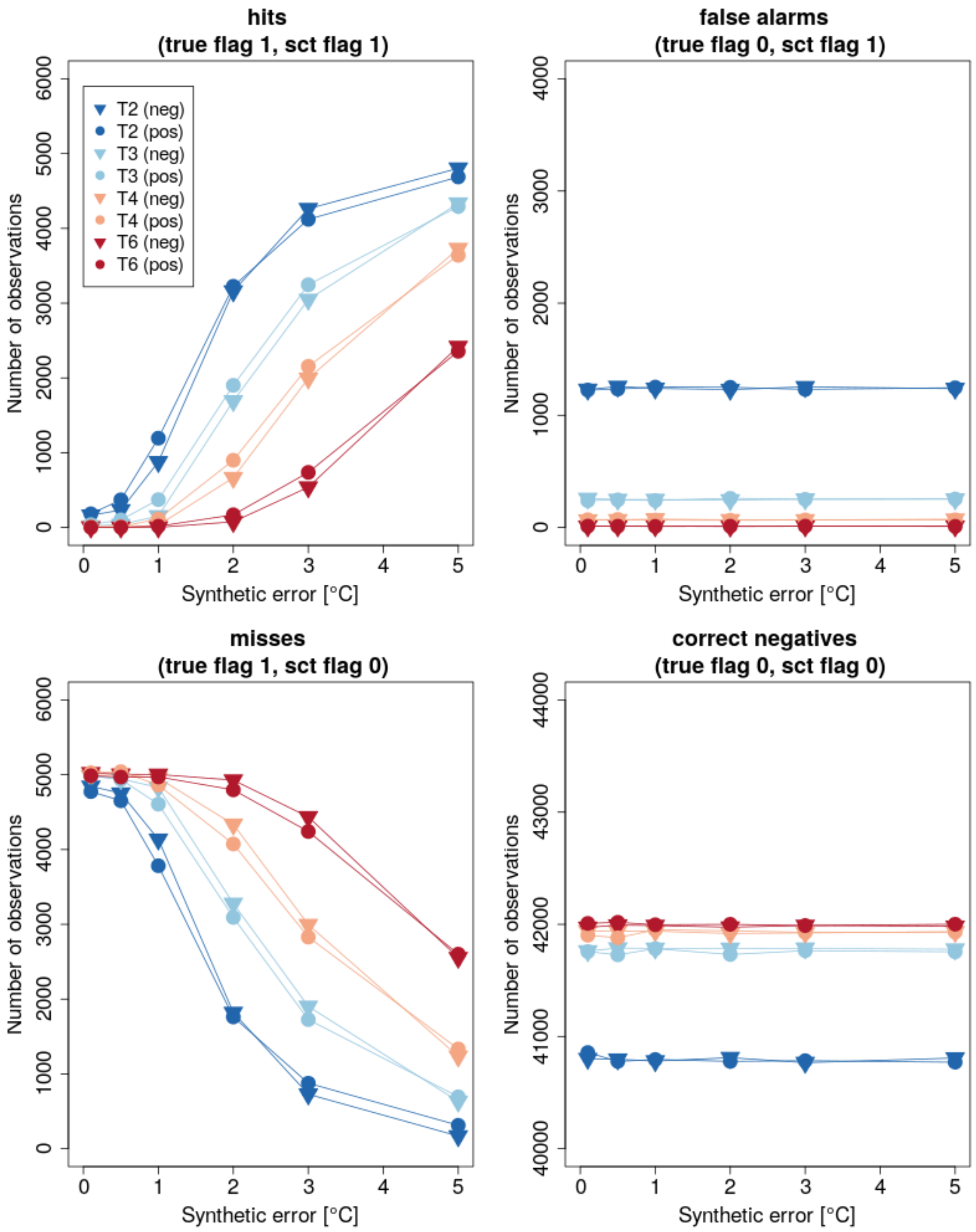


Figure 5: Hits (upper left panel), false alarms (upper right panel), misses (lower left panel), and correct negatives (lower right panel) for February 2020 for the positive (circles) and negative (triangles) synthetic errors 0.1, 0.5, 1, 2, 3, and 5. The different SCT thresholds can be seen in blue (T2), light blue (T3), peach (T4) and red (T6).

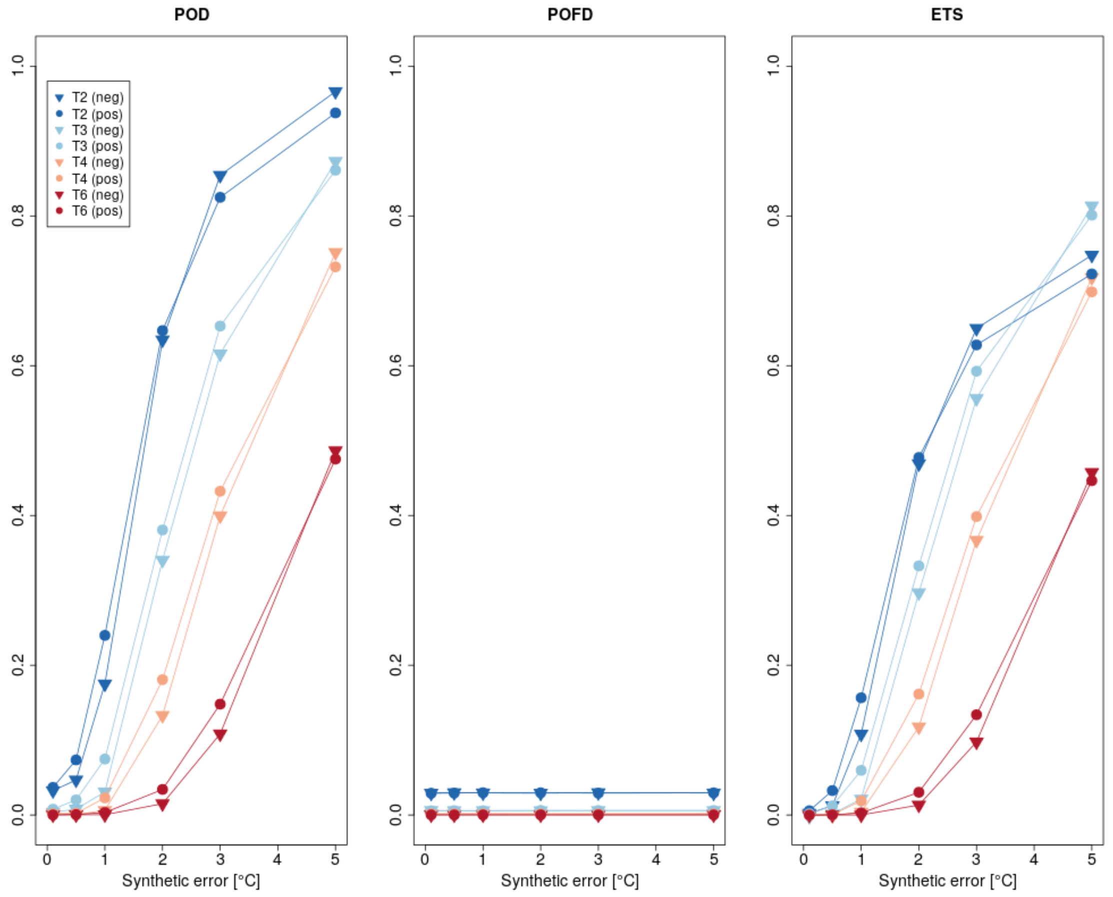


Figure 6: Probability of detection (POD, left panel), probability of false detection (POFD, middle panel), and equitable threat score (ETS, right panel) for February 2020 for the positive (circles) and negative (triangles) synthetic errors 0.1, 0.5, 1, 2, 3, and 5. The different SCT thresholds can be seen in blue (T2), light blue (T3), peach (T4) and red (T6).

There are some differences with the performance of the SCT if we separate into stations within the major cities or look at the most rural stations. The results can be seen in Figs. 7 and 9, and Figs. 8 and 10, respectively. The city stations chosen are Oslo-Blindern, Bergen-Florida, and Tromsø. They all have lots of neighbouring stations. The rural stations chosen are Fokstugu, Utsira Fyr, and Karasjok-Markannjarga. The latter has few stations within the inner circle of the SCT (about 10 stations) and the outer circle does not add many more. Fokstugu has some stations to the South-West, and the area is also characterized by mountainous terrain. Utsira Fyr has few stations in the inner circle, the outer circle however, is better covered as it extends over parts of the the city of Haugesund. See Tab. 2 for the sizes of the inner and outer circles used, and *Lussana and Båserud* (2021) for more in-depth information.

The number of hits and misses for the city stations follows a similar pattern as for all stations. The false alarm rate is highest for T2 and zero for the rest. The number of correct negatives varies around the same values for all thresholds and synthetic error amounts, but T2 is slightly lower. The results are similar for the rural stations, but the overall performance is poorer. Thresholds T3 and T4 have false alarms for the rural stations. Threshold T6 still has zero false alarms.

The ETS is again similar, and very low, for the smallest synthetic errors (0.1 and 0.5), for the rest it is higher (about 10% higher) for the city stations compared to all stations (Fig. 5), and in general, lower for the rural stations. There is a larger difference between the positive and negative errors for the city stations compared to all stations. The SCT performs better for the negative errors.

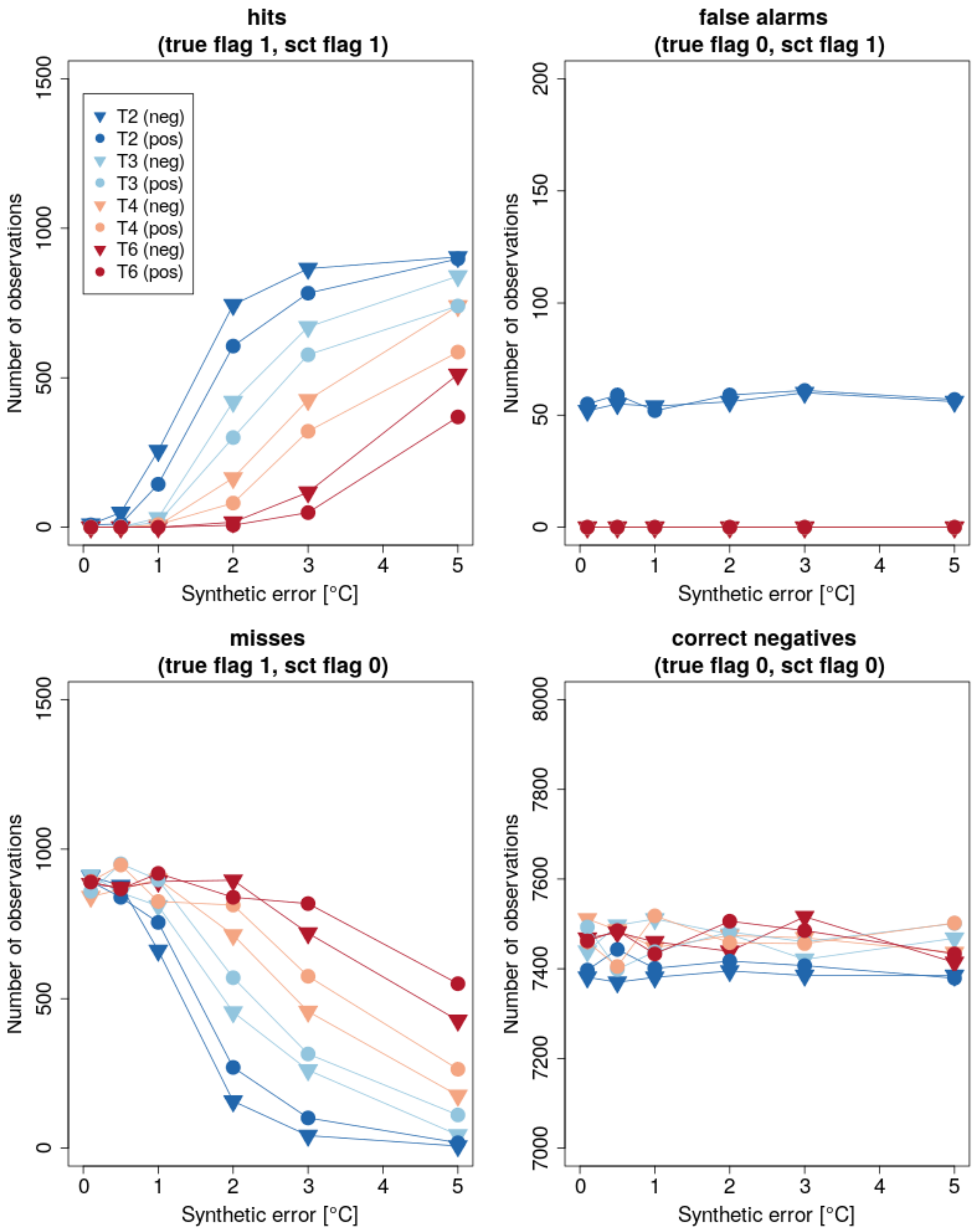


Figure 7: Hits (upper left panel), false alarms (upper right panel), misses (lower left panel), and correct negatives (lower right panel) for February 2020 for a smaller selection of city stations (Oslo-Blindern, Bergen-Florida, and Tromsø) for the positive (circles) and negative (triangles) synthetic errors 0.1, 0.5, 1, 2, 3, and 5. The different SCT thresholds can be seen in blue (T2), light blue (T3), peach (T4) and red (T6).

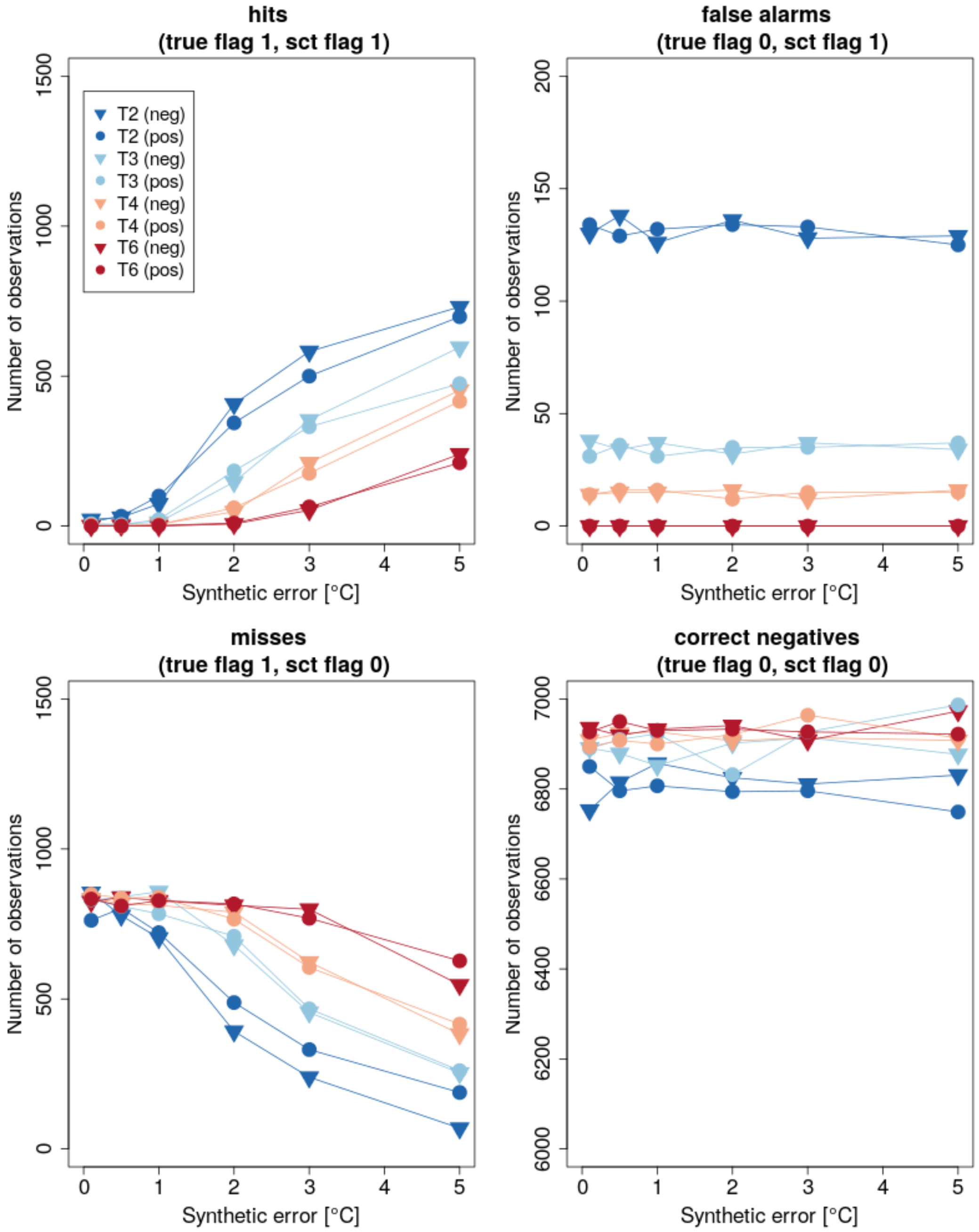


Figure 8: Hits (upper left panel), false alarms (upper right panel), misses (lower left panel), and correct negatives (lower right panel) for February 2020 for a smaller selection of rural stations (Fokstugu, Utsira Fyr, and Karasjok-Markannjarga) for the positive (circles) and negative (triangles) synthetic errors 0.1, 0.5, 1, 2, 3, and 5. The different SCT thresholds can be seen in blue (T2), light blue (T3), peach (T4) and red (T6).

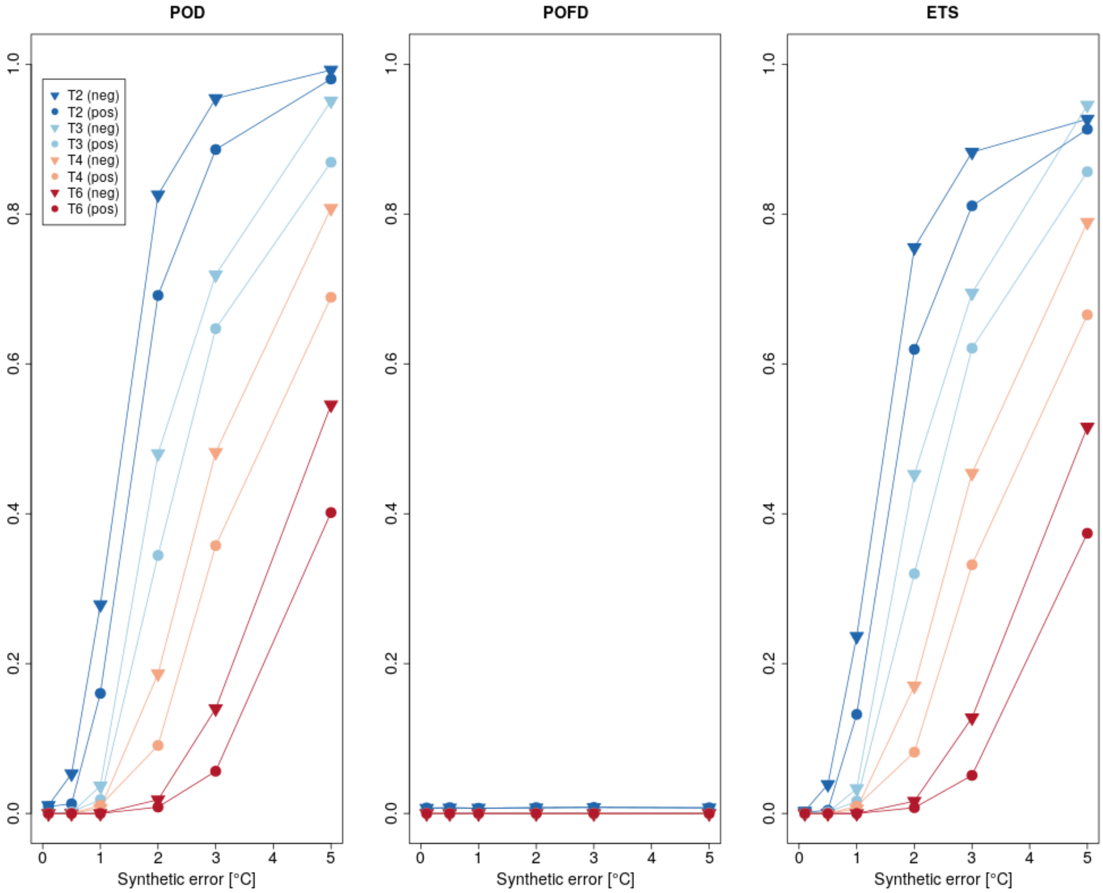


Figure 9: Probability of detection (POD, left panel), probability of false detection (POFD, middle panel), and equitable threat score (ETS, right panel) for February 2020 for a smaller selection of city stations (Oslo-Blindern, Bergen-Florida, and Tromsø) for the positive (circles) and negative (triangles) synthetic errors 0.1, 0.5, 1, 2, 3, and 5. The different SCT thresholds can be seen in blue (T2), light blue (T3), peach (T4) and red (T6).

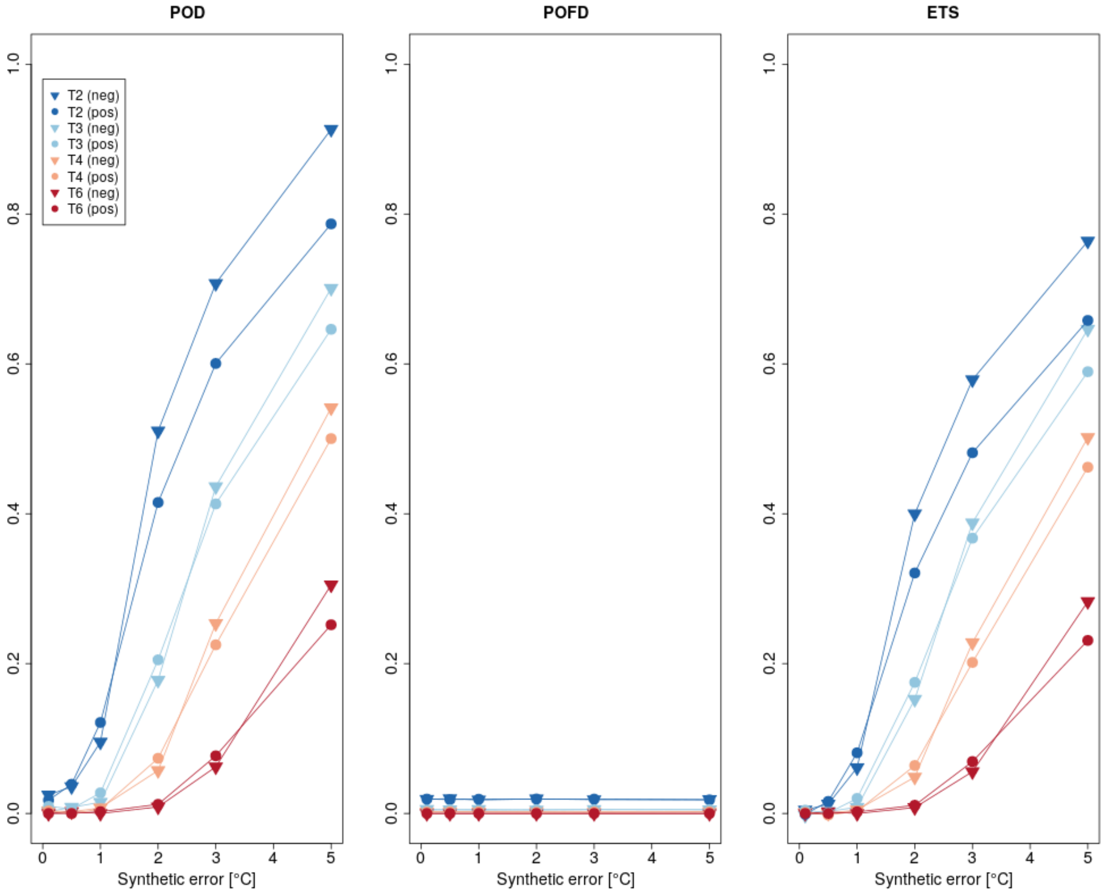


Figure 10: Probability of detection (POD, left panel), probability of false detection (POFD, middle panel), and equitable threat score (ETS, right panel) for February 2020 for a smaller selection of rural stations (Fokstugu, Utsira Fyr, and Karasjok-Markannjarga) for the positive (circles) and negative (triangles) synthetic errors 0.1, 0.5, 1, 2, 3, and 5. The different SCT thresholds can be seen in blue (T2), light blue (T3), peach (T4) and red (T6).

3 Precipitation experiments on real-world data

3.1 Statistical comparison against a reference truth

The precipitation data were prepared in the same way as for temperature (for details see Sect. 2). The SCT was again run over the domain in Fig. 1. Fifteen of the RBCN stations have precipitation observations. The settings used for precipitation can be found in Tab. 3. It is worth noting that a Box-Cox transformation with a transformation parameter equal to

0.5 was applied to the precipitation data before running the SCT. The valid and admissible ranges for precipitation are set up in such a way that we use either ranges of fixed values or a percentage of the precipitation amount, depending on the amount of precipitation to be checked. The SCT thresholds 2, 3, 4, and 6, were also used for precipitation.

parameter	value
ε^2	0.1
r_{in}	10 km
r_{out}	50 km
k	3
$p_{\text{out,mn}}$	5
$p_{\text{out,mx}}$	50
D_{z}	-
D_{mn}	500 m
D_{mx}	25 km
$\boldsymbol{v}^{\text{mn}}$	min($\boldsymbol{y}^{\text{o}} - 0.2\,\text{mm}$ OR $\boldsymbol{y}^{\text{o}} - 5\%$)
$\boldsymbol{v}^{\text{mx}}$	max($\boldsymbol{y}^{\text{o}} + 0.2\,\text{mm}$ OR $\boldsymbol{y}^{\text{o}} + 5\%$)
$\boldsymbol{a}^{\text{mn}}$	min($\boldsymbol{y}^{\text{o}} - 5\,\text{mm}$ OR $\boldsymbol{y}^{\text{o}} - 33\%$)
$\boldsymbol{a}^{\text{mx}}$	max($\boldsymbol{y}^{\text{o}} + 5\,\text{mm}$ OR $\boldsymbol{y}^{\text{o}} + 33\%$)
$\boldsymbol{y}^{\text{b}}$	Median outer circle
$\lambda_{\text{Box\_Cox}}$	0.5

Table 3: Parameters used for precipitation.

Figure 11 shows the number of precipitation observations from 2020 that the SCT flags as hits (upper left panel), false alarms (upper right panel), misses (lower left panel), and correct negatives (lower right panel), in reference to the official QC for the different SCT thresholds 2, 3, 4, and 6.

As expected the SCT flags more observations, and misses fewer observations, for the lowest thresholds. The lowest thresholds give also more false alarms and less correct negatives. The rate of change between the thresholds is largest going from threshold 2 to threshold 3, for all panels. The larger amount of misses dominate over the hits, leading to an ETS very close to zero (Fig. 12).

As for temperature, the evaluation was also done on a monthly basis. The results can be seen in Figs. 25– 36 in the supplementary material in Appendix B.

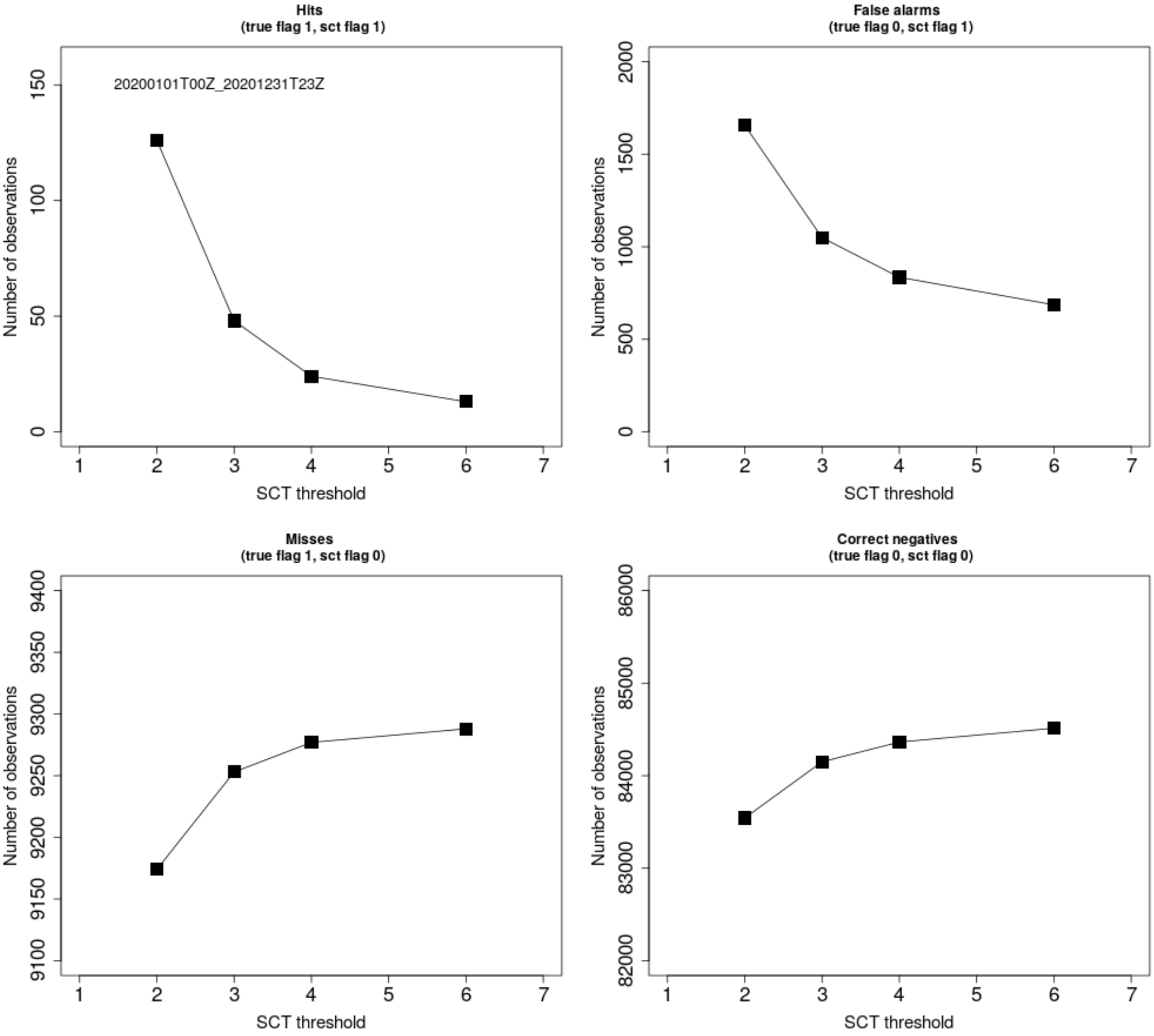


Figure 11: Number of precipitation observations that are found by the SCT to be hits (upper left panel), false alarms (upper right panel), misses (lower left panel), and correct negatives (lower right panel), for the different SCT thresholds 2, 3, 4, and 6.

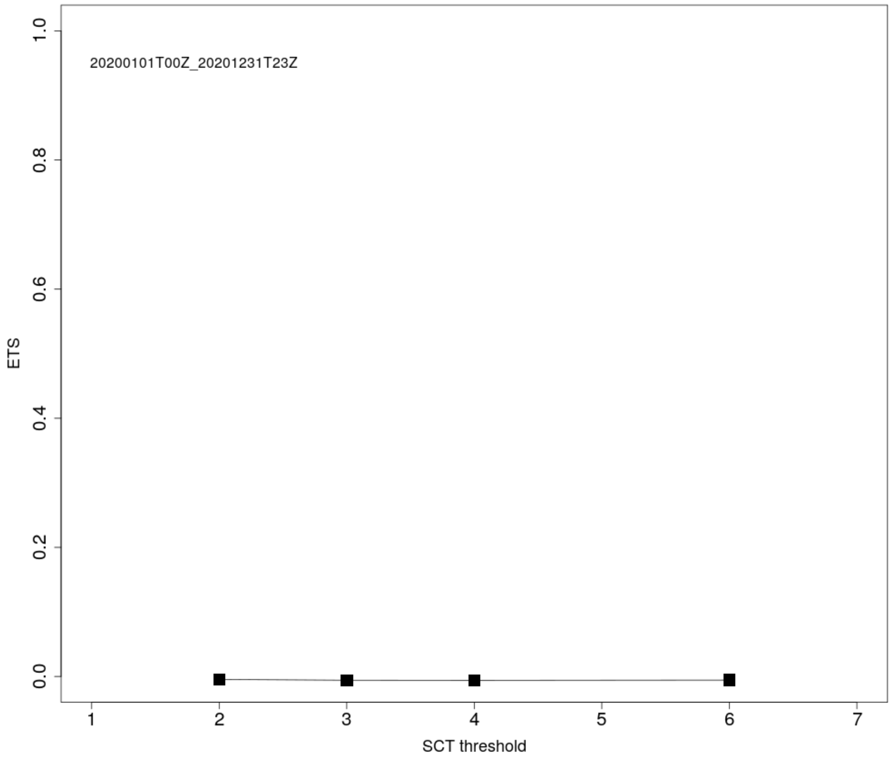


Figure 12: Equitable threat score (ETS) for precipitation for the SCT against the different SCT thresholds 2, 3, 4, and 6.

4 Conclusions and future work

The results from using the SCT for the application of quality control of temperature for high quality stations using stations with lesser quality, show promising results with the current settings. The number of hits is higher and the number of misses is lower for the lowest threshold, while both are fairly constant for the rest. The amount for false alarms and correct negative vary more between the thresholds, but the lowest threshold stands out. This leads to an increase in the ETS as the thresholds increase.

The RBCN data for temperature contain few bad observations in the reference data, giving us not too many possibilities for hits with the SCT. A large portion of the bad observations are only small corrections, which are more challenging to catch with the SCT. The chosen range of valid values also plays a role here, and values lower than 1 °C will in the future be tested.

For the introduction of random synthetic errors the lowest threshold gives a higher

POD, but also the highest POFD. Since the POFD is lower than for all other thresholds, the resulting ETS is highest for threshold T3 for the largest errors. For synthetic errors below 5 °C, the lowest threshold gives the highest ETS values.

In the future we also want to run the SCT with random synthetic errors for June 2020. Like February, this is also a month with no errors in the reference data, making it suitable for such an analysis. Preliminary testing for June has showed that the difference between the positive and negative errors is larger than was found for February. The hypothesis being that a higher variability is present for the citizen stations during summer due to differences in the amount of sun exposure.

We have also started an analysis on temperature observations from stations that are expected to have more errors in the original data (stations from Statens Vegvesen) in order to give us the possibility for more hits with the SCT when evaluating against the official QC for the full year.

For precipitation the results show more hits and fewer misses for the lowest thresholds, but also a higher false alarm rate and fewer correct negatives. The amount of hits vs false alarms and misses gives us an ETS that is close to zero independent of the threshold chosen. In order to increase the ETS, we want in the future to test a wider range of settings for precipitation. Preliminary testing has showed differences in the amount of flagged data when varying the *k*. The size of the inner and outer circles is another option. Like for temperature, the range of valid and admissible values should also be investigated further. Several case studies for precipitation over the city of Oslo have been started, and the results from these should help shape the settings used in the future.

Lastly, it will be interesting to run the SCT in combination with some of the other tests within titanlib. A combination of tests has the potential of both good performance and decreased run time. Another step in decreasing the run time for the SCT for testing purposes is to make the test run only for sub-regions around the stations to be evaluated and not for the entire domain.

A Categorical statistics used for evaluation

A useful reference is the website `https://www.cawcr.gov.au/projects/verification/` and references reported there.

The contingency table used in our study can be seen in Tab. 4.

	observation is bad	observation is good
SCT flags observation as bad	hits	false alarms
SCT flags observation as good	misses	correct negatives

Table 4: Contingency table used in quality control.

A hit is defined as when the SCT flags a bad observation as bad, while a correct negative is when the SCT flags a good observation as good. A miss is when the SCT flags a bad observation as good, and a false alarm is when the SCT flags a good observation as bad.

The statistics we use are reported in the following.

Probability of detection (POD):

$$\mathrm{POD} = \frac{\mathrm{hits}}{\mathrm{hits} + \mathrm{misses}} \tag{1}$$

Probability of false detection (POFD):

$$\mathrm{POFD} = \frac{\mathrm{false\ alarms}}{\mathrm{correct\ negatives} + \mathrm{false\ alarms}} \tag{2}$$

Equitable threat score (ETS):

$$\mathrm{ETS} = \frac{\mathrm{hits} - \mathrm{hits}_{\mathrm{random}}}{\mathrm{hits} + \mathrm{misses} + \mathrm{false\ alarms} - \mathrm{hits}_{\mathrm{random}}} \tag{3}$$

where

$$\mathrm{hits}_{\mathrm{random}} = \frac{(\mathrm{hits} + \mathrm{misses})(\mathrm{hits} + \mathrm{false\ alarms})}{\mathrm{total}} \tag{4}$$

B Supplementary material

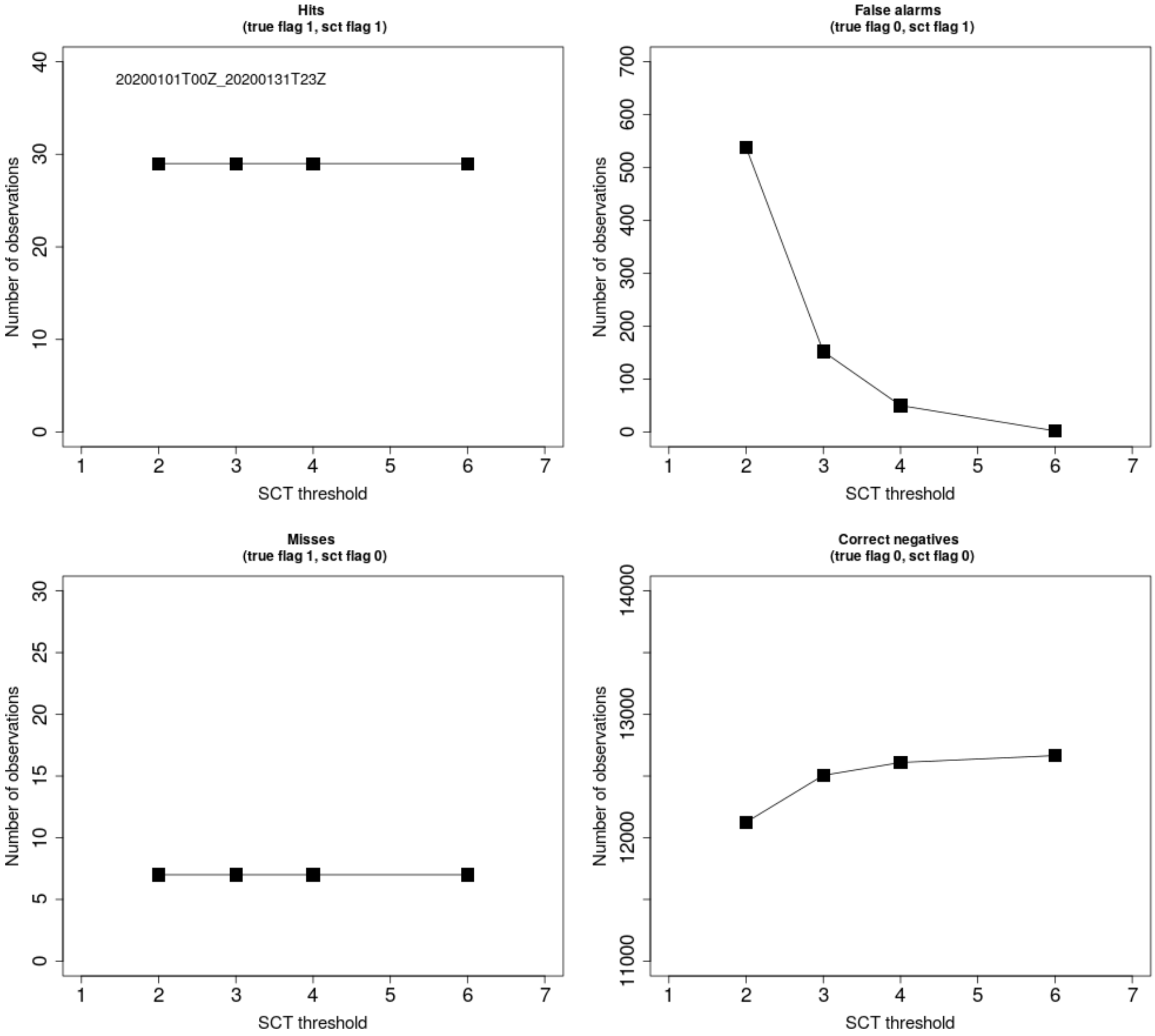


Figure 13: Number of temperature observations found to be hits (upper left panel), false alarms (upper right panel), misses (lower left panel), and correct negatives (lower right panel), for the different SCT thresholds 2, 3, 4, and 6, for January 2020.

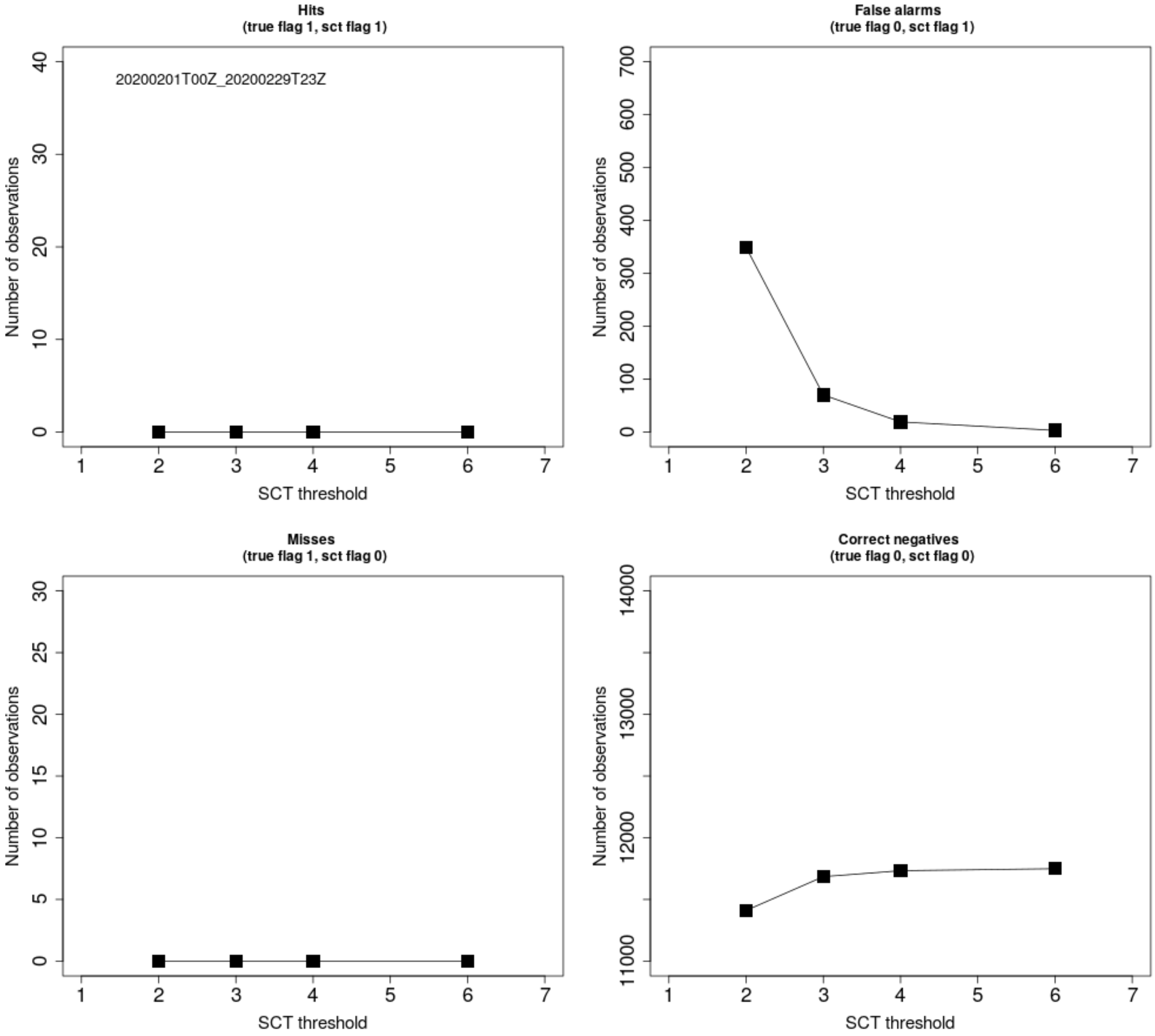


Figure 14: Number of temperature observations found to be hits (upper left panel), false alarms (upper right panel), misses (lower left panel), and correct negatives (lower right panel), for the different SCT thresholds 2, 3, 4, and 6, for February 2020.

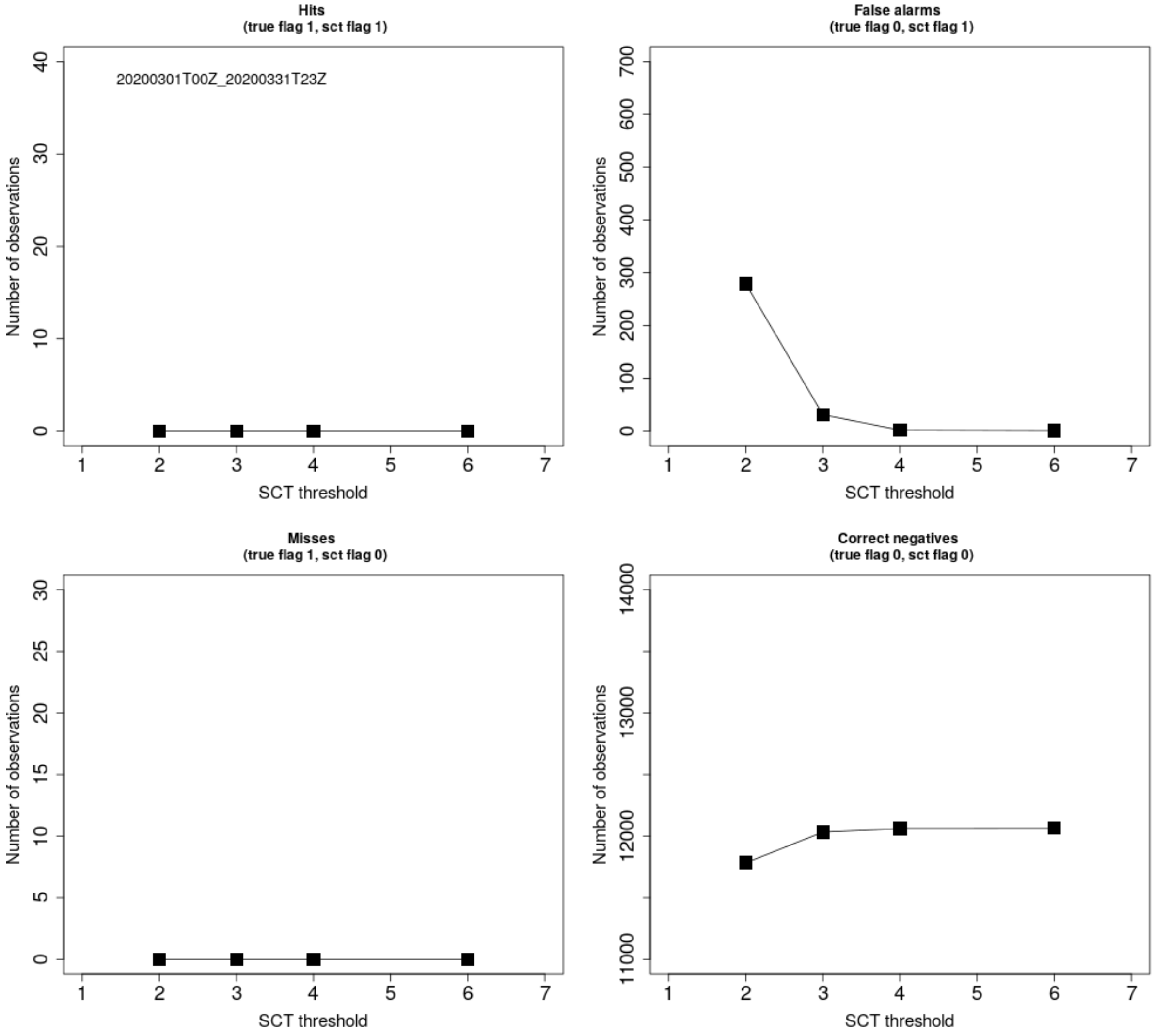


Figure 15: Number of temperature observations found to be hits (upper left panel), false alarms (upper right panel), misses (lower left panel), and correct negatives (lower right panel), for the different SCT thresholds 2, 3, 4, and 6, for March 2020.

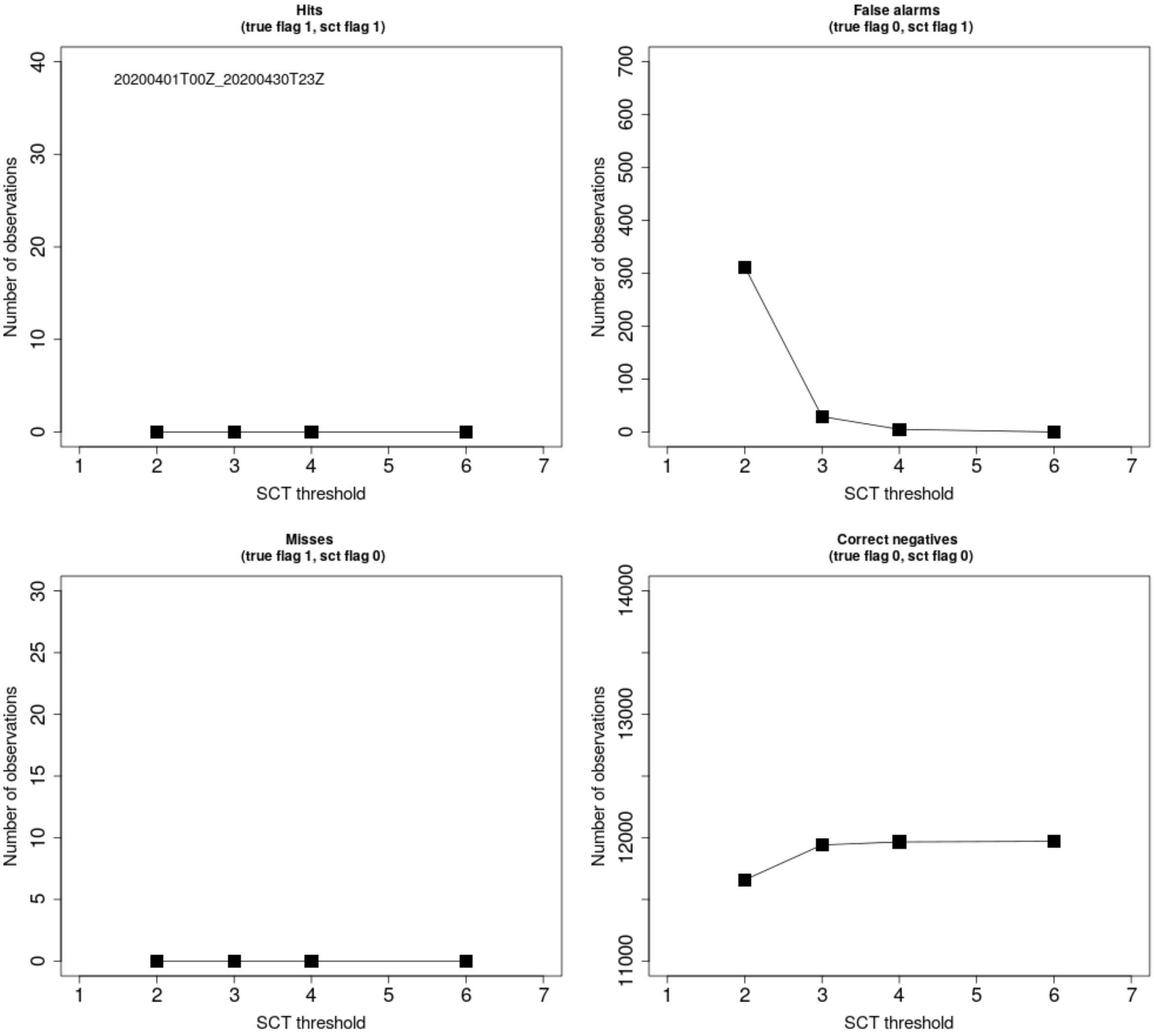


Figure 16: Number of temperature observations found to be hits (upper left panel), false alarms (upper right panel), misses (lower left panel), and correct negatives (lower right panel), for the different SCT thresholds 2, 3, 4, and 6, for April 2020.

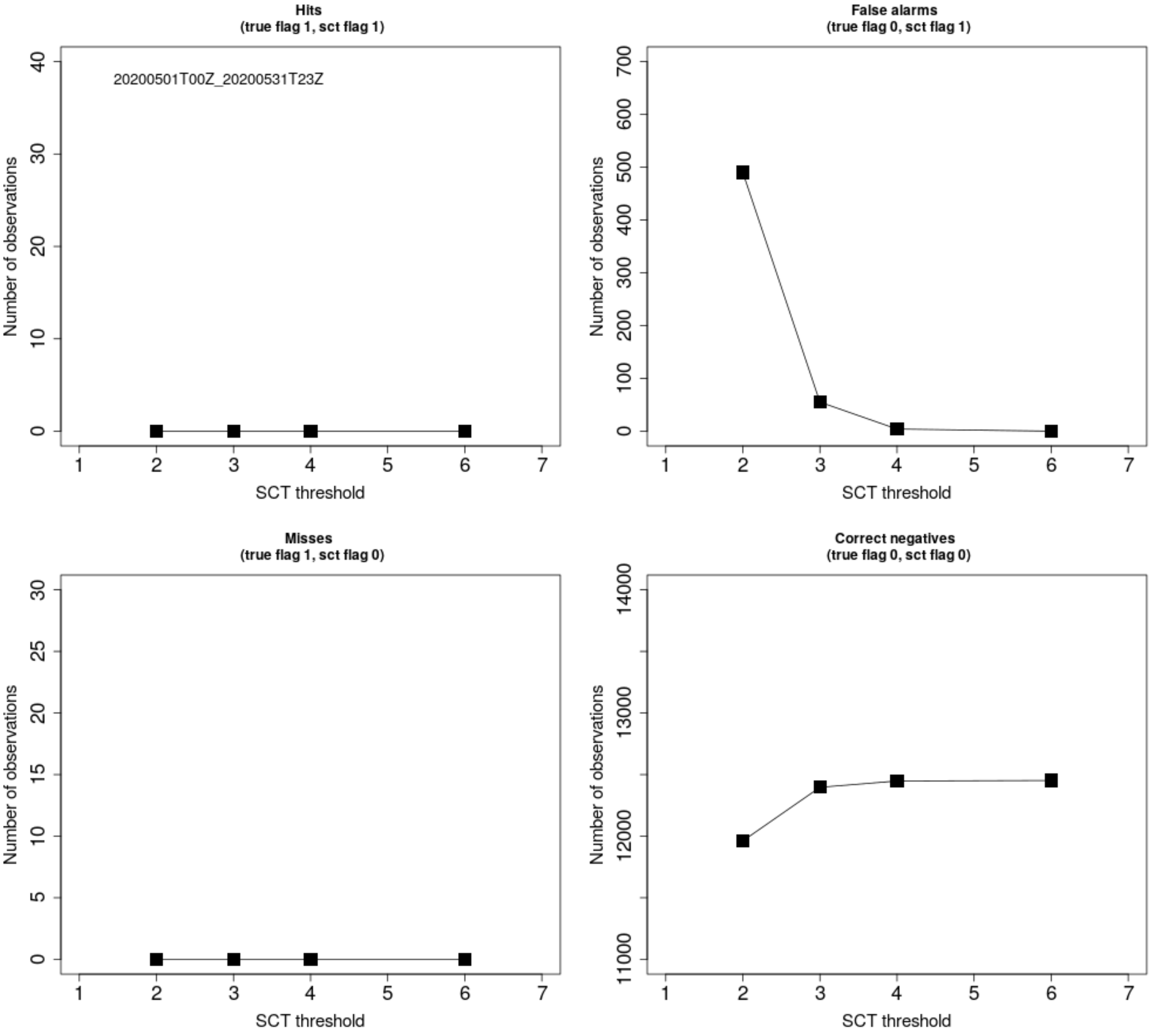


Figure 17: Number of temperature observations found to be hits (upper left panel), false alarms (upper right panel), misses (lower left panel), and correct negatives (lower right panel), for the different SCT thresholds 2, 3, 4, and 6, for May 2020.

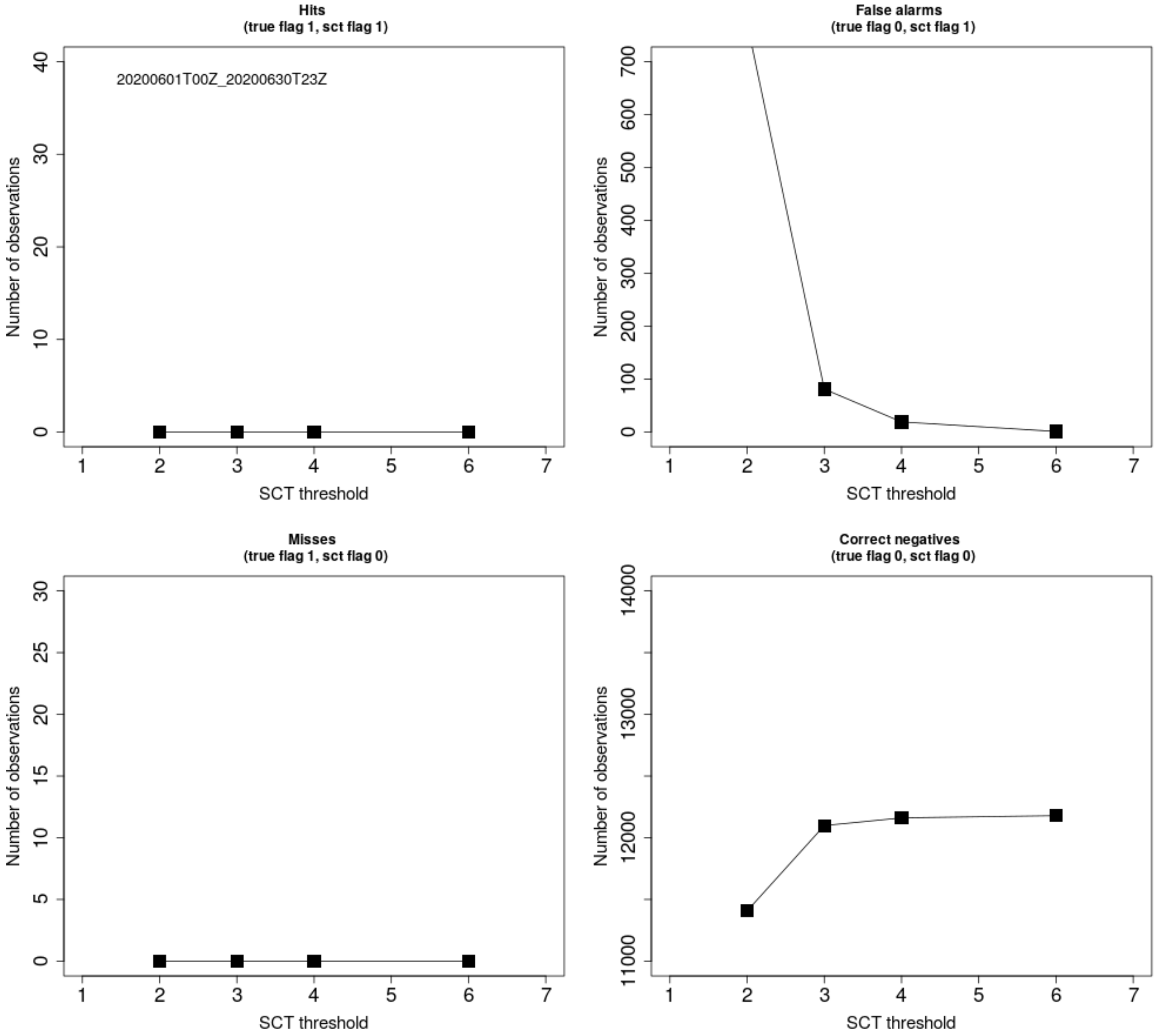


Figure 18: Number of temperature observations found to be hits (upper left panel), false alarms (upper right panel), misses (lower left panel), and correct negatives (lower right panel), for the different SCT thresholds 2, 3, 4, and 6, for June 2020.

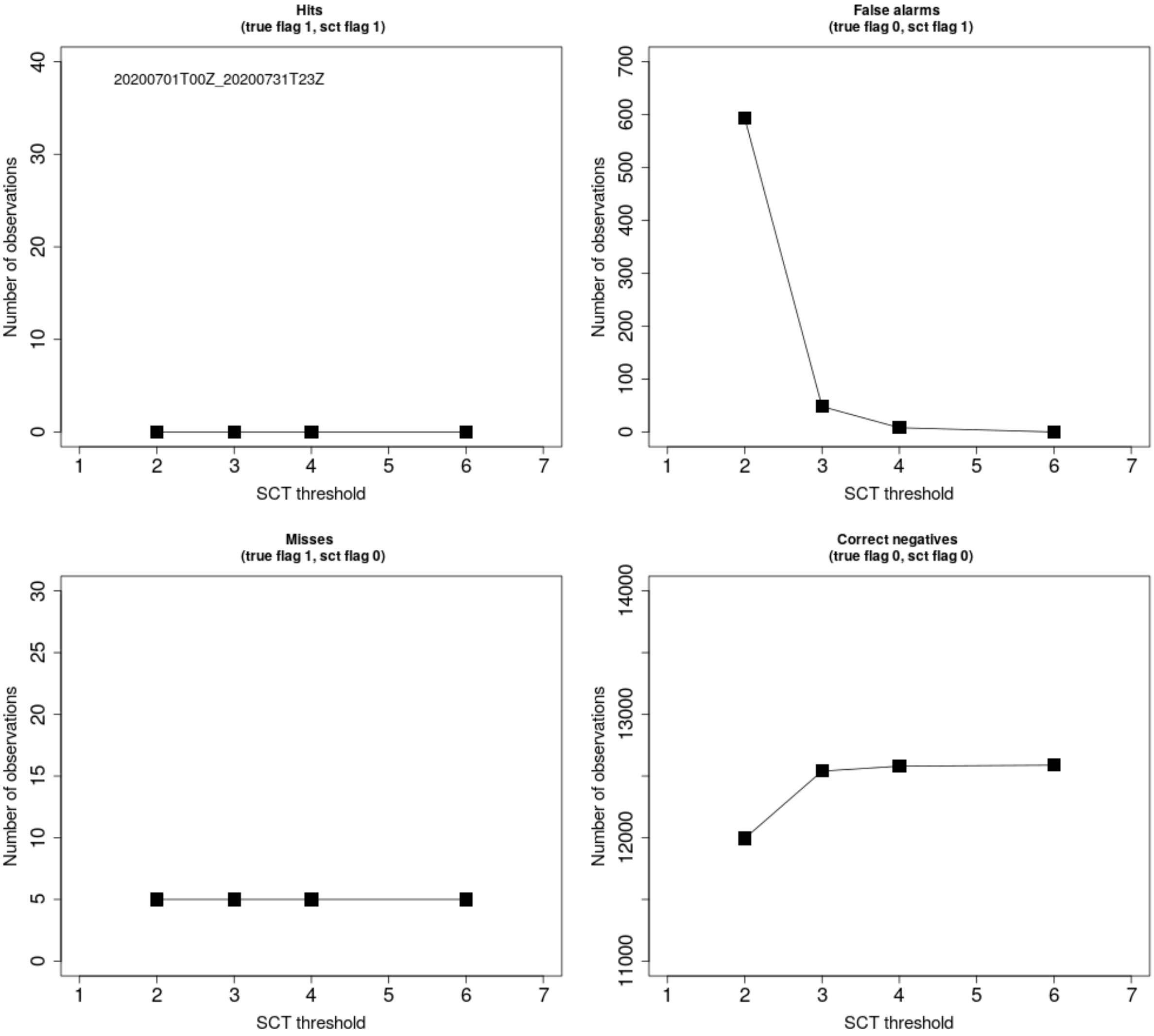


Figure 19: Number of temperature observations found to be hits (upper left panel), false alarms (upper right panel), misses (lower left panel), and correct negatives (lower right panel), for the different SCT thresholds 2, 3, 4, and 6, for July 2020.

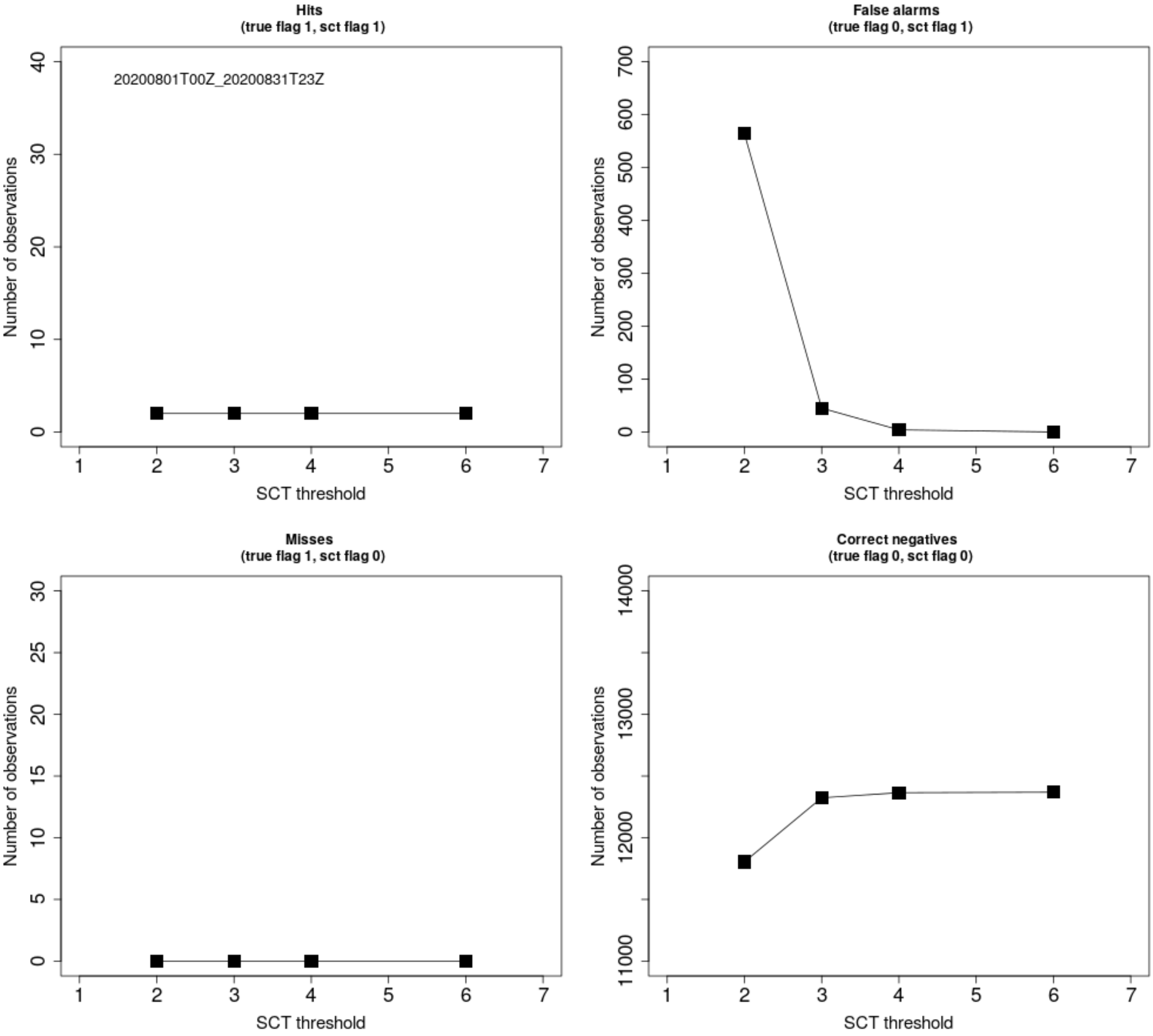


Figure 20: Number of temperature observations found to be hits (upper left panel), false alarms (upper right panel), misses (lower left panel), and correct negatives (lower right panel), for the different SCT thresholds 2, 3, 4, and 6, for August 2020.

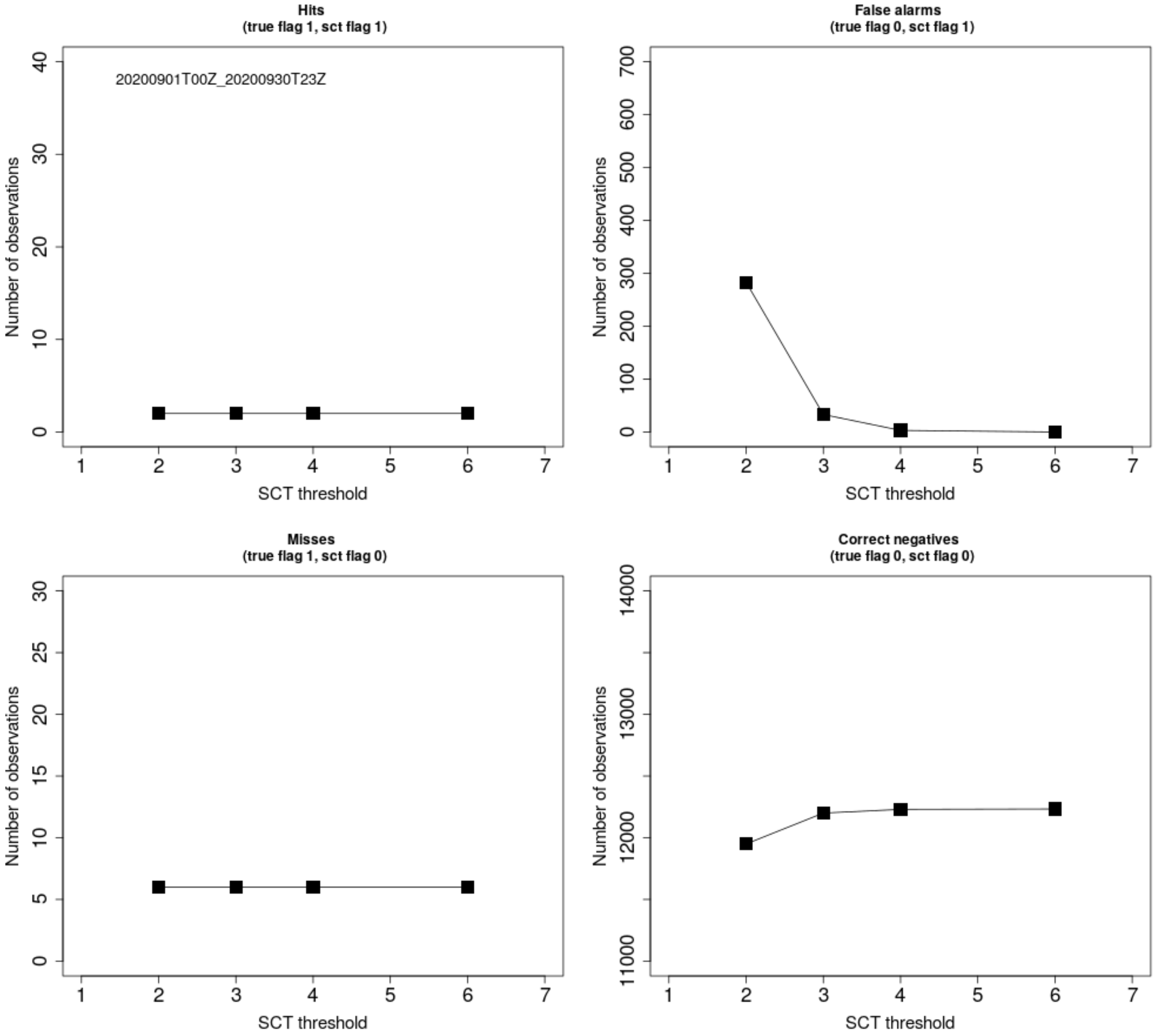


Figure 21: Number of temperature observations found to be hits (upper left panel), false alarms (upper right panel), misses (lower left panel), and correct negatives (lower right panel), for the different SCT thresholds 2, 3, 4, and 6, for September 2020.

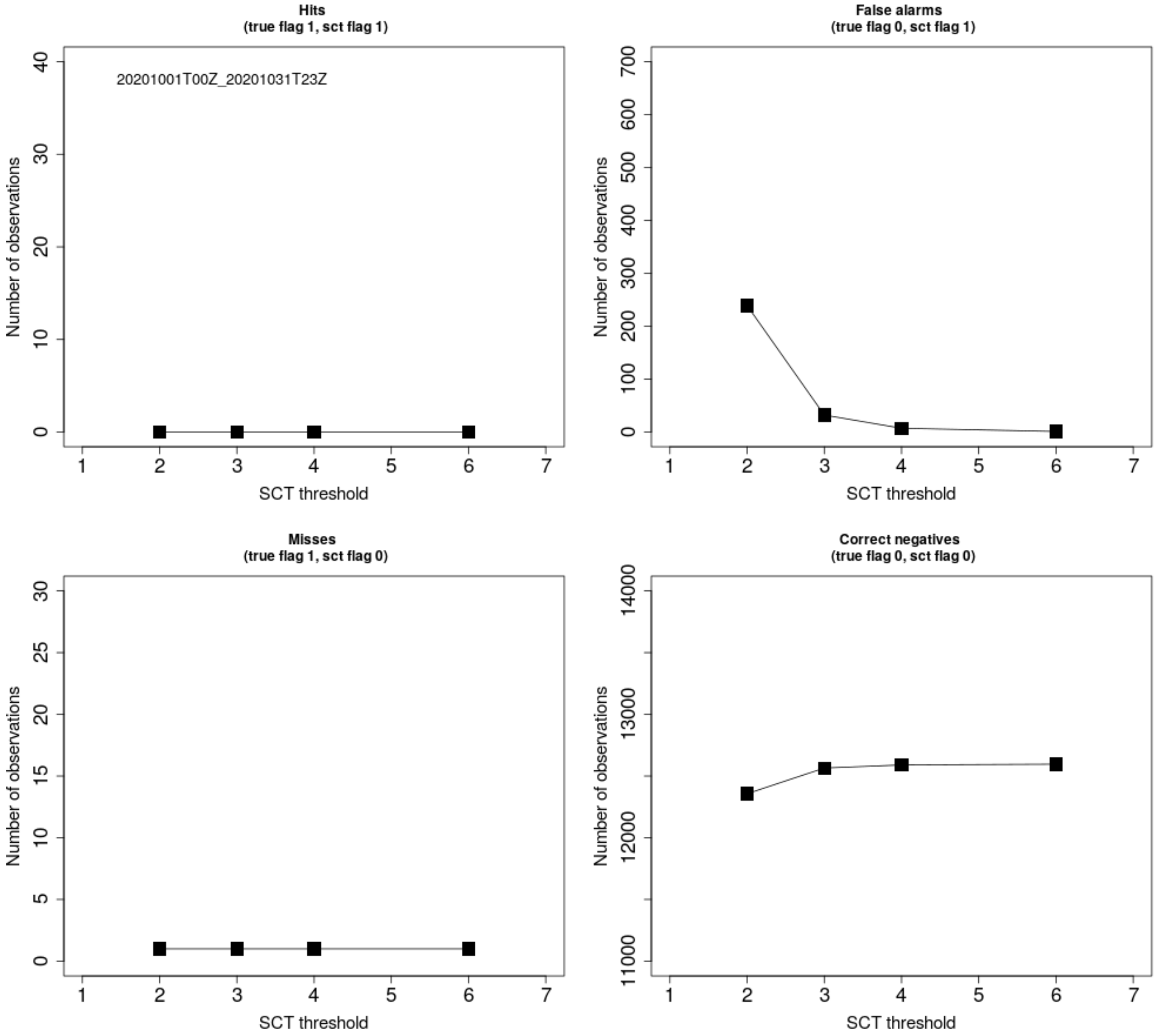


Figure 22: Number of temperature observations found to be hits (upper left panel), false alarms (upper right panel), misses (lower left panel), and correct negatives (lower right panel), for the different SCT thresholds 2, 3, 4, and 6, for October 2020.

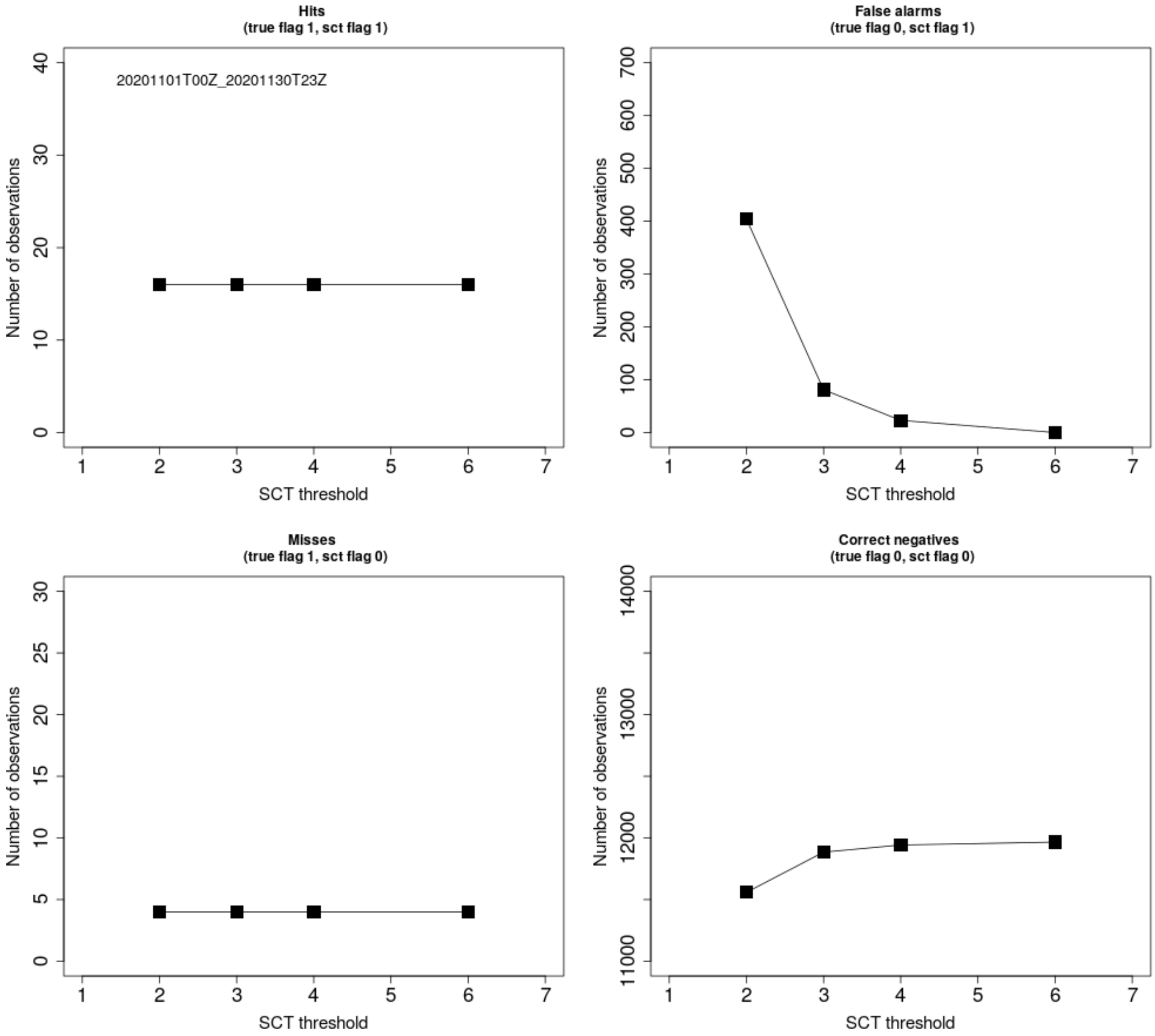


Figure 23: Number of temperature observations found to be hits (upper left panel), false alarms (upper right panel), misses (lower left panel), and correct negatives (lower right panel), for the different SCT thresholds 2, 3, 4, and 6, for November 2020.

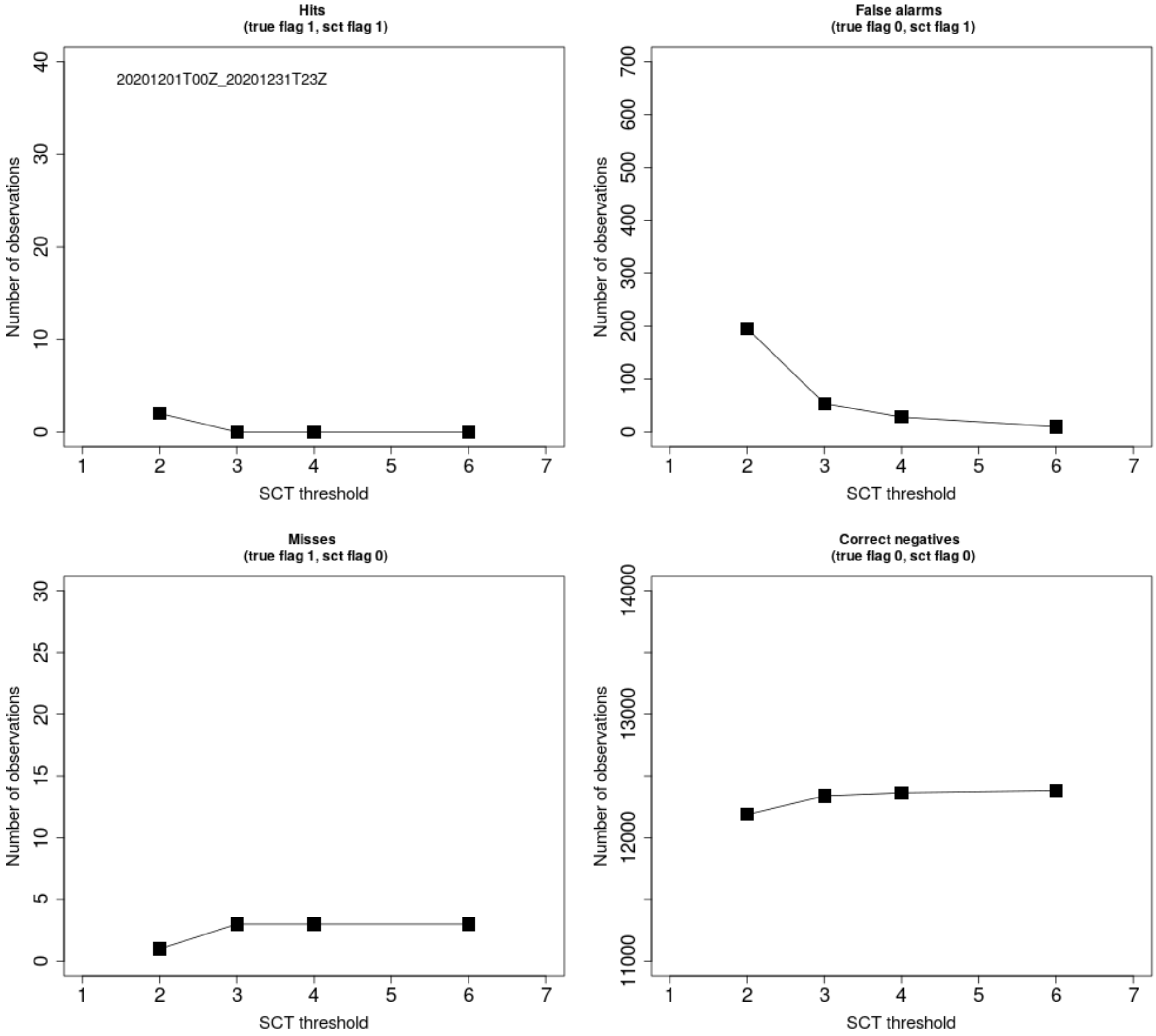


Figure 24: Number of temperature observations found to be hits (upper left panel), false alarms (upper right panel), misses (lower left panel), and correct negatives (lower right panel), for the different SCT thresholds 2, 3, 4, and 6, for December 2020.

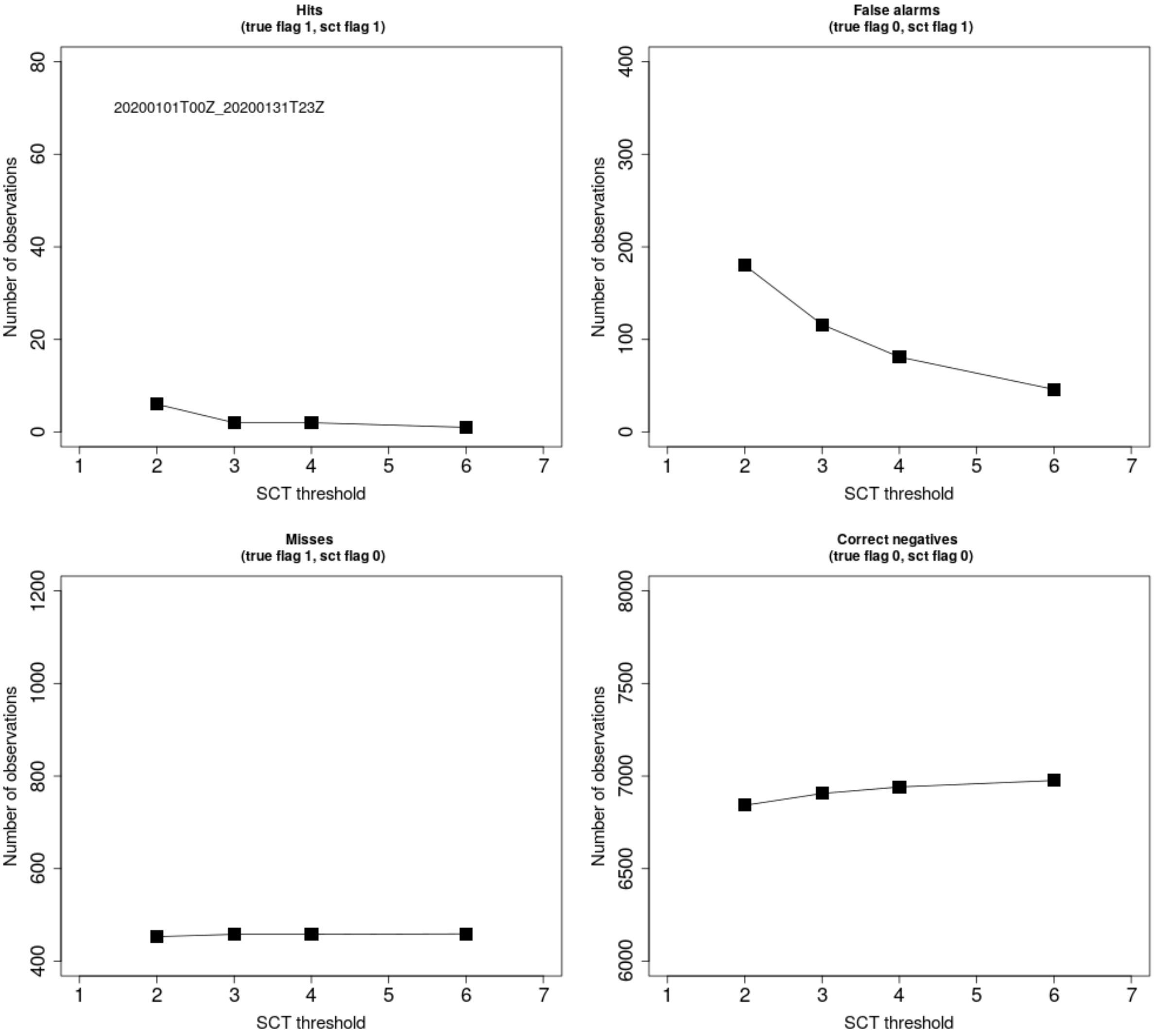


Figure 25: Number of precipitation observations found to be hits (upper left panel), false alarms (upper right panel), misses (lower left panel), and correct negatives (lower right panel), for the different SCT thresholds 2, 3, 4, and 6, for January 2020.

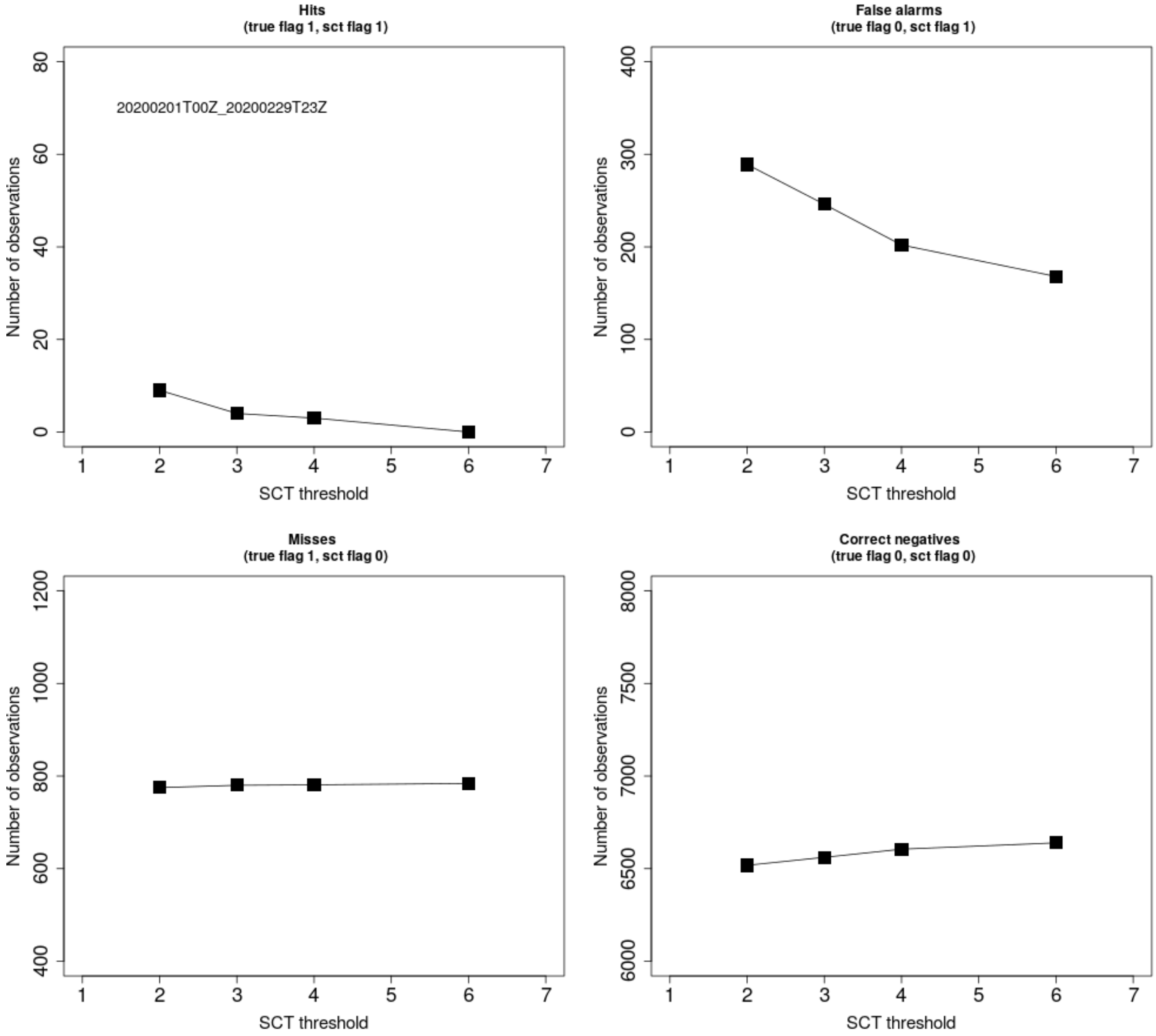


Figure 26: Number of precipitation observations found to be hits (upper left panel), false alarms (upper right panel), misses (lower left panel), and correct negatives (lower right panel), for the different SCT thresholds 2, 3, 4, and 6, for February 2020.

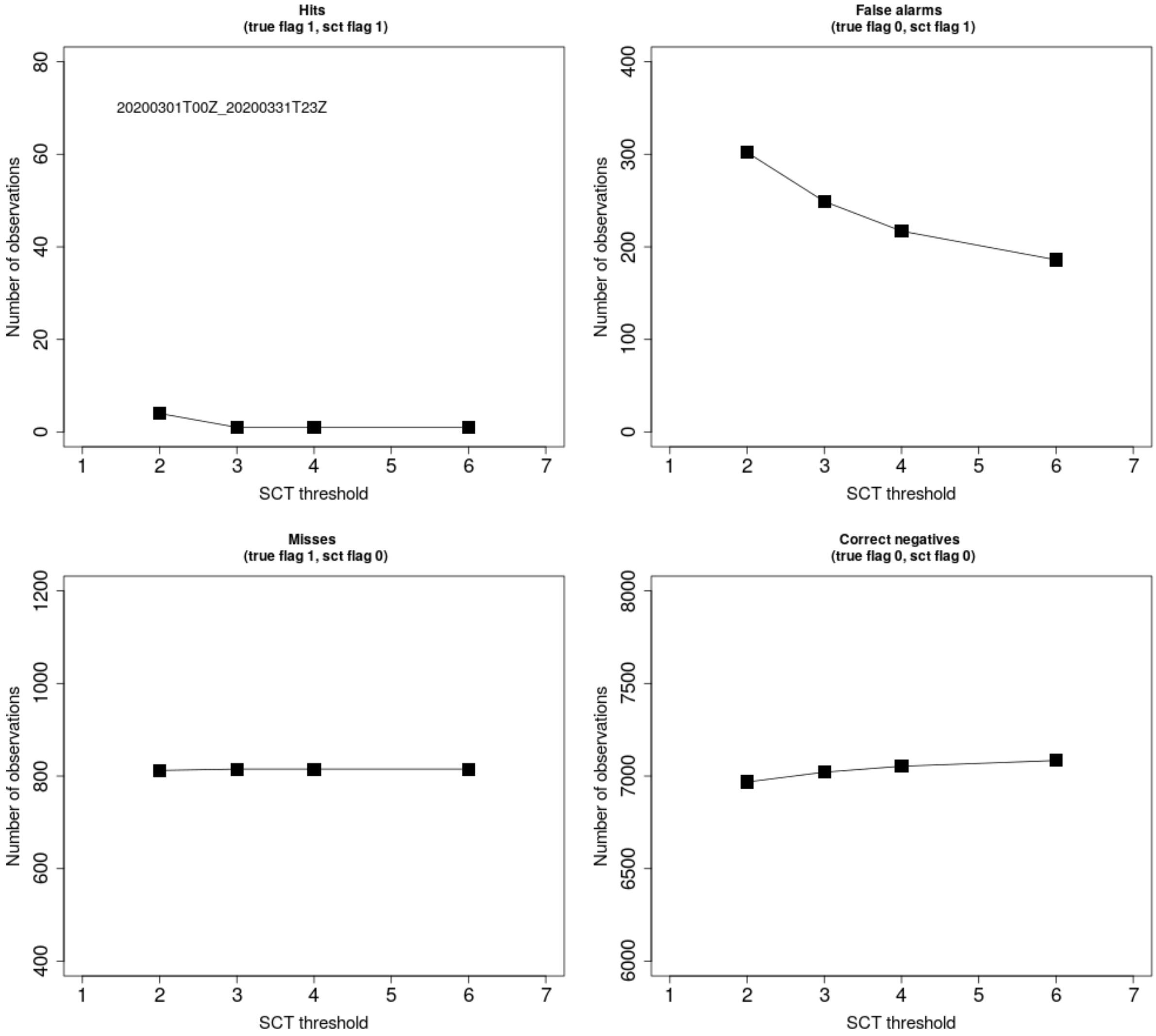


Figure 27: Number of precipitation observations found to be hits (upper left panel), false alarms (upper right panel), misses (lower left panel), and correct negatives (lower right panel), for the different SCT thresholds 2, 3, 4, and 6, for March 2020.

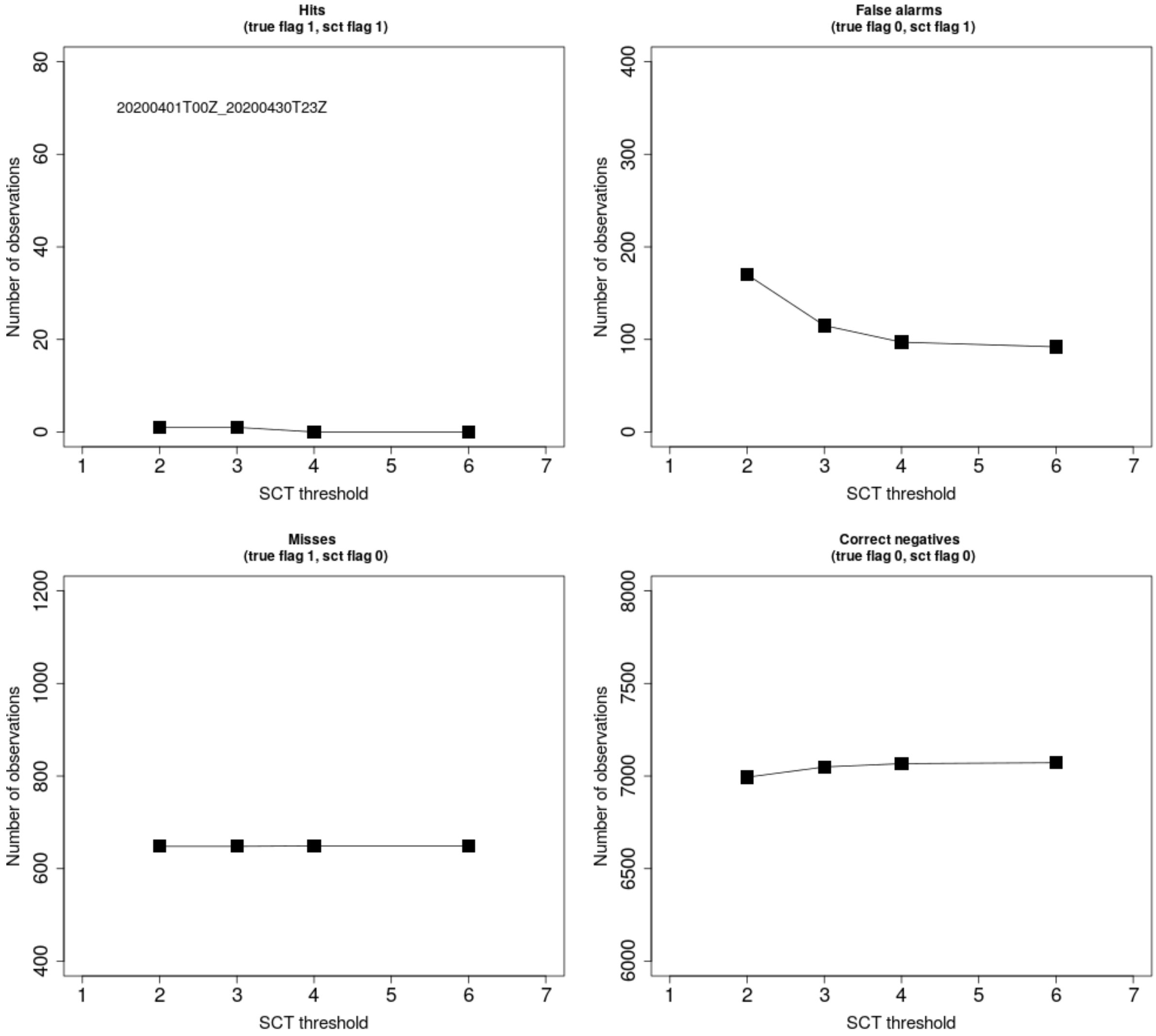


Figure 28: Number of precipitation observations found to be hits (upper left panel), false alarms (upper right panel), misses (lower left panel), and correct negatives (lower right panel), for the different SCT thresholds 2, 3, 4, and 6, for April 2020.

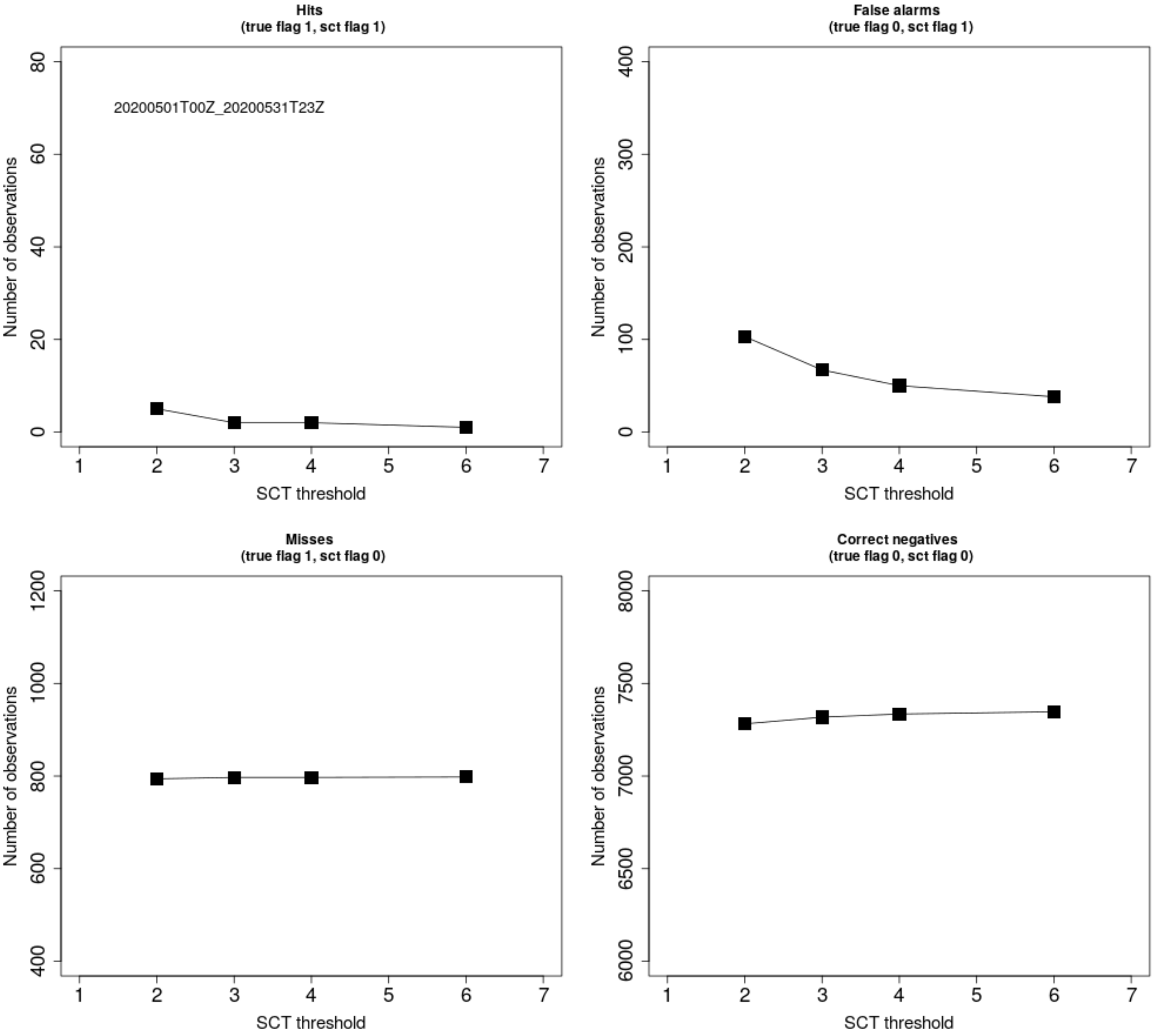


Figure 29: Number of precipitation observations found to be hits (upper left panel), false alarms (upper right panel), misses (lower left panel), and correct negatives (lower right panel), for the different SCT thresholds 2, 3, 4, and 6, for May 2020.

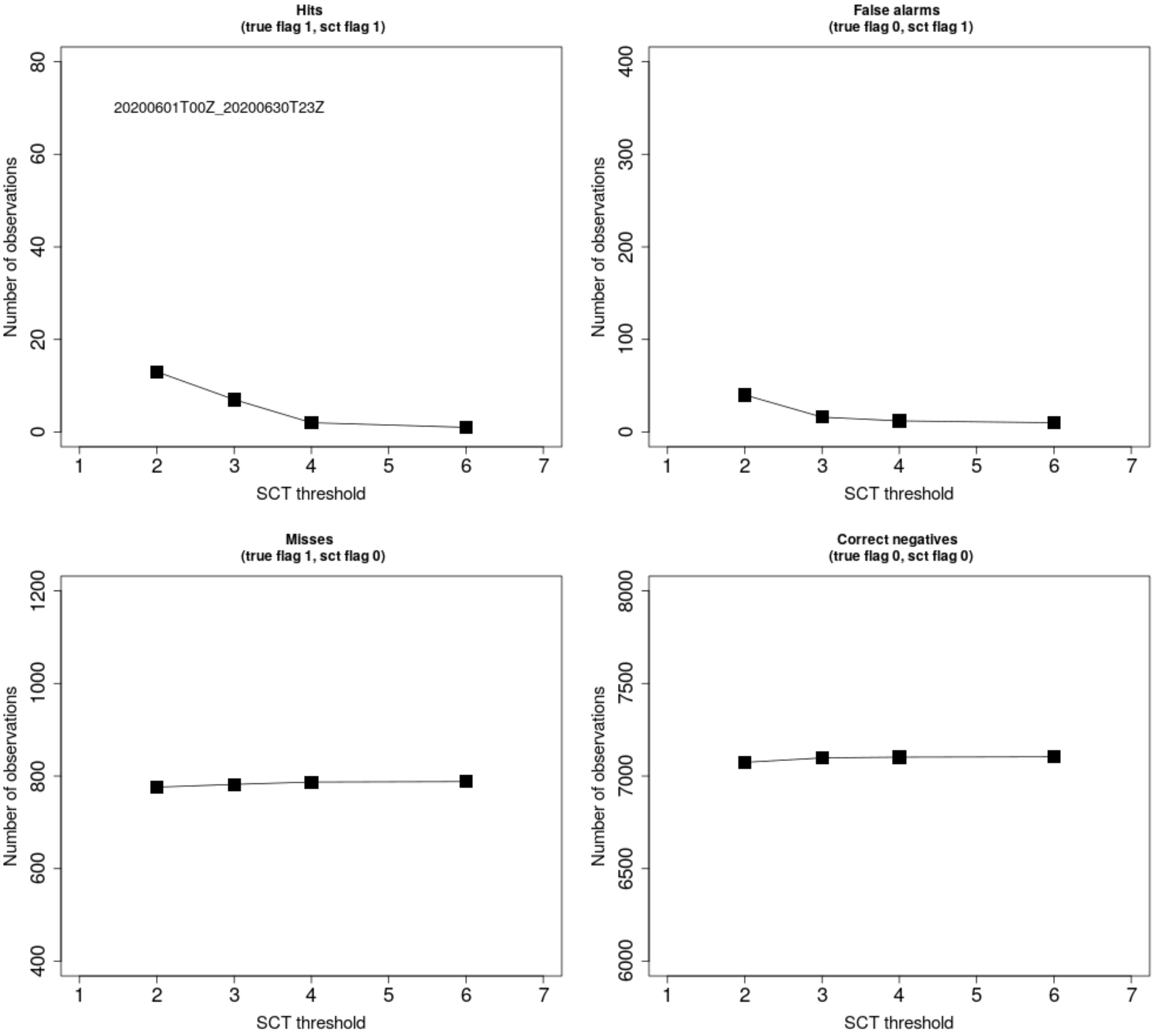


Figure 30: Number of precipitation observations found to be hits (upper left panel), false alarms (upper right panel), misses (lower left panel), and correct negatives (lower right panel), for the different SCT thresholds 2, 3, 4, and 6, for June 2020.

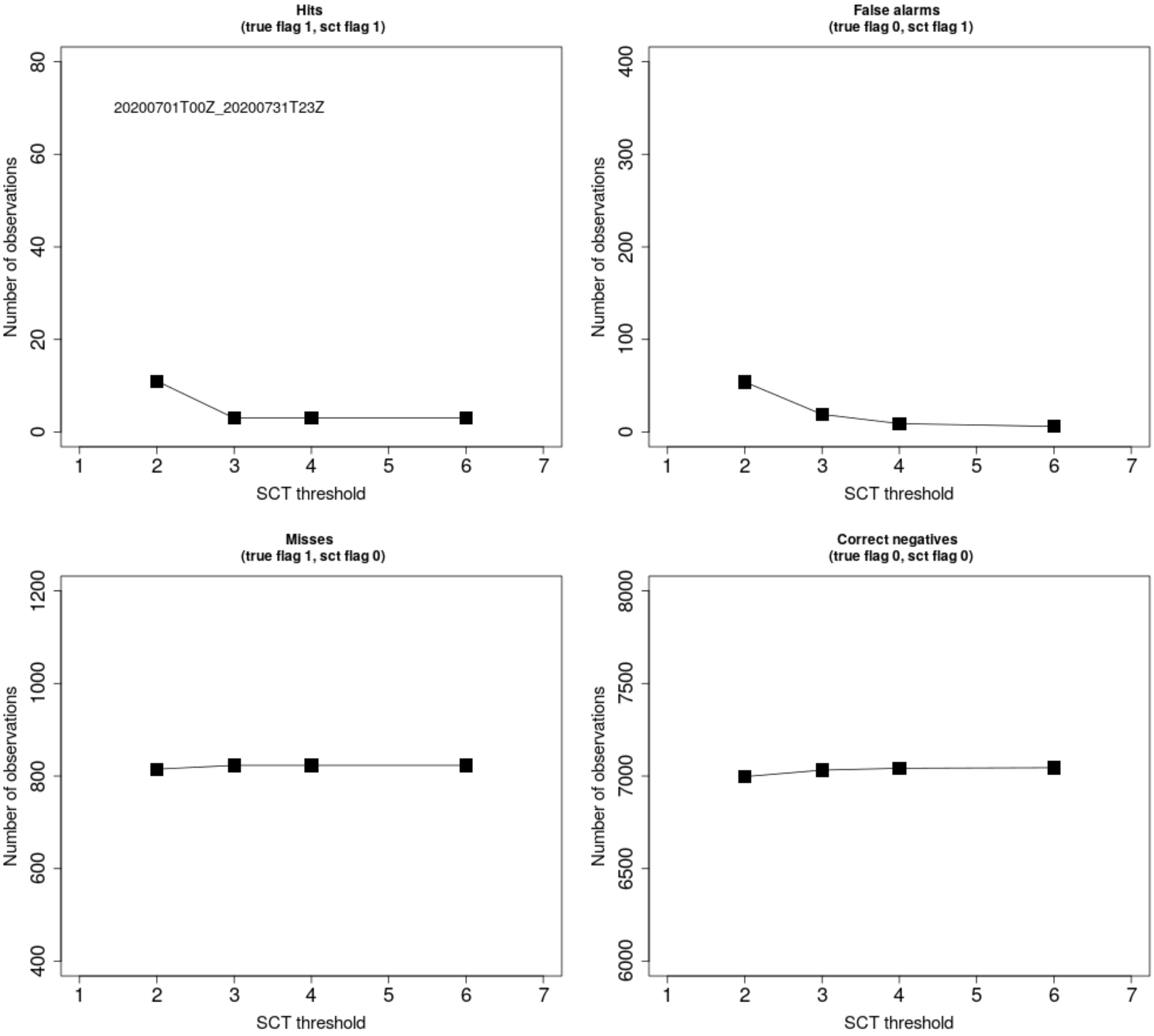


Figure 31: Number of precipitation observations found to be hits (upper left panel), false alarms (upper right panel), misses (lower left panel), and correct negatives (lower right panel), for the different SCT thresholds 2, 3, 4, and 6, for July 2020.

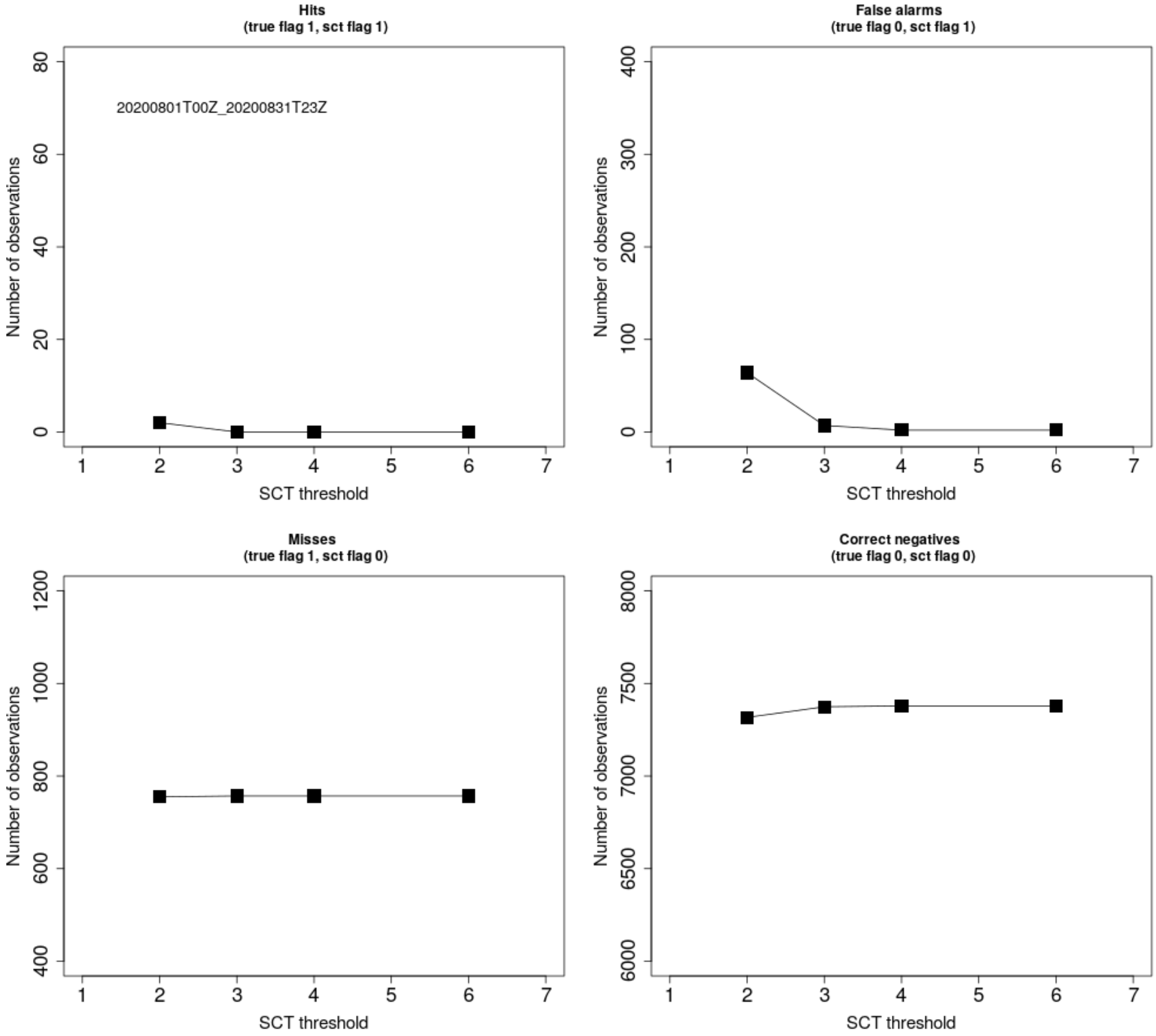


Figure 32: Number of precipitation observations found to be hits (upper left panel), false alarms (upper right panel), misses (lower left panel), and correct negatives (lower right panel), for the different SCT thresholds 2, 3, 4, and 6, for August 2020.

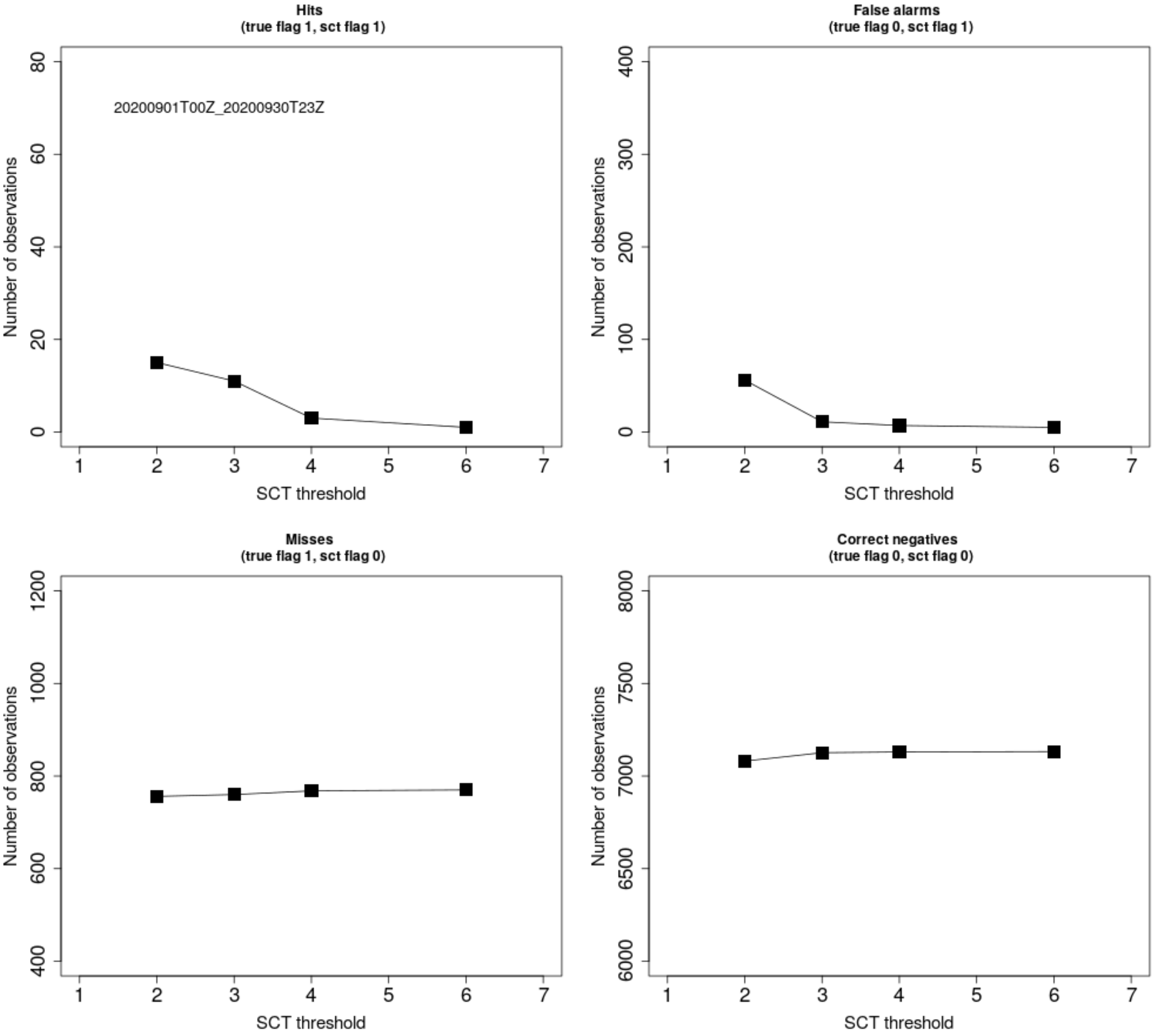


Figure 33: Number of precipitation observations found to be hits (upper left panel), false alarms (upper right panel), misses (lower left panel), and correct negatives (lower right panel), for the different SCT thresholds 2, 3, 4, and 6, for September 2020.

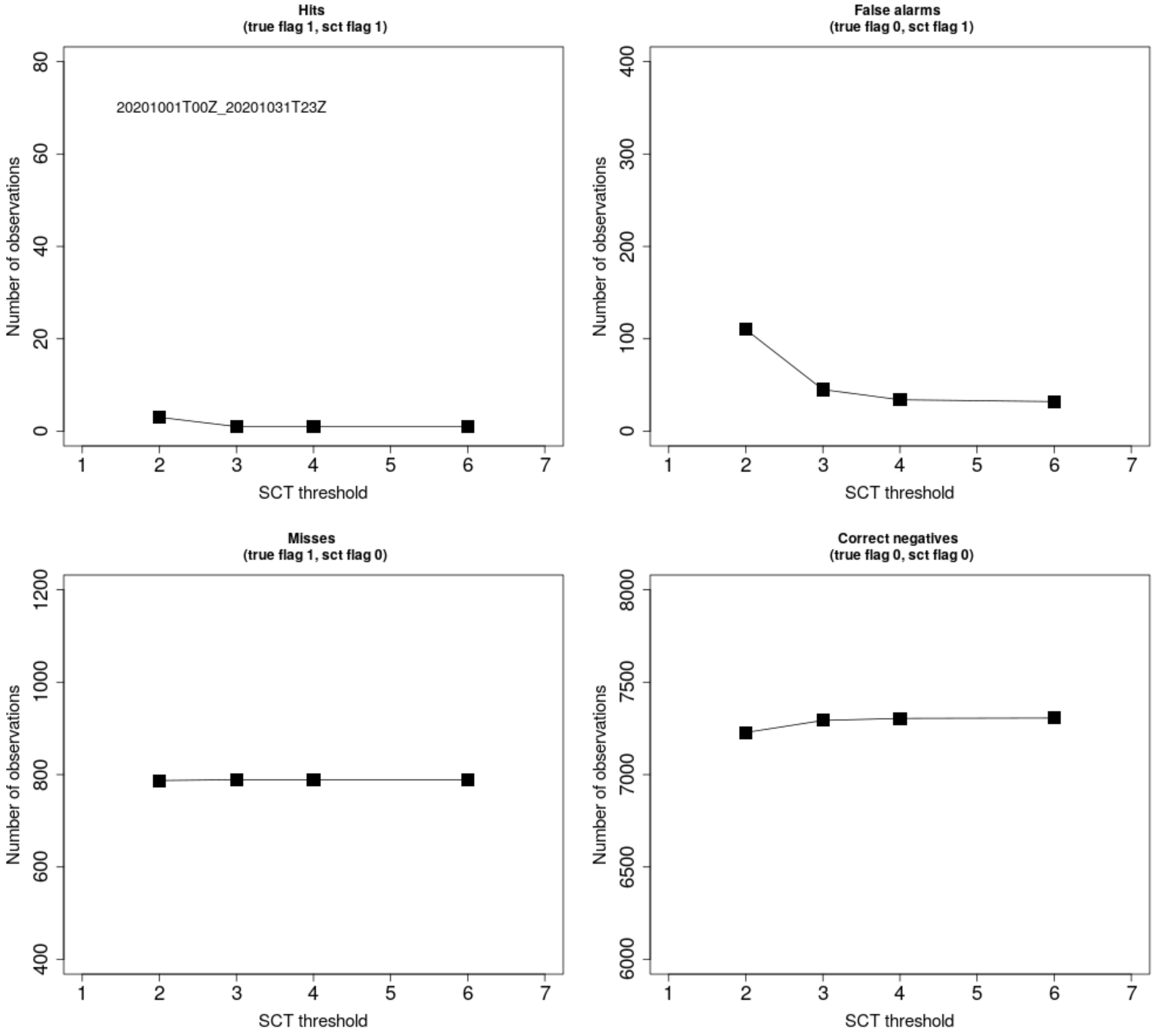


Figure 34: Number of precipitation observations found to be hits (upper left panel), false alarms (upper right panel), misses (lower left panel), and correct negatives (lower right panel), for the different SCT thresholds 2, 3, 4, and 6, for October 2020.

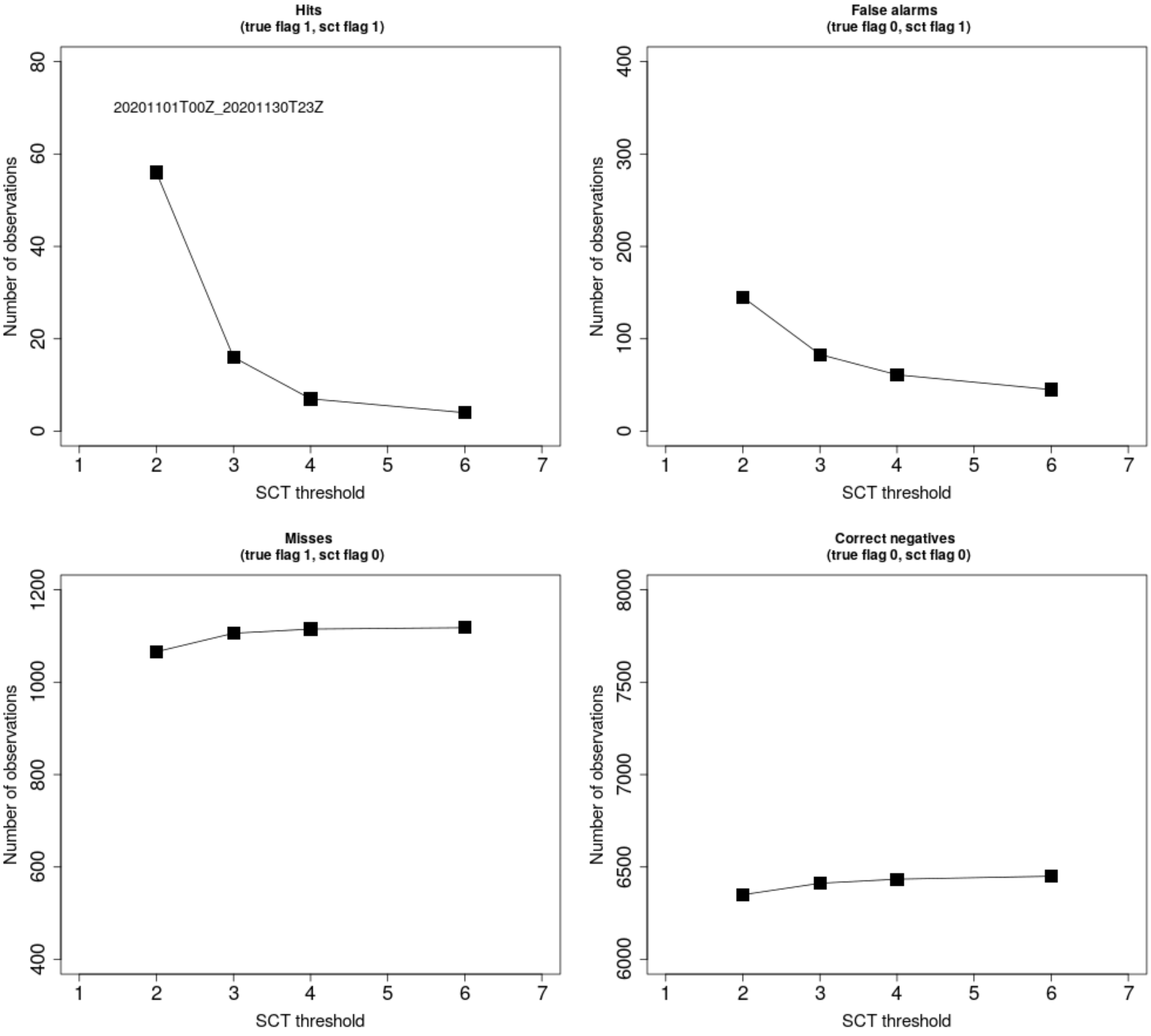


Figure 35: Number of precipitation observations found to be hits (upper left panel), false alarms (upper right panel), misses (lower left panel), and correct negatives (lower right panel), for the different SCT thresholds 2, 3, 4, and 6, for November 2020.

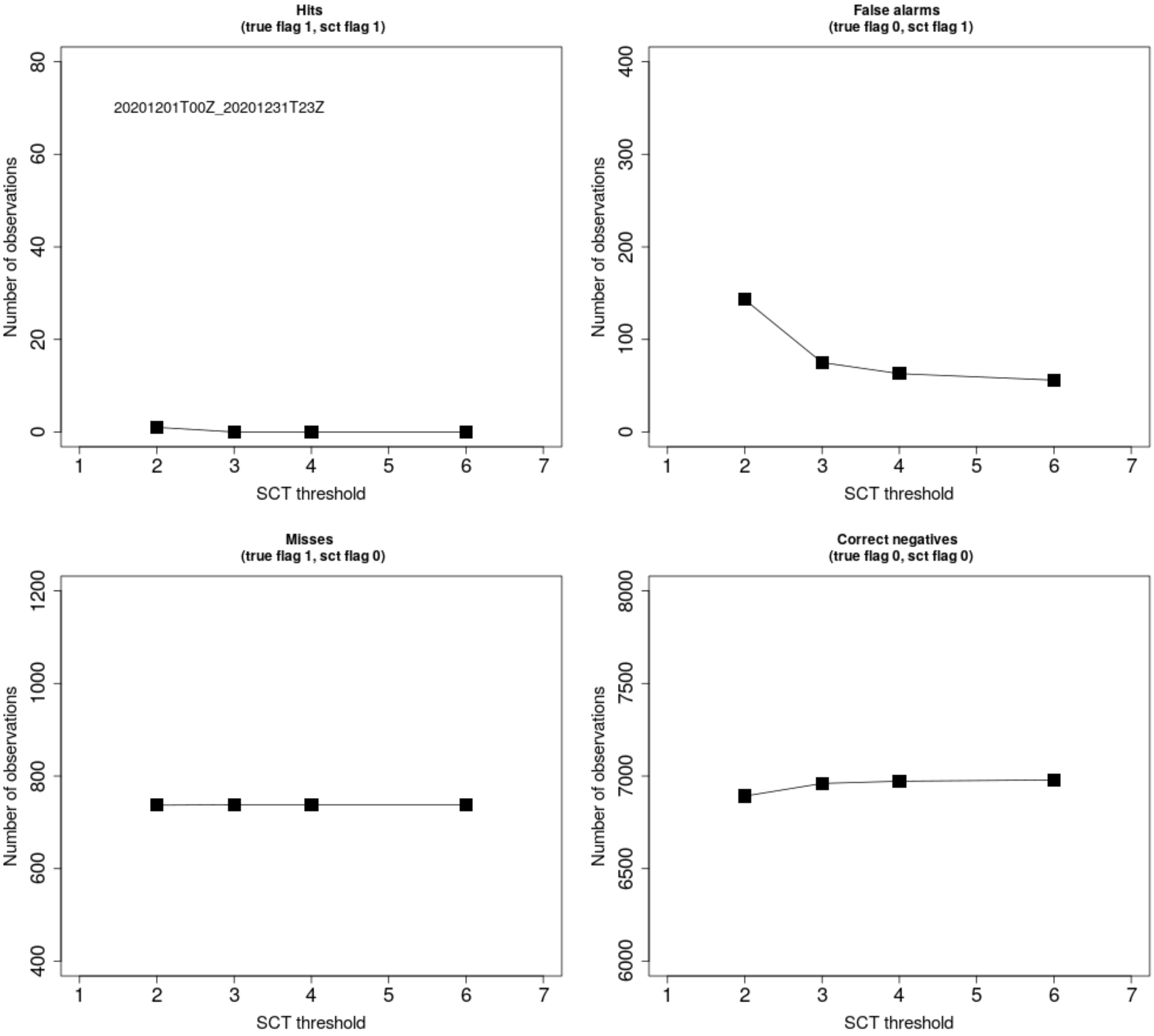


Figure 36: Number of precipitation observations found to be hits (upper left panel), false alarms (upper right panel), misses (lower left panel), and correct negatives (lower right panel), for the different SCT thresholds 2, 3, 4, and 6, for December 2020.

References

Lussana, C., and L. Båserud (2021), A spatial consistency test for the quality control of meteorological observations, Part I: Methodology, *Tech. rep.*, Norwegian Meteorological Institute.