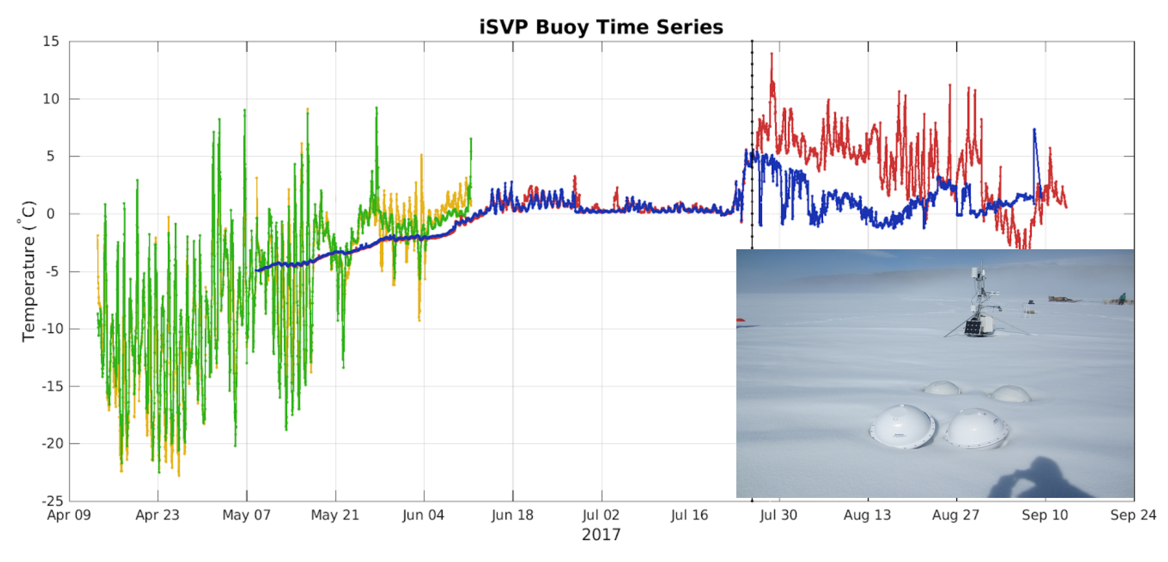
Technical Report

Detailed investigation of the uncertainty budget for Non-recoverable IST observations and their SI traceability



ESA Contract No. 4000113848_15I-LG

Jacob L. Høyer and Emy Alerskans, Pia Nielsen-Englyst, Peter Thejll, Gorm Dybkjær and Rasmus Tonboe,

DECEMBER 2017

INTENTIONALLY BLANK

Fiducial Reference Measurements for validation of Surface Temperature from Satellites (FRM4STS): Ice Surface Temperature Comparison of Participants Radiometers

Detailed investigation of the uncertainty budget for Non-recoverable IST observations and their SI traceability

Jacob L. Høyer and Emy Alerskans, Pia Nielsen-Englyst, Peter Thejll, Gorm Dybkjær and Rasmus Tonboe,

Danish Meteorological Institute

DMI
Danish Meteorological Institute

Danish Meteorological Institute
Lyngbyvej 100, 2100 Copenhagen Ø

CONTENTS

DOCUMENT VERSION HISTORY

DOCUMENT APPROVAL

DOCUMENT MANAGEMENT

Issue	Revision	Date of Issue/revision	Description of Changes
1	1	20-Oct-15	Creation of document

DOCUMENT APPROVAL

Contractor Approval

Name	Role in Project	Signature & Date (dd/mm/yyyy)
Dr Nigel Fox	Technical Leader	
Mr David Gibbs	Project Manager	

CUSTOMER APPROVAL

Name	Role in Project	Signature	Date (dd/mm/yyyy)
C Donlon	ESA Technical Officer		

APPLICABLE DOCUMENTS

AD Ref.	Ver. /Iss.	Title
EOP-SM/2642	1	Fiducial Reference Measurements for Thermal Infrared Satellite Validation (FRM4STS) Statement of Work

1 INTRODUCTION

The work presented in this report focus on the uncertainty budget for non-recoverable ice surface temperature (IST) observations when used to validate and calibrate satellite observations. The work with establishing fiducial reference measurements for ice surface temperature and agree upon protocols within the community is much less mature than for SST and LST. This means that today there is no consensus on a FRM IST data set that can be used for routine validation of the satellite IST products. Candidates for a routine IST FRM are the buoy data from the Data Buoy Cooperation Panel (DBCP) and the International Arctic Buoy Programme (IABP) that deploy iSVP buoys on the sea ice that typically report hourly temperature observations from the sea ice (see http://iabp.apl.washington.edu/index.html). However, very few studies have been performed to assess the uncertainty of these non-recoverable observations (e.g Rigor et al., 2000). Dybkjær et al., 2012 found that manual inspection of each buoy was needed in order to increase the quality of the ice drifting observations obtained from the Global Telecommunication System (GTS). In addition, they ended up discarding a vast majority of the observations. The conclusions in these papers pointed towards the need for an in-depth analysis of the uncertainties on the ice drifter observations.

Due to the nature of the ice surface temperature observations, the establishment of an IST FRM to be used for satellite climate data record validation must include both assessment of the sensor performance and the representativeness effects due to spatial and temporal variability. In particular, the vertical transformation of the ice drifting observations to the skin IST is very different if the sensor is 20 cm above the sea ice or covered with 5 cm of snow, this due to the large vertical gradients within the snow. To obtain a reliable IST FRM data set to be used for climate data record and routine validation, the magnitude of all these effects need to be assessed. In this option we quantify the uncertainty components related to the temperature observations from the ice surface drifters. The work is connected to option-1 in the FRM4STS proposal: “Study of SI Traceability for non-recoverable SST and IST FRM instruments” in terms of sensor degradation and performance of the SVP buoys. The results from this work have also been reported on the international FRM4STS workshop held at NPL in October, 2017.

2 QUANTIFICATION OF UNCERTAINTY BUDGET FOR SATELLITE AND IN SITU COMPARISONS

To establish the uncertainty budget for instruments in the Arctic, an inter-comparison of four ice Surface Velocity Program (iSVP) drifting buoys deployed on the sea ice off Qaanaaq, western Greenland, was performed. Furthermore, the iSVP buoy data were compared with data from an Automatic Weather Station (AWS) located only a few meters from the buoys. The first part of the analysis will focus on these observations to assess the differences between the different observations. See Figure 2.1 for the area of the instrument deployment and Høyer et al., 2017 for more details on the deployment site.

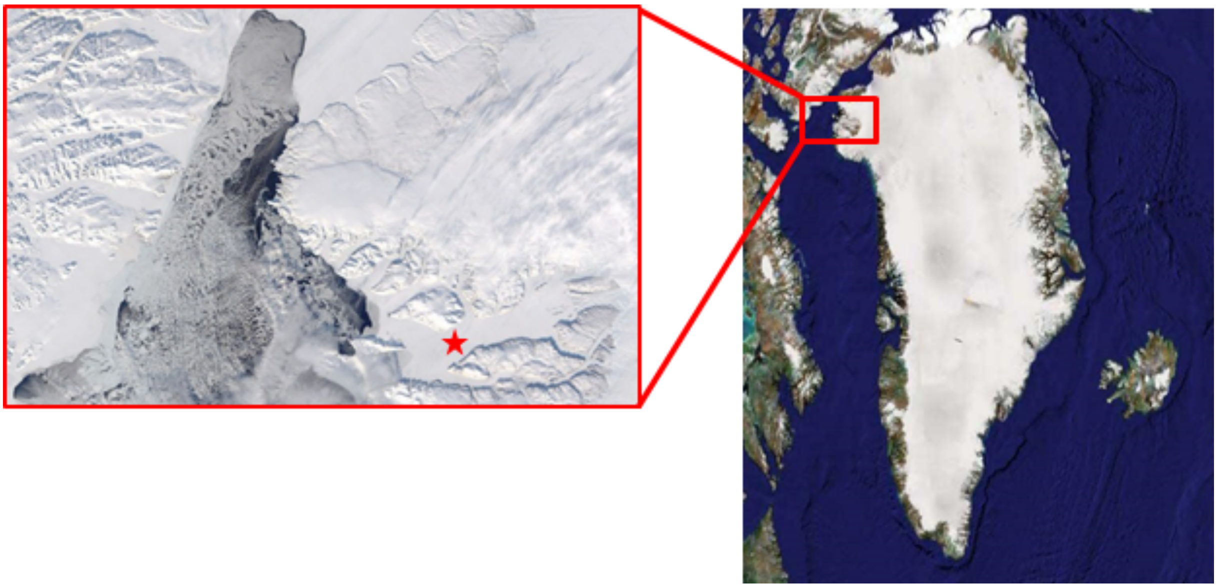


Figure 2.1: Location of the deployment of the iSVP buoys and the Automatic Weather Station (AWS). Qaanaaq is situated in Northwest Greenland at 77^{o}N. The instruments were deployed on the sea ice off Qaanaaq in Inglefield Bredning (left) marked with a red star.

2.1 BUOY ANALYSIS

2.1.1 Buoy deployment

Two sets of two iSVP buoys that measure temperature every hour were deployed in Qaanaaq, a few meters from the AWS (Figure 2.2). The first two buoys, which hereafter will be referred to as old buoy 1 and 2, were deployed on 25 January, 2017. However, it was discovered that these buoys could only measure temperatures above -5 °C due to a configuration error during the construction the buoys. If the temperature was below -5 °C, the two buoys would only register -5 °C. Therefore, two new iSVP buoys, hereafter referred to as new buoy 1 and 2, were deployed on 13 April, 2017 and later on collected on 11 June, 2017. The temporal coverage for all four buoys is shown in Figure 2.3. Note that even though the old buoys were deployed at the end of January, no observations are available before May. This is because the old buoys are not able to measure temperatures below -5 °C and therefore, all observations at -5 °C have been excluded, thus shortening the actual temporal coverage from approx. 8 months to about 5 months.

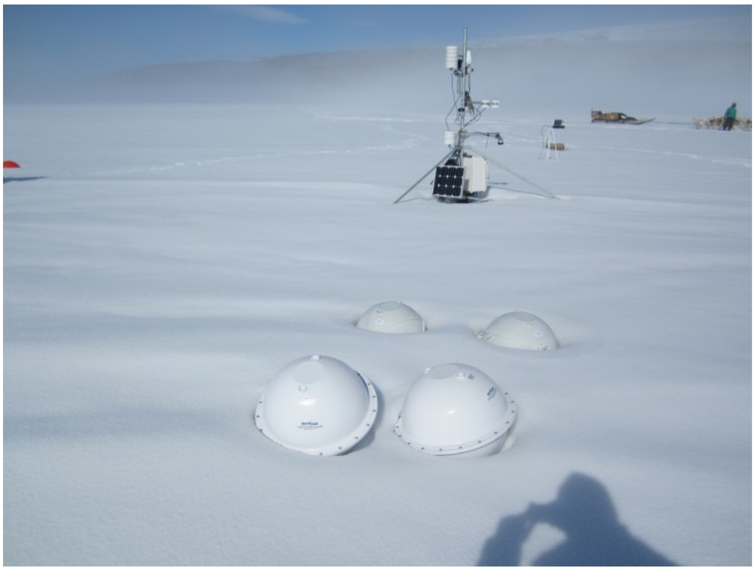


Figure 2.2: The four Surface Velocity Program (iSVP) ice buoys (white) with the Automatic Weather Station (AWS) in the background.

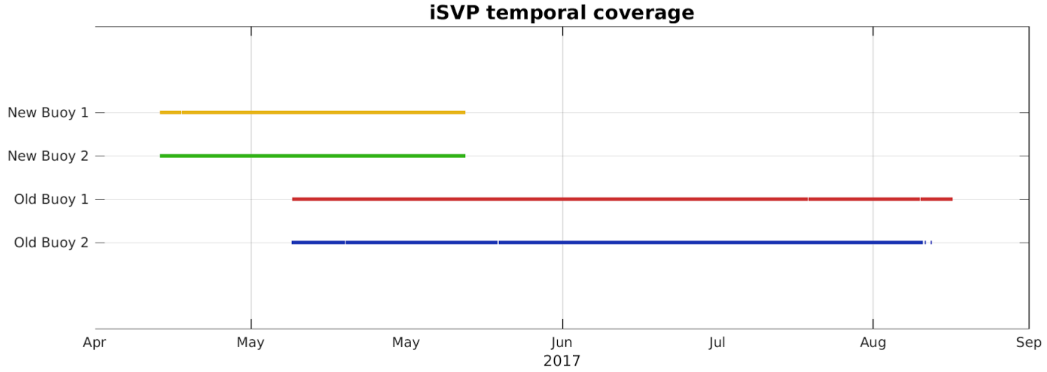


Figure 2.3: Temporal coverage of the four iSVP buoys: new buoy 1 (yellow), new buoy 2 (green), old buoy 1 (red) and old buoy 2 (blue). Note that even though the old buoys were deployed at the end of January, no observations are available before May. This is because the old buoys are not able to measure temperatures below -5 °C and therefore, all observations at -5 °C have been excluded, thus shortening the actual temporal coverage from approx. 8 months to about 5 months.

All four buoys were deployed at the same position, 77.465°N, 69.218°E, in pairs of two, and were stationed only a few meters from each other on the sea ice (see Figure 2.2). The two new buoys did not drift far from the original location and neither did they drift apart from each other as they were recovered before the ice broke up. The two old buoys, however, drifted apart when the ice broke up and began to drift in the water. Old buoy 1 stranded northwest of the original location, whereas old buoy 2 drifted southwest and ended up outside Devon Island, Canada. The time of the ice breakup drifting event and the position of the old buoys during the ice breakup event were calculated based on the distance between the buoys. Due to slightly different observational periods, a pairwise matching in time of the observations was made. Thus, the distance between the buoys were calculated whenever matching observations in time existed for the two buoys, allowing for a maximum difference in time of 30 min between corresponding observations. The ice breakup drifting event was defined to take place when the distance between the buoys was greater than 1 km. The time for the drifting event is marked in Figure 2.5 (black, dotted line) and the position for the buoys at the time of drifting is shown in the right panel of Figure 2.4 (black triangle).

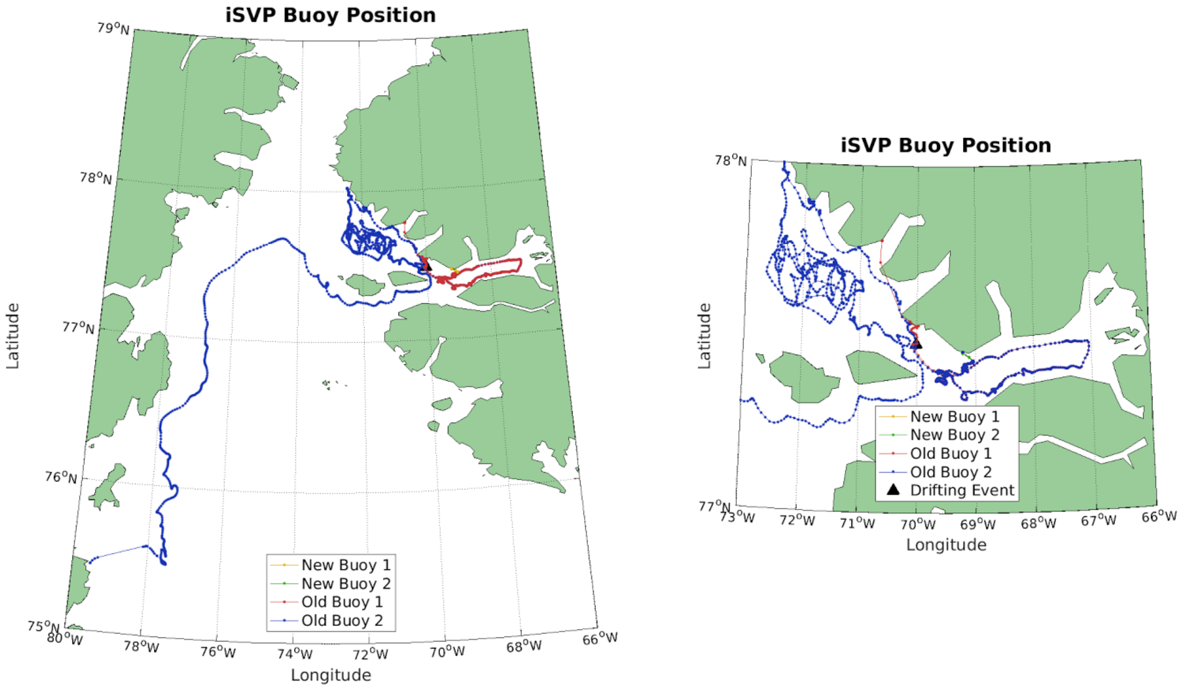


Figure 2.4: Left: position for all four iSVP buoys; new buoy 1 (yellow), new buoy 2 (green), old buoy 1 (red) and old buoy 2 (blue), and right: zoomed in figure of the left panel with the drifting event included (black triangle). All four buoys were deployed at the same position: 77.465°N, 69.218°E, in pairs of two.

The time series of temperature for all four buoys is shown in Figure 2.5. The temperature observations of the two new buoys show a large variability at the start of the observation period, whereas toward the end, the temperature curves become smoother. This is most likely due to the buoys becoming covered in snow. The old buoys, however, show a smooth temperature curve from the beginning, which indicates that they are already covered in snow, with an increase in variability over time. Furthermore, the temperature curves of the two old buoys follow each other closely in the beginning but at the end of June, they begin to deviate more and more from each other. This is because the two old buoys beginning to drift further and further apart, as can be seen in Figure 2.4.

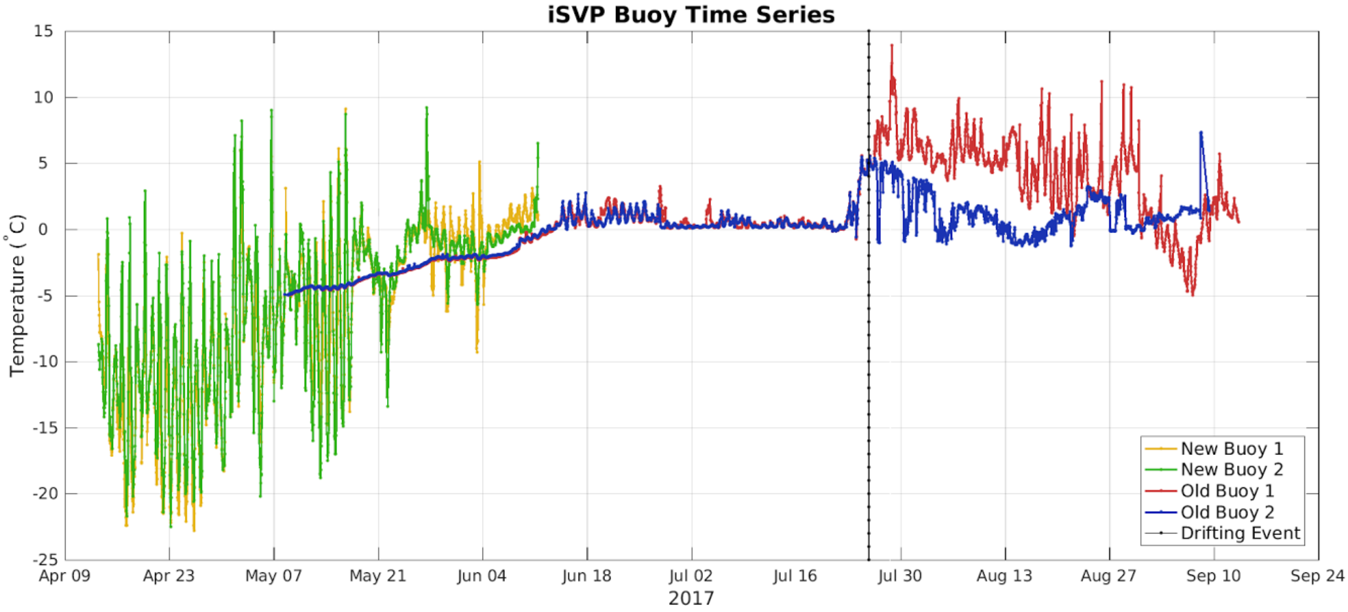


Figure 2.5: Time series of temperature for all four iSVP buoys; new buoy 1 (yellow), new buoy 2 (green), old buoy 1 (red) and old buoy 2 (blue). Marked is also the drifting event for the two old buoys (black, dotted line). Note that no observations are available for old buoy 1 and 2 before start of May, even though they were deployed in end of January. This is because the old buoys are not able to measure temperatures below -5 °C and therefore, all observations at -5 °C have been excluded.

The following part of the report will focus on the time period before the ice broke up.

2.1.2 Buoy vs. buoy inter-comparison

In the buoy vs. buoy comparison, all four buoys are compared to each other. Due to the buoys' different observational periods, three time periods with overlapping observations can be defined (see Figure 2.3); the period in which all four buoys have overlapping observational records in time, 9 May-11 June (hereafter referred to as period NO4); the period in which the two new buoys have overlapping observational records in time (hereafter referred to as period N12), 13 April-11 June; and the period in which the two old buoys have overlapping observational records in time, 9 May-9 September, for which it should be remembered that the buoys begin to drift apart around 25 July (hereafter referred to as period O12 and O12b for exclusion of observations after the drifting event took place).

A pairwise inter-comparison was made between the buoys by matching the buoy observations in time. Hence, the difference in temperature between two buoys was calculated whenever matching observations in time existed for the buoys, allowing for a maximum difference in time of 30 min between corresponding observations. The time series of the temperature difference for all four buoys, the two new buoys and the two old buoys are shown in Figure 2.6, 2.7 and 2.8, respectively.

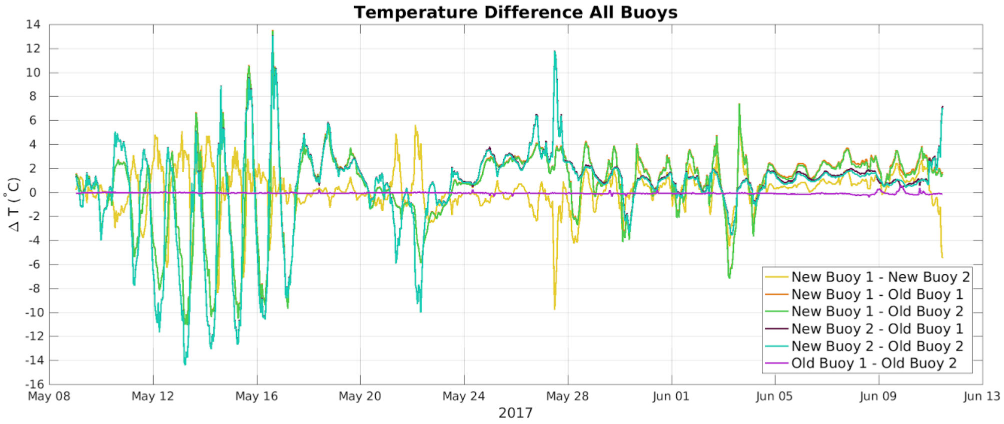


Figure 2.6: Time series of the temperature differences between the buoys for period NO4; new buoy 1-new buoy 2 (yellow), new buoy 1-old buoy 1 (orange), new buoy 1-old buoy 2 (green), new buoy 2-old buoy 1 (brown), new buoy 2-old buoy 2 (turquoise) and old buoy 1-old buoy 2 (magenta).

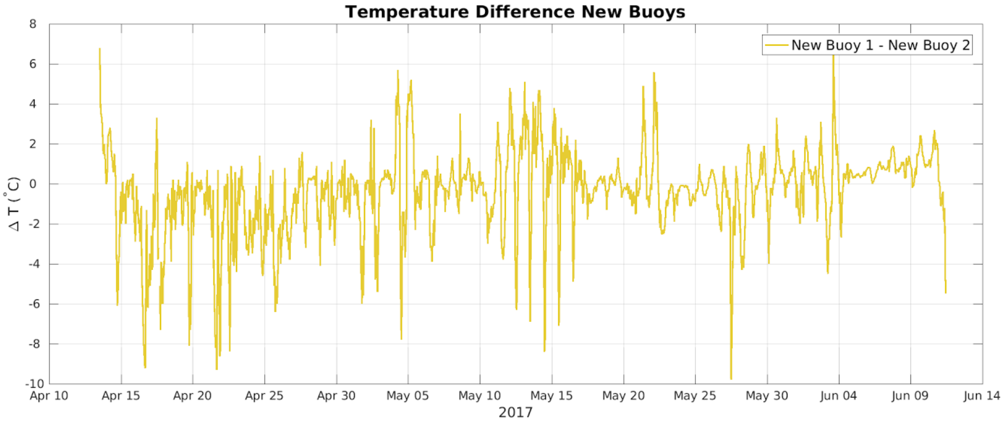


Figure 2.7: Time series of the temperature difference between the two new buoys for period N12.

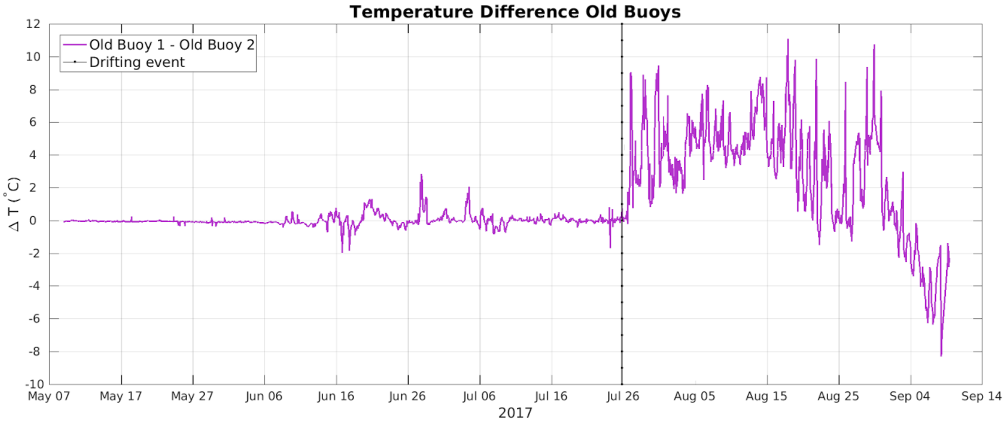


Figure 2.8: Time series of the temperature difference between the two old buoys for period O12 (magenta). The drifting event at 25 July is also marked (black, dotted line).

Pairwise statistics, mean and standard deviation of differences, for all four buoys were calculated for period NO4 (Figure 2.9). The same statistics were also calculated between the new buoys and old buoys for period N12 and O12b, respectively (Table 2.1).

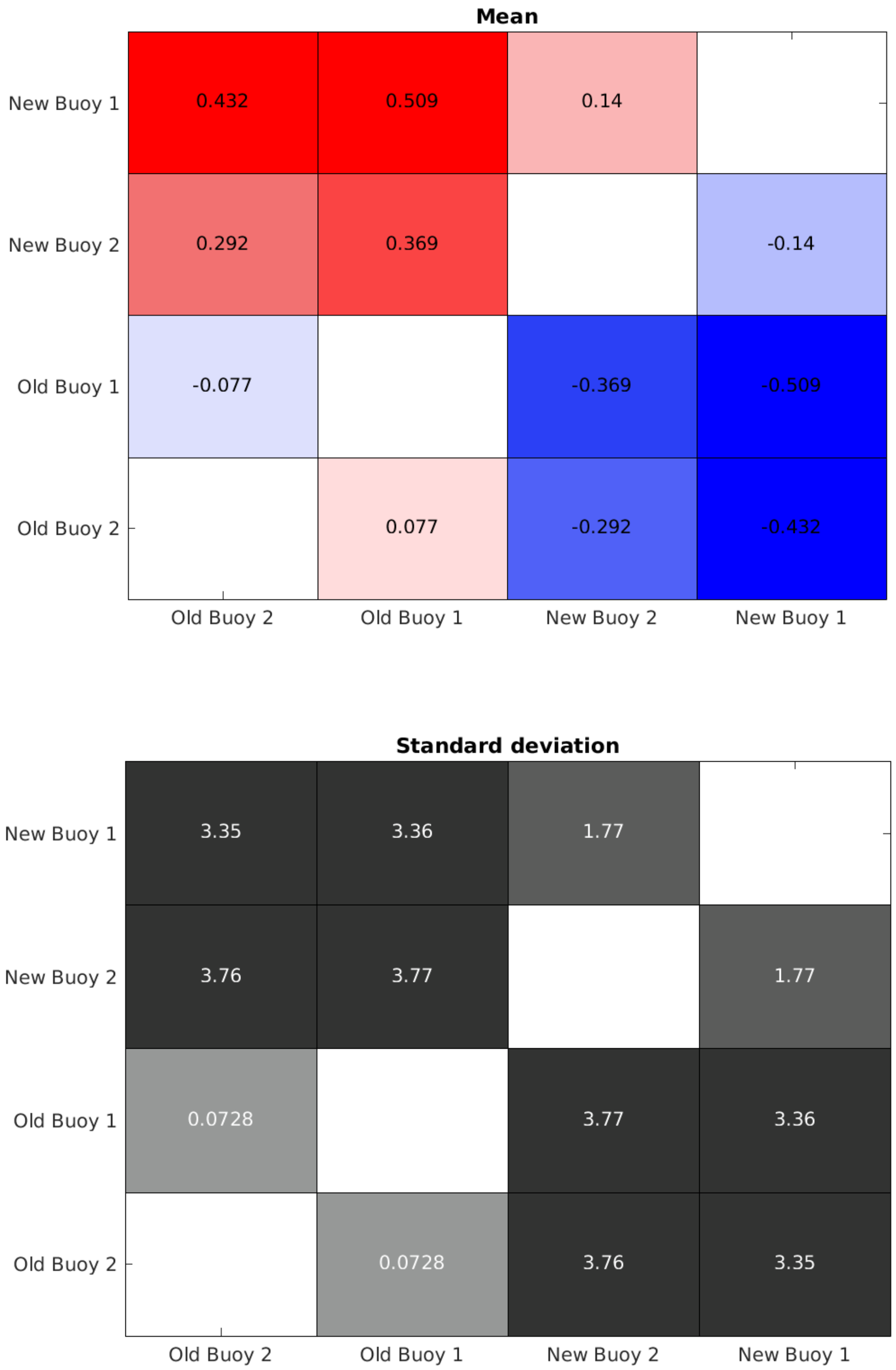


Figure 2.9: Pairwise statistics of all four buoys for period NO4: (top) mean of differences, and (bottom) standard deviation of differences. The signs of the mean of differences should be read from y to x, e.g. new buoy 1 is on average 0.432 °C warmer than old buoy 2.

Table 2.1: Statistics of: new buoys for period N12, and (bottom) old buoys for period O12b.

Buoy Inter-comparisons	Mean difference (°C)	Standard deviation (°C)
New Buoy 1 vs. New Buoy 2	-0.404	2.12
Old Buoy 1 vs. Old Buoy 2	-0.005	0.33

The mean differences between the datasets range between -0.0057-0.51 °C, depending on the respective buoy pairs and time periods; best agreement is found between the two old buoys in period O12b, with a mean difference of -0.0057 °C. This can be compared with the mean difference of the two new buoys of 0.14 °C in period NO4. In contrast, the mean difference between new buoy 1 and old buoy 1 is, with 0.51 °C, highest between all pairs. On the other hand, the standard deviation of the differences is highest between new buoy 2 and the old buoys, with 3.8 °C, whereas best agreement is again found between the two old buoys, this time in period NO4.

2.1.3 iSVP measurement uncertainty

The iSVP buoys are equipped with NTC sensors to measure the temperature. The sensors have been rated to work in temperatures ranging from -55 to +80 °C. Calibration certificates from Metocean show the sensors to have an accuracy of 0.05 °C. Very little degradation in the performance is expected for the sensors for the months considered here (D. M*eldrum, personal communication, 2017*).

2.2 AWS VS. BUOY COMPARISON

In the AWS vs. buoy comparison, the AWS ice surface skin temperature, IST_{skin}, was compared to the temperature observations made by the iSVPs. Due to different observational periods (see Figure 2.10), measurement times (iSVP buoys measure every hour and AWS every 10 min) and data gaps, no common time period was used in the calculations. Instead, the individual buoys were compared with the AWS for the period in which the AWS and buoy have overlapping observational records in time. Comparisons were made between AWS and buoy by matching the observations in time. Thus, the difference in temperature between buoy and AWS were calculated whenever matching observations in time existed, allowing for a maximum difference in time of 15 min between corresponding observations. The time series of the temperature differences are shown in Figure 2.11.

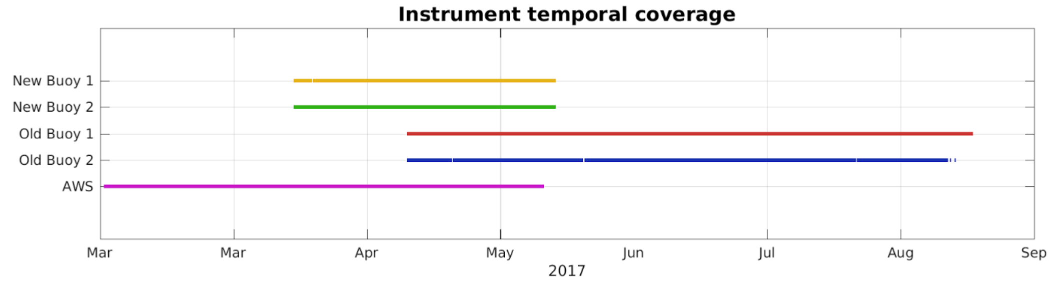


Figure 2.10: Temporal coverage of instruments: new buoy 1 (yellow), new buoy 2 (green), old buoy 1 (red), old buoy 2 (blue) and AWS (magenta).

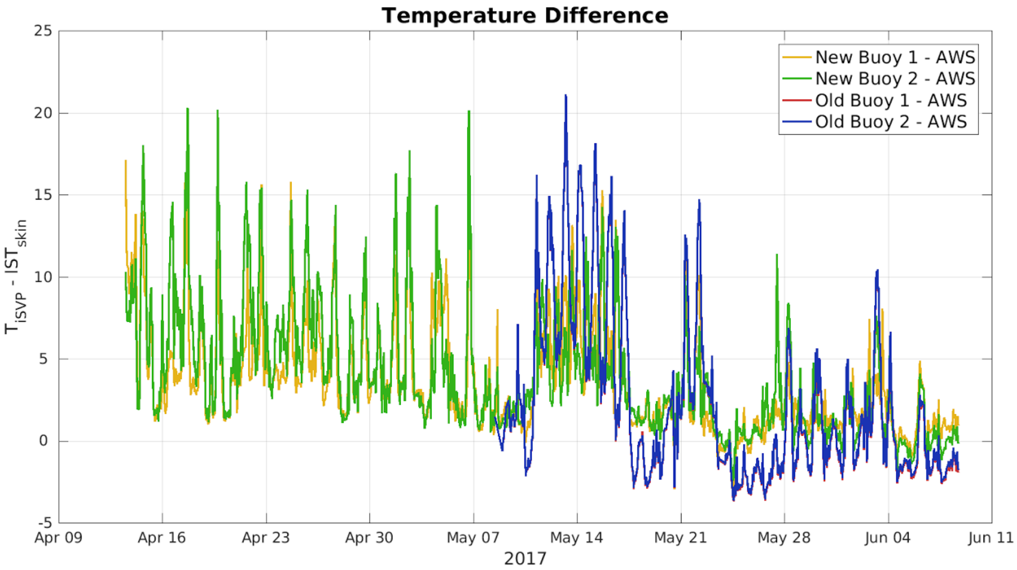


Figure 2.11: Time series of temperature differences between new buoy 1 and AWS (yellow), new buoy 2 and AWS (green), old buoy 1 and AWS (red) and old buoy 2 and AWS (blue).

Statistics, mean and standard deviation of the difference $T_{buoy} - IST_{skin}$, was calculated between all four buoys and the AWS for the period in which the AWS and respective buoy have an overlapping observational record in time (see Figures 2.12 and 2.13). The statistics were calculated for three different ranges of data; all data, daytime data (defined as a solar zenith angle < 90°) and nighttime data (defined as a solar zenith angle ≥ 90°).

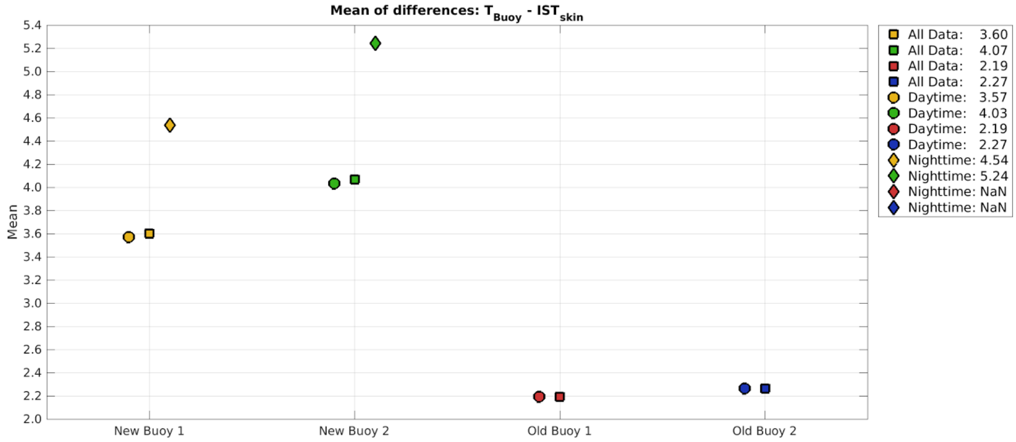


Figure 2.12: Mean of temperature difference between new buoy 1 and AWS (yellow), new buoy 2 and AWS (green), old buoy 1 and AWS (red) and old buoy 2 and AWS (blue) for the period in which the AWS and the individual buoys have overlapping observational records. Squares represent all data, circles daytime data and diamonds nighttime data.

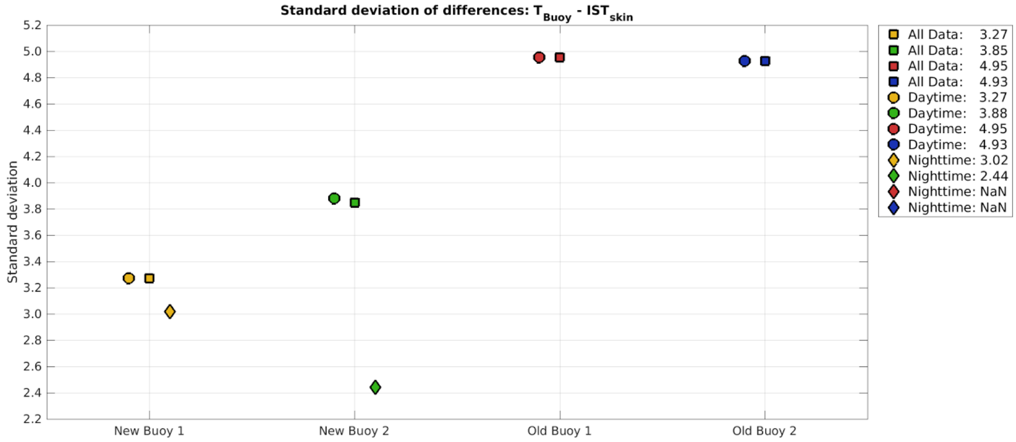


Figure 2.13: Standard deviations of temperature difference between new buoy 1 and AWS (yellow), new buoy 2 and AWS (green), old buoy 1 and AWS (red) and old buoy 2 and AWS (blue) for the period in which the AWS and the individual buoys have overlapping observational records. Squares represent all data, circles daytime data and diamonds nighttime data.

It should be noted that no nighttime observations are available for the old buoys, meaning that the category “all data” only includes daytime observations for the two old buoys. For the new buoys, on the other hand, both daytime and nighttime observations are available. The mean difference between the datasets range between 2.2 and 5.2 °C, i.e. they are all warmer than the AWS, depending on the buoy and type of data used. The best agreement is found between the AWS and old buoy 1 and 2, whereas the mean difference between new buoy 2 and the AWS, with 5.2 °C, is the highest. In addition, by comparing daytime and nighttime data for the two old buoys, it can be seen that the buoys perform better during daytime. The standard deviation of difference, on the contrary, is highest between the AWS and the old buoys, whereas lowest variability, with 2.4 °C, is found for nighttime data for new buoy 1.

2.3 TSKIN VERSUS T1M AND T2M

The effect of vertical sampling was investigated by comparing the temperature in one meter, T_{1m}, and the temperature in two meters, T_{2m}, with the skin temperature. For this comparison AWS data were used. Figure 2.14 shows the time series of the temperature difference $T_{1m} - IST_{skin}$ and $T_{2m} - IST_{skin}$, whereas Figure 2.15 shows the mean and standard deviation of the differences. The best agreement for the mean of differences is found between the one meter temperature and the skin temperature, with 1.30 °C, and likewise the best agreement for the standard deviation of differences is found between the one meter temperature and the skin temperature, with 1.34 °C.

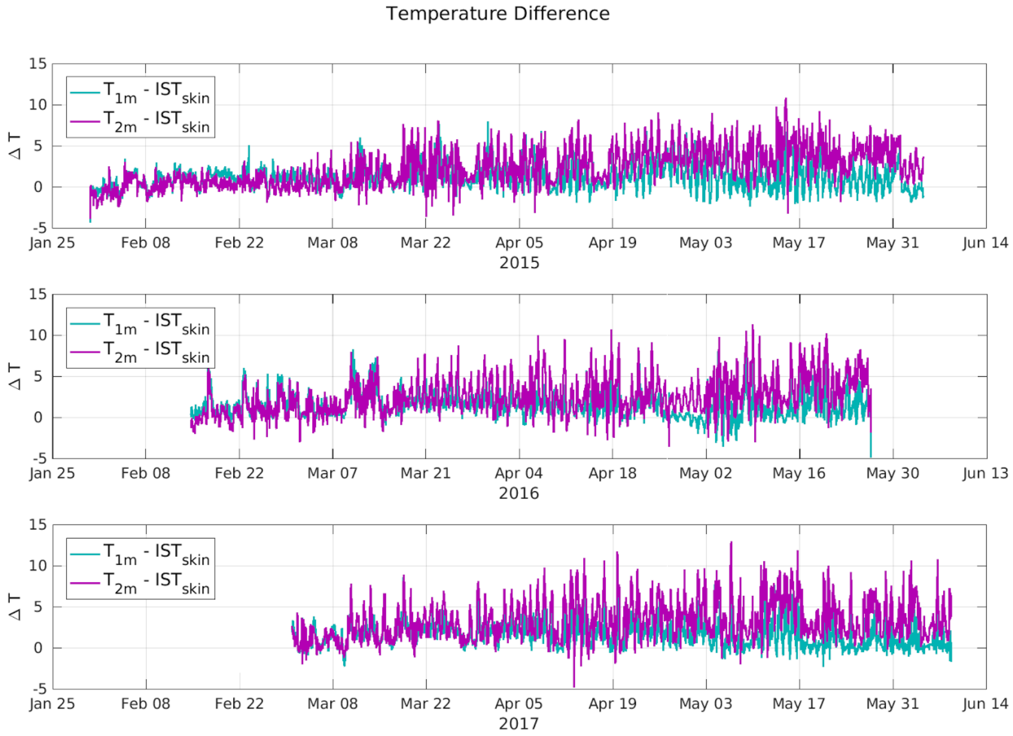


Figure 2.14: Time series of the difference between the temperature in one meter (T_{1m}) and IST_{skin} (blue) and the temperature in two meters (T_{2m}) and IST_{skin} (green) for the Qaanaaq AWS for 2015 (top), 2016 (mid) and 2017 (bottom).

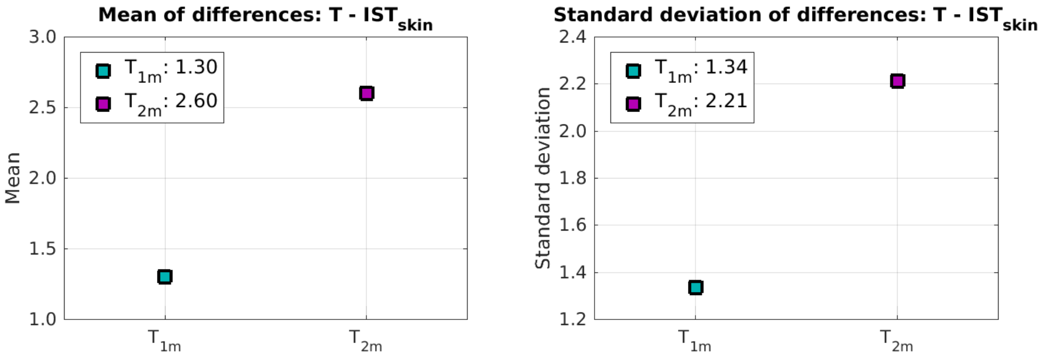


Figure 2.15: Mean (left) and standard deviations (right) of differences $T_{1m} - IST_{skin}$ (cyan) and $T_{2m} - IST_{skin}$ (magenta).

The mean and standard deviation of the difference $T_{2m} - IST_{skin}$ and the distribution of observations, as well as the number of observations, was calculated as a function of the wind speed. The results can be seen in Figure 2.16, where a significant wind speed dependence is evident.

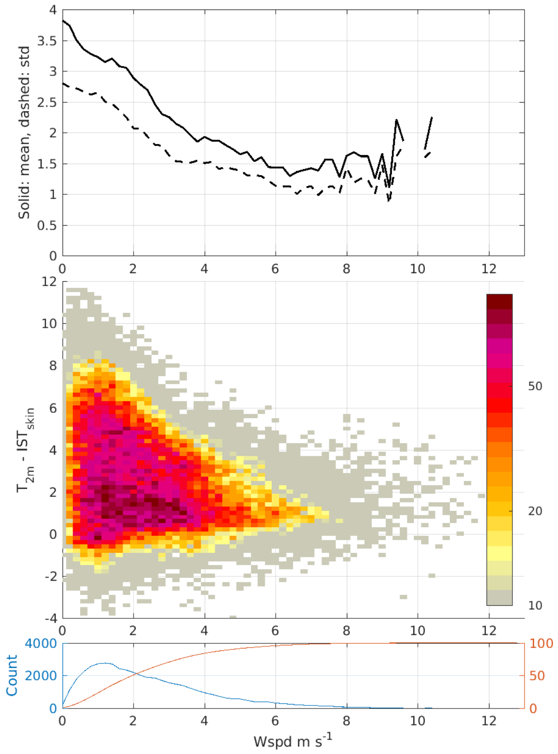


Figure 2.16: 3-panel dependency plot of the difference $T_{2m} - IST_{skin}$ as a function of wind speed. Top panel shows the mean (solid) and standard deviation (dashed) of differences for each bin, the heat plot in the mid panel shows the distribution of observations, and the bottom panel shows the number of observations (blue) and the cumulative number of observations (orange) in each bin.

2.4 TEMPORAL SAMPLING

The effect of a temporal difference between satellite and in situ observations was investigated by comparing the skin surface temperature to itself, shifted in time, T_{shift}. For this, observations were used from the Qaanaaq AWS, the ARM observations from Barrow, Oliktok and Atqasuk. In addition, observations from the Tara experiment were used. The skin temperature observational records were shifted in time ± 3 hours, in steps of 1 minute for the ARM data and in steps of 10 minutes for the AWS and Tara data. Mean and standard deviation of the $T_{shift} - T_{skin}$ difference were calculated for each shift in time. The calculations were performed for daytime and nighttime data separately. Figure 2.17 shows the results of the temporal sampling investigation.

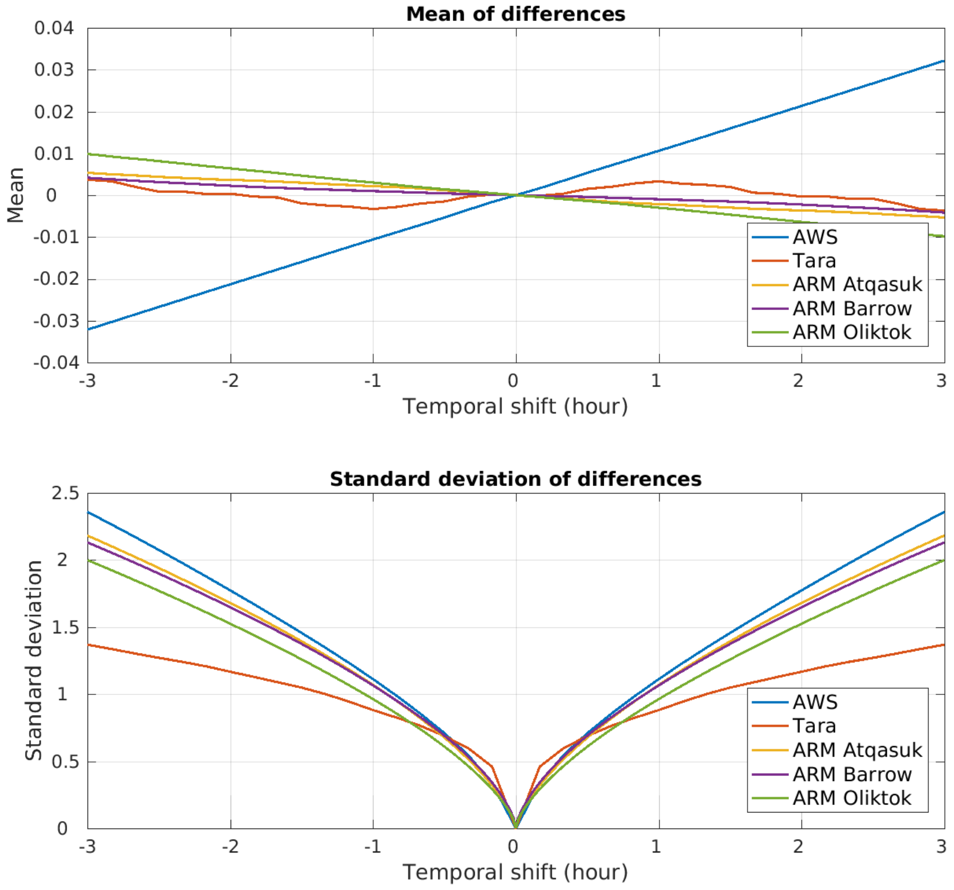


Figure 2.17: (top) Mean and (bottom) standard deviation of differences $T_{shift} - T_{skin}$ for both daytime and nighttime data vs. the temporal shift (h) for AWS data (blue), Tara data (orange), ARM Atqasuk data (yellow), ARM Barrow data (purple) and ARM Oliktok data (green).

2.5 SPATIAL SAMPLING EFFECTS

The temperature variability associated with the satellite observing a spatially averaged footprint versus a point measurement was discussed and assessed in Høyer et al., 2017. The table is shown below:

Table 2.2: Statistics for the spatial variability experiment. See Høyer et al., 2017 for more information.

	Nob s	Distance (km)	Duration	Stdv (°C)	Bias to DMI-AWS IR120 (°C)	Spatial stdv (°C)
Part 1 (Apr-02)	718	4.08	00:59:45	0.69	-0.01	0.25
Part 2 (Apr-03)	709	3.04	00:59:00	0.42	0.50	0.12

The table shows that the estimated variability related to the spatial variability of IST is from 0.12 to 0.25 °C.

2.6 UNCERTAINTY BUDGET

When the satellite and in situ observations are compared in order to validate the satellite algorithms, it is evident from the sections above, that many other factors introduce differences. If we assume all the components to have a Gaussian distribution and not to be correlated, the satellite versus in situ difference is given as:

$$\sigma_{SAT-In\,situ} = \sqrt{\mu^2_{SAT} + \mu^2_{In\,situ} + \mu^2_{\Delta x} + \mu^2_{\Delta t} + \mu^2_{\Delta z}}$$

Where μ_{SAT} and $\mu_{In\,situ}$ are the uncertainties on the satellite observations and in situ observations, respectively. $\mu_{\Delta x}$, $\mu_{\Delta t}$ and $\mu_{\Delta z}$ are the sampling contributions introduced by the geophysical variability associated with point versus footprint (Δx), difference between satellite and observations time (Δt) and the difference between the vertical level of observations (Δz). With the work performed within FRM4STS, we can now estimate the sampling contributions, based on the Qaanaaq observations.

Several assumptions have been made when constructing Table 2.3. First of all, these numbers are based on average values derived from the Qaanaaq observations. There are obvious daily and seasonal dependencies that have not been considered here. The most prominent dependency is the wind dependency related to the IST_{skin} versus T_{2m} difference. In addition, there is a small day/night dependency in the temporal sampling error.

Table 2,3: Estimates for components in the uncertainty budget, when comparing the satellite versus in situ observations.

$\boldsymbol{\Delta x}$ (km)	$\boldsymbol{\Delta t}$ (min)	$\boldsymbol{\Delta z}$ (m)	$\boldsymbol{\mu_{in\,situ}}$ (°C)	$\boldsymbol{\mu_{\Delta x}}$ (°C)	$\boldsymbol{\mu_{\Delta t}}$ (°C)	$\boldsymbol{\mu_{\Delta z}}$ (°C)	$\sqrt{\mu^2_{In\,situ} + \mu^2_{\Delta x} + \mu^2_{\Delta t} + \mu^2_{\Delta z}}$ (°C)
1.0	**10**	$\boldsymbol{IST_{skin}}$	**0.2**	**0.12-0.25**	**0.34**	**0**	**0.41-0.47**
1.0	**30**	$\boldsymbol{IST_{skin}}$	**0.2**	**0.12-0.25**	**0.71**	**0**	**0.75-0.78**
1.0	**60**	$\boldsymbol{IST_{skin}}$	**0.2**	**0.12-0.25**	**1.11**	**0**	**1.13-1.16**
1.0	**10**	$\boldsymbol{T_{2m}}$	**0.05**	**0.12-0.25**	**0.34**	**1.45 - 2.38**	**1.49-2.42**
1.0	**30**	$\boldsymbol{T_{2m}}$	**0.05**	**0.12-0.25**	**0.71**	**1.45 - 2.38**	**1.62-2.50**
1.0	**60**	$\boldsymbol{T_{2m}}$	**0.05**	**0.12-0.25**	**1.11**	**1.45 - 2.38**	**1.83-2.64**
1.0	**10**	$\boldsymbol{T_{buoy}}$	**0.05**	**0.12-0.25**	**0.34**	**3.27 - 4.95**	**3.29-4.97**
1.0	**30**	$\boldsymbol{T_{buoy}}$	**0.05**	**0.12-0.25**	**0.71**	**3.27 - 4.95**	**3.35-5.01**
1.0	**60**	$\boldsymbol{T_{buoy}}$	**0.05**	**0.12-0.25**	**1.11**	**3.27 - 4.95**	**3.46-5.08**

In addition, some specific assumptions have been made in filling in table 2.2. These include:

- The radiometric uncertainty is estimated to 0.2 °C (see Høyer et al., 2017)
- The iSVP buoy sensor uncertainty is estimated to 0.05 °C, according to specifications
- The AWS pt100 sensor uncertainty is estimated to 0.05 °C for T_{2m}
- The range for $\boldsymbol{\mu_{\Delta z}}$ for T_{2m} arise from two intervals: wspd ≥ 3 m/s and wspd < 3 m/s, where the $\boldsymbol{\mu_{in\,situ}}$ contributions have been subtracted
- The range for $\mu_{\Delta z}$ for T_{buoy} are derived from the four iSVP buoys compared against the IST_{skin} from the AWS where the $\boldsymbol{\mu_{in\,situ}}$ have been subtracted

The table above demonstrates the challenges involved in calibrating and validating satellite IST products. When using traditional observations, such as T_{2m} and T_{buoy} for satellite versus in situ matchups, the cumulated effects of the components not associated with uncertainties in the satellite IST retrievals can reach more than 5 °C. It is also clear from the table that the most suitable FRM observations for satellite validation are traceable radiometric observations from an FRM radiometer measuring with at subhourly intervals (e.g. 1 minute).

3 AUTOMATIC PROCEDURES AND METHODS TO QUALITY CONTROL IN SITU OBSERVATIONS

The previous section demonstrated that the temperatures reported from iSVP buoy observations placed on the sea ice may differ several degrees from the skin temperature observed from satellites. This is unfortunate, since the majority of automatic reported observations obtained through the global communications network (GTS) consist of these observations. In addition, the extreme environmental conditions in the polar regions result in many observations with poor quality and obvious bad observations. In order to increase the usage of the GTS iSVP buoy observations for satellite validation, automatic procedures and a software package have therefore been developed for quality controlling the NRT IST observations. The results are described in this section on a test data set obtained from the GTS.

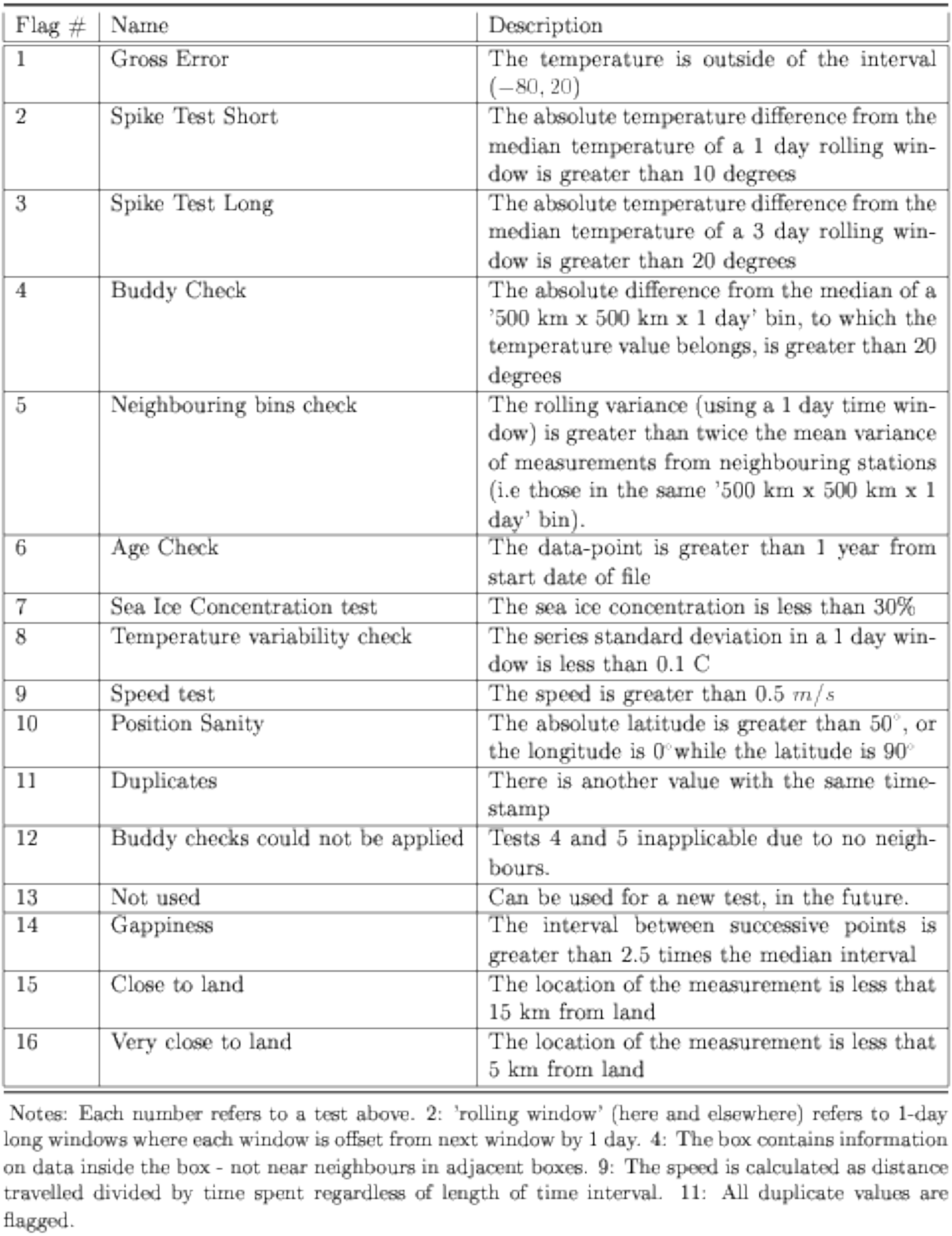
3.1 INTRODUCTION

A quality check of individual data points from ice buoys has to be performed. The software presented here reads each indicated input file and produces output that helps identify data points that fail to pass 16 quality checks. Tests are generally based on how good and believable the individual values appear to be by themselves; compared to the rest of the time series; and when compared to neighboring observations in time and space.

3.2 QUALITY TESTS

The 16 quality tests are individually described in Table 1. The software runs through all selected data files, and for each file all points are checked in various ways. 16 flags are set on the basis of the checks and written into a 2-byte word. In Table 1, the conditions for setting the flag are described. The representation of the 16 flag values is binary (or ASCII text additionally, if selected by the user) - two bytes must be read and interpreted by the user's software. Test 1 is encoded in bit one, while test 16 is encoded in bit 16 -- so that test 1 is the least significant bit while test 16 is the most significant bit: flag 1 is encoded as the factor of 2^1 while flag 16 is the factor on 2^{16}.

Table 3.1: Automatic quality control tests, developed within FRM4ST for IST buoy observations.



Flag #	Name	Description
1	Gross Error	The temperature is outside of the interval $(-80, 20)$
2	Spike Test Short	The absolute temperature difference from the median temperature of a 1 day rolling window is greater than 10 degrees
3	Spike Test Long	The absolute temperature difference from the median temperature of a 3 day rolling window is greater than 20 degrees
4	Buddy Check	The absolute difference from the median of a '500 km x 500 km x 1 day' bin, to which the temperature value belongs, is greater than 20 degrees
5	Neighbouring bins check	The rolling variance (using a 1 day time window) is greater than twice the mean variance of measurements from neighbouring stations (i.e those in the same '500 km x 500 km x 1 day' bin).
6	Age Check	The data-point is greater than 1 year from start date of file
7	Sea Ice Concentration test	The sea ice concentration is less than 30%
8	Temperature variability check	The series standard deviation in a 1 day window is less than 0.1 C
9	Speed test	The speed is greater than 0.5 m/s
10	Position Sanity	The absolute latitude is greater than 50°, or the longitude is 0°while the latitude is 90°
11	Duplicates	There is another value with the same timestamp
12	Buddy checks could not be applied	Tests 4 and 5 inapplicable due to no neighbours.
13	Not used	Can be used for a new test, in the future.
14	Gappiness	The interval between successive points is greater than 2.5 times the median interval
15	Close to land	The location of the measurement is less that 15 km from land
16	Very close to land	The location of the measurement is less that 5 km from land

Notes: Each number refers to a test above. 2: 'rolling window' (here and elsewhere) refers to 1-day long windows where each window is offset from next window by 1 day. 4: The box contains information on data inside the box - not near neighbours in adjacent boxes. 9: The speed is calculated as distance travelled divided by time spent regardless of length of time interval. 11: All duplicate values are flagged.

3.2 RUNNING THE QC SOFTWARE

The software code for the automatic QC is freely available upon request. The following sections will aid the users in using and interpreting the results from the software. To run the QC software, ensure that python 2.7 and all required libraries are installed -- see the 'README.md' file provided with the code.

Run code with

```
python write_flags.py 'inputdir/' 'outdir/' vartotest [speed quality_array]
```

where the argument *'vartotest'* indicates whether the air temperature 'TA' or the skin temperature 'IT' are to be tested on # allowed values for this argument are 'TA' or 'IT', the argument *'inputdir/*' is the directory with input files. The configuration file 'con#g.py' further helps specify the files to be read. The *'outdir/*' argument is the directory (must exist) for the output files. Output is in *netcdf* format, and is in the form of the original input file but with the result of the quality tests given in the file. The single quotes on the file and directory arguments are necessary. The first item 'speed' is given if you want the station speed written to the output file. '*quality_array'* is used when the flags are to be shown as an array of 1's and 0's in the output file, for visual inspection, for instance.

3.2.1 Usage notes

The first time the code is run, an array of aggregated data is written out to a file, and is used for quality checks that involve neighboring data points. The aggregated file uses, at the moment, bins or boxes that are 500x500 km and 1 day long. Once this file (named aggregated.h5) is written out it is reused in subsequent runs. For instance, if the full set of some files are used the first time, then on wanting to test a single file later it will be tested against that aggregation. If you want to extend the set of data used in the aggregate file, remove it, and run the code on a fuller set of data and the file will be rewritten with the new information.

3.2.2 Graphical display of results

To aid in the visualization of the flags on each data series maps and plots of all the stations represented by the input files are also produced. Example output is shown in Figures 3.1 and 3.2 for observations from an arbitrary buoy.

3.2.3 Test of quality improvement

We finally tested how large the improvement in data quality is if flagged data-points are removed. We did this by considering the difference in RMSE with and without removal of flagged data. All flags were employed, although not all flags ‘fired’ equally, as seen in Figure 3.3. The improvement was then quantified in a histogram, shown in Figure 3.4, showing the percentage decrease in RMSE as a positive number, indicating quality-increase. The majority of data have undergone an improvement. Most seem to improve up to 30% or so while a substantial number improve from 30 to 90%. A minority actually have undergone a decrease in quality, which is due to removal of data points and the dependence of RMSE on number of data-points in the denominator.

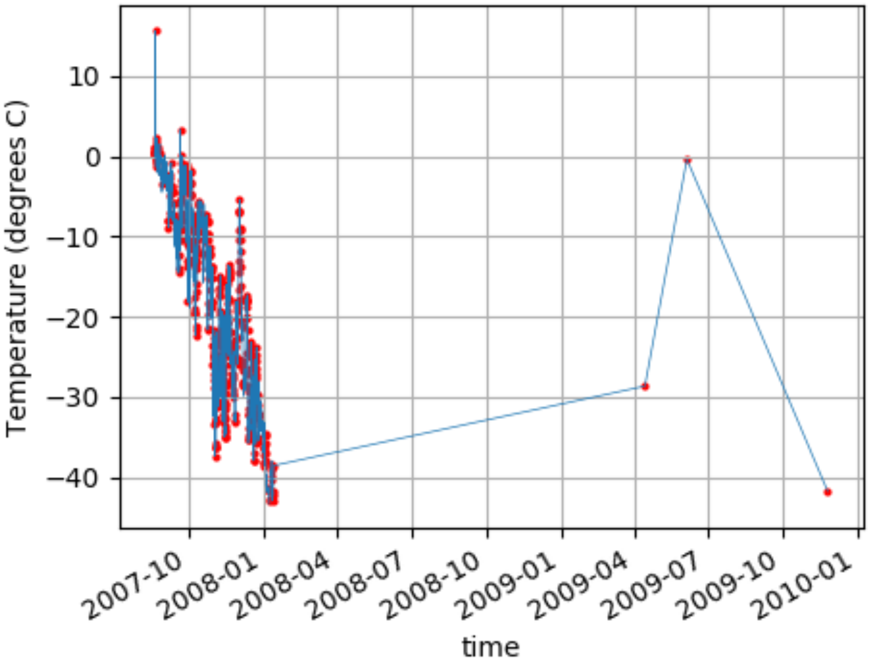


Figure 3.1: Buoy observations. Red dots indicate values flagged by one of the 16 tests.

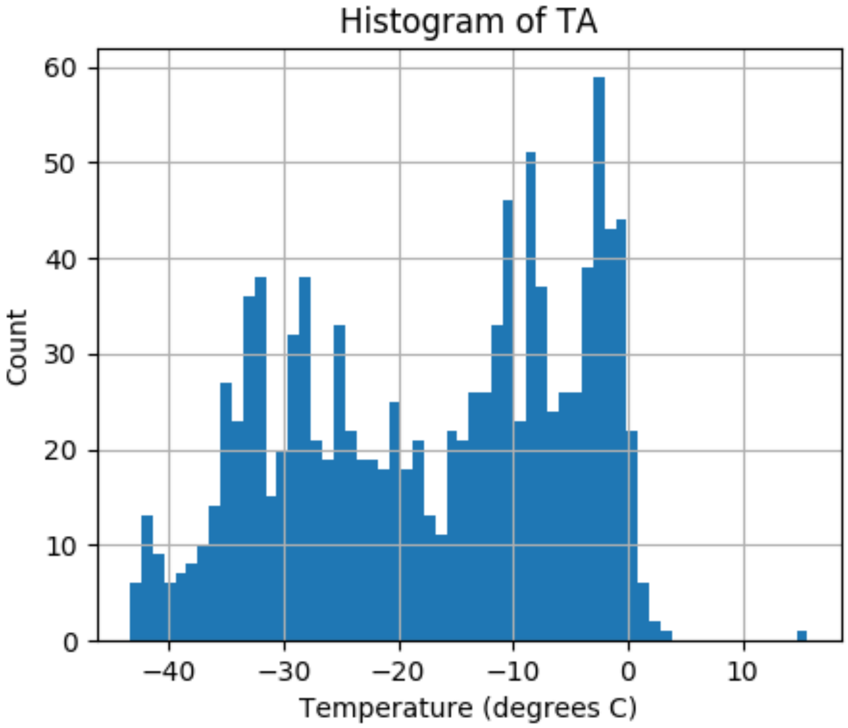


Figure 3.2: Histogram of observed temperature observations.

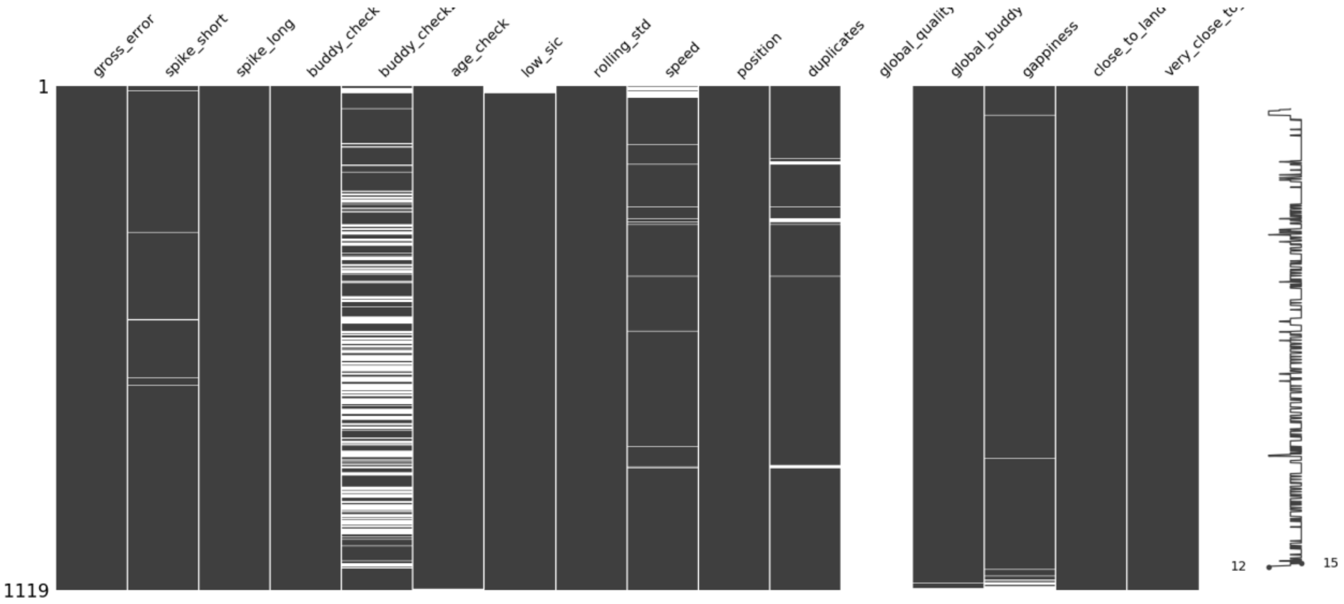


Figure 3.3: The 16 tests summarized. For each column representing a test, a white stripe indicates the flag has been set.

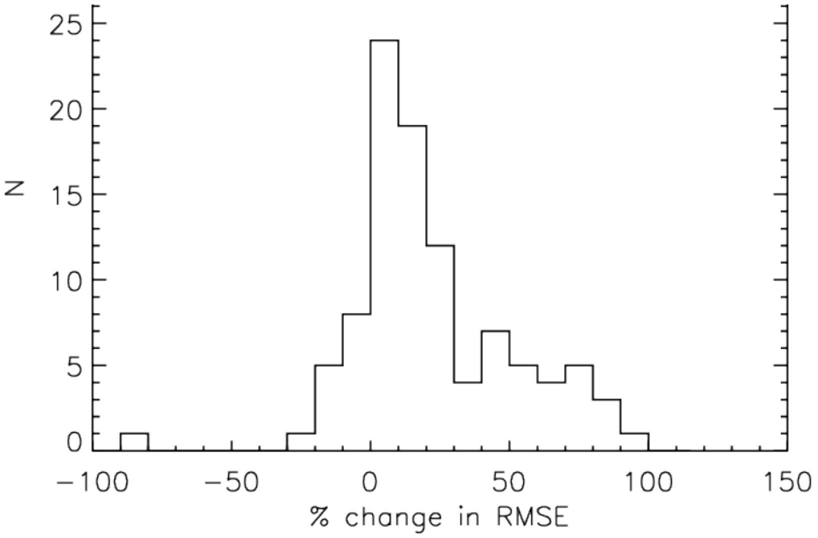


Figure 3.4: Percentage improvement in RMSE when flagging bad data using all flags described in this report. The change in RMSE is positive if quality is improved.

4 CONCLUSIONS AND RECOMMENDATIONS

The results from this options shows that using iSVP as a FRM for validating satellite ice surface temperature measurements over sea ice bears several complications. The accuracy of the sensors is adequate to be used for validation. Through data analysis on iSVP buoys set out on the sea ice off Qaanaaq, western Greenland, each of the components in the uncertainty budget is assessed, when satellite and iSVP buoys observations are validated. A table is provided for typical validation conditions, showing sampling differences to range from 0.36 °C for radiometric observations to more than 5 °C, when comparing satellite and iSVP buoys analysis. The larges effects arise from the different sampling components, where the vertical displacement of the in situ observations from the skin into the snow and ice or in the air accounts for the largest contribution.

In addition to the Qaanaaq data analysis, automatic quality control routines have been developed to filter out the observations that do not represent the skin IST. These procedures consist of 15 tests that range from to buddy checks on mean value and variability. Application of these QC tests on a test data set obtained from the GTS leads to improvements in the data quality, but the fundamental problem still exists, namely that the IST skin temperature easily can differ several degrees from temperature observations within a few centimetres of snow in the air. These large discrepancies severely limit the usefulness of the observations for validating and improving upon the satellite IST retrieval algorithms. Several land-based snow and ice radiometric observations are available today from e.g. the ARMS sites but the snow and ice melts during summer and no all-year round radiometers exist today that can be used to monitor and validate the existing satellite IST algorithms.

The lack of FRM radiometric observations over sea ice is limiting the development within the IST algorithms, which for most cases rely upon an algorithm type developed in (Key et al, 1992). It would therefore be very beneficial for all satellite ice produces, if a calibrated all-year round radiometers were deployed.at Summit, Greenland, and on the sea ice in the Arctic.

It is therefore highly recommended that all-year FRM radiometric observations of IST are being performed over a homogeneous permanent ice area, such as Summit, Greenland, for the calibration and validation of satellite IST observations.

5 REFERENCES

Dybkjær, G., Tonboe, R., & Høyer, J. L. (2012). Arctic surface temperatures from Metop AVHRR compared to in situ ocean and land data. *Ocean Science*, *8*(6), 959-970.

Høyer, J. L., Lang, A., Tonboe, R. T., Eastwood, S., Wimmer, W., Dybkjær, G. (2017) Report from Field Inter-Comparison Experiment (FICE) for ice surface temperature, *FRM4STS report*, available at: www.frm4sts.org.

Rigor, I. G., Colony, R. L., & Martin, S. (2000). Variations in surface air temperature observations in the Arctic, 1979–97. *Journal of Climate*, *13*(5), 896-914.