Uncertainty of temperature estimates from fiber-optic distributed temperature sensing: Comparison of calibration methods

Bachelor Thesis

Leon Steinmeier   (1355194)

June 19, 2018

Abstract

This thesis explores the possibilities of data quality enhancement for distributed temperature sensing (DTS) measurements through the use of different calibration techniques and aggregation of measurement data. The data used for the analysis had originally been gathered during two different studies in 2010 ad 2011. Also within the scope of this work the development of a software package for processing and calibration of DTS measurement data took place.

Contents

Contents

List of Figures

List of Tables

1 Introduction

Distributed Temperature Sensing (DTS) as a measurement technique has been around in geoscience for more than a decade now (Hausner et al., 2011). It was first used by the oil and gas industry in the 1980s to monitor pipelines. The ability to cover a measurement transect of multiple kilometers with only one device and a glass fiber was what made this technique the tool of choice for that task. The technique's potential for scientific applications was first tested in the field of hydrology (S. et al., 2006). Since the beginning of this decade it has also been used in boundary layer meteorology to characterize non-turbulent air movement near the surface (Thomas et al., 2012).

Depending on the research question, the quality (i.e. trueness, precision and accuracy) of temperature measurements often needs to be quite high to convey meaningful results (Hausner et al., 2011). While the manufacturers of DTS devices usually equip their products with an automatically applied calibration routine to translate the measured signals into temperature values, the quality of these device-reported temperatures often fails to reach the necessary level. In these cases an own calibration has to be conducted. For that purpose multiple methods exist with varying complexity and requirements to the measurement setup (Hausner et al., 2011; Thomas et al., 2012).

1.1 Research questions

In this thesis two of these calibration methods and the device-calibration will be compared with each other. One of the non-device methods will be a rather simple one that uses the device calibrated temperatures and does an offset and span correction on them. The other one relies on the measurement's raw signal and is considered to provide the highest data quality currently achievable. The comparison aims at quantifying the quality differences between the temperature values resulting from the different calibration methods.

The hypothesis is that in terms of data quality, the device-calibration provides the worst quality of the three, followed by the offset-span calibration on the already device-calibrated temperatures and both of them being topped by the more complex calibration of the raw signal.

Furthermore the influence of aggregation of contiguous data points on the data quality will be investigated. Here the hypothesis is that the data quality increases with the size of the aggregation interval, but only until a certain threshold.

1.2 Theory

To conduct a measurement a light pulse is emitted into an optical fiber. While the pulse travels through the fiber, a fraction of the photons is scattered by irregularities in the material. Because of this some photons are able to travel back through the fiber to the device that emitted the light. Within that device the energy of the returning photons is measured at different wavelengths. While most of the photons return at the same wavelength as the initial pulse as a result of Rayleigh scattering a few of them do so

because of Raman scattering which causes them to be at either lower (Stokes band) or higher (anti-Stokes band) wavelengths than the initial pulse. As the intensity of the so called "Stokes" is less dependent on temperature than the "anti-Stokes' " intensity their ratio can be used to calculate temperature. Together with travel-time information on the laser pulse this temperature signal can then be located on the fiber.

2 Materials and Methods

2.1 Analyzed Data Sets

The measurement data used in the analysis was gathered during two different experiments in 2010 (Thomas et al., 2012) and 2011 (Zeeman et al., 2015). These experiments took place in Oregon, USA from June until November 2010 and from August until October 2011 respectively.

2.1.1 Deployed DTS Devices

The DTS device used in 2010 was a prototype of the Ultima brand produced by Silixa Ltd. (`https://silixa.com/`). In 2011 a newer model of the same brand was used.

Unfortunately manufacturer specifications could not be acquired for any of these instruments. However the hard copy of a data sheet for the 2km variant of the Ultima-S Model which was produced a few years later was made available to the authors of this thesis. That model is similar to the ones used in the afore mentioned experiments. The maximum spatial resolution for this instrument is declared as 12.5 cm and the minimum measurement time as one second. Without any temporal averaging the measurement resolution is given as about 0.28 degrees Celsius near the instrument and about 0.46 degrees Celsius at a distance of 2 km along the fiber. This data refers to the device-calibrated measurements.

For both experiments the devices were configured to measure at a spatial resolution of about 12.7 cm and a temporal resolution of 3 seconds in 2010 and 5 seconds in 2011. Despite that the time series aren't equidistant as the devices sometimes measured at slightly different interval sizes.

2.1.2 Reference Baths

For both experiments water baths of two different temperatures were deployed to provide reference sections for the calibration of fiber temperatures. While the colder baths were both cooled with ice, the warmer bath in 2010 was kept at ambient temperature unlike the warmer bath in 2011 which was actively heated and therefore had a temperature around 24 degrees Celsius. During both experiments the water baths were also constantly mixed by aquarium pumps to prevent the forming of temperature gradients.

2.1.3 Reference Sensors

For the experiment in 2010 two RM Young PT-1000s were used to measure the water baths' temperatures. The PT-1000s' accuracy at zero degrees Celsius is reported as 0.3 degrees Celsius (`http://www.youngusa.com/Manuals/41342-90(c).pdf`).

During the experiment in 2011 the PT100s delivered together with the DTS device were used in both baths. While the specifications of these devices are missing there is still data about the PT100s that are delivered together with the Ultima-S model. Their absolute tolerance is stated as 0.15 degrees Celsius.

2.1.4 Fiber Arrangement

The general setup of a DTS experiment usually routes the fiber through two water baths of different temperatures shortly after the starting point at the DTS device. After that the fiber is arranged in the actual measurement area where at the end it comes back going through the water baths again (figure 1). This results in four reference sections along the fiber: two at the beginning and two at the end.
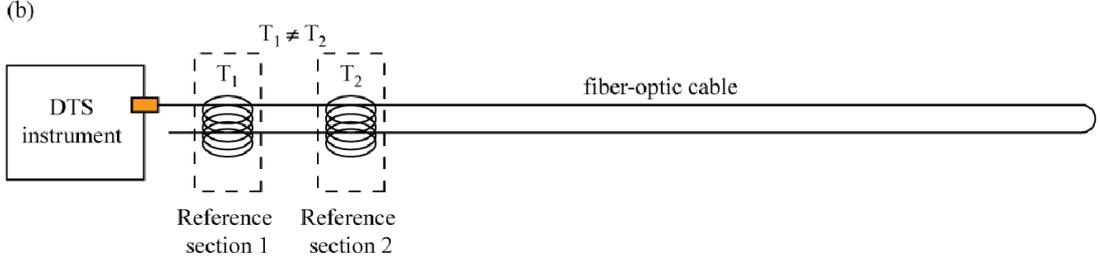
Figure 1: General Setup of a DTS Experiment (Hausner et al., 2011)

Within the baths the fiber was coiled up. In 2011 that coil was arranged horizontally. The fiber coil's orientation in the 2010 experiment is unknown.

Detailed information on bath order and the length of fiber in between and within baths for both experiments is given in table 1. The fiber lengths in between two contiguous baths lies between five and fifteen meters.

Table 1: Length of Fiber Sections [m] and Temperature Category of Both Water Baths

	Beginning of the fiber				End of the fiber	
	DTS device → Bath 1	Bath 1	Bath 2	Environment	Bath 2	Bath 1
2010	150	5.5	3.6	318.4	3.3	1.9
	temperature category →	cold	warm		warm	cold
2011	236	27	25	1507	21	22
	temperature category →	warm	cold		cold	warm

2.1.5 Data Processing and Calibration Software

Within the context of this thesis a software package was developed to aid in the tasks of data processing and calibration. The code was written in C++ to achieve a high processing speed as the previously used Matlab and Python scripts performed unsatisfactorily in that regard.

The package contains functionality to transfer and aggregate the measurement data conveyed by the Ultima devices as per-measurement XML files to files of the netCDF format (`https://www.unidata.ucar.edu/software/netcdf/docs/`). Multiple XML files

selected by the user can be aggregated into a single netCDF file containing measurement and meta data. Naming and structure of the meta data was informed by the netCDF CF meta data conventions (`http://cfconventions.org/`).

Furthermore the program can perform the two different calibrations mentioned in 2.2 on the measurement values in the XML files. To solve the linear equation system of the single ended calibration the LAPACK++ library (`https://sourceforge.net/projects/lapackpp/`) and it's function LaLinearSolve is used.

Additionally, the user can pass the software fiber mapping data to have it being contained in the netCDF files' meta data for a more convenient analysis later on.

The interface between user and software is a configuration file that contains text in the form of comments and "tags". These tags are strings that values are assigned to. Tags and their values are read in on startup and define key values (e.g. which meta data should be extracted from the XML files or which calibration should be performed, if any at all) that the program runs on.

Performance tests for the software were run on a note book with an Intel Core i5-6267U CPU under the Ubuntu 18.04 operating system while standard office software was running in the background.

For the XML data files from the 2010 experiment where each file contained data of around 690 meters of fiber, about 15 kilometers of data or 21 XML files per second could be processed while doing both calibrations. The same numbers for the 2011 experiment where every file contained around 1990 meters of data, resulted in around 46 kilometers of data or 23 files per second.

A class diagram depicting the program's components can be seen in figure 2.
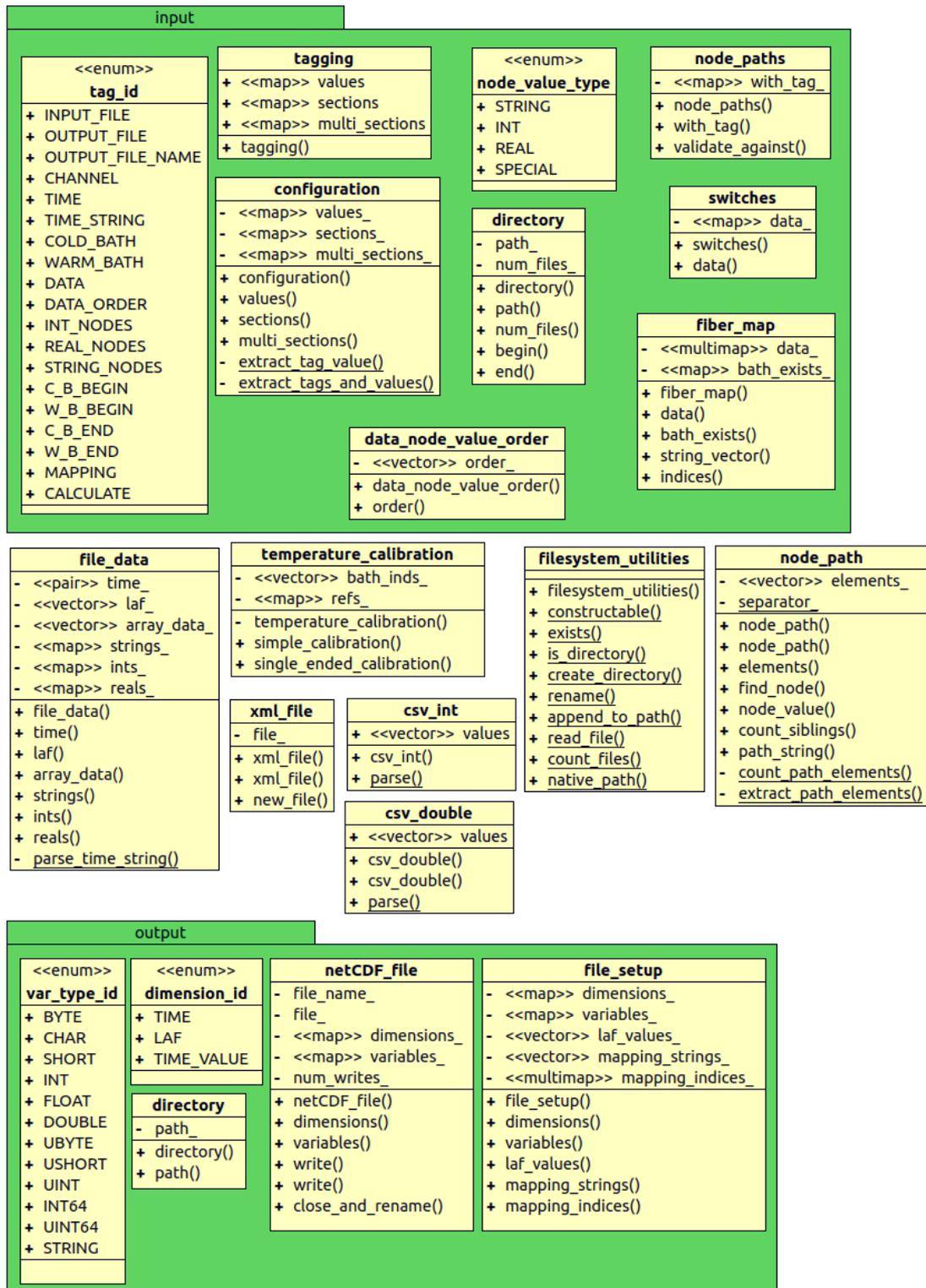
Figure 2: Class Diagram of Program Components

2.2 Calibration Methods

Every calibration was conducted four times while always using only three of the four reference bath sections to reserve the fourth section for validation purposes.

2.2.1 Device Calibration

Of the three calibration methods compared in this thesis the device calibration requires the least effort as it is conducted completely automatically by the device using an internal reference. Therefore all four bath sections on the fiber can be used for the validation of this calibration's results without constraint.

2.2.2 Offset-Span Calibration

The second calibration works similar to a two-point calibration and is applied to each measurement of device-calibrated temperatures along the fiber. Two water baths of different temperatures on opposing ends of the fiber are used as the two reference points.

For the first part of the offset correction the difference between one of those baths' reference temperature and it's mean fiber temperature is used to shift all values of the measurement along the temperature dimension such that the difference at that bath is reduced to zero Kelvin.

The span correction includes calculating the same difference for the second bath and dividing it by one plus the number of values in between both bath sections. A multiple of that fraction is added to every value after the first bath (excluding the second bath) and subtracted from every value before the first bath. The size of the multiple is calculated as one plus the number of values along the fiber (excluding those of the second bath) in between the respective value and the first calibration bath.

For the second part of the offset correction the values of the second bath section are shifted to have their mean match with that bath's reference temperature.

Because of the order of the bath sections of the two analyzed experiments (table 1) only the "inner" two bath sections on the fiber could lie in between calibration sections. Still the "outer" two were also used for validation as they were only a few decameters away from the calibration sections next to them.

2.2.3 Single Ended Calibration

This calibration uses the measured intensities of the Stokes and anti-Stokes and three reference sections with at least two of them having different temperatures and being on opposing ends of the fiber. The calculation is done using the second algorithm described in (Hausner et al., 2011) that employs explicitly calculated parameters and reference sections.

For the validation all bath sections were used even so the "outer" ones could not be positioned in between calibration sections similar to the offset-span calibration.

2.3 Data Selection

Because it appeared unphysical, data from 2010 between 3:00 and 3:20 am was excluded together with data at times where the warm bath's reference temperature dropped below 10 degrees Celsius after 7:00 am (figures 3 and 4).
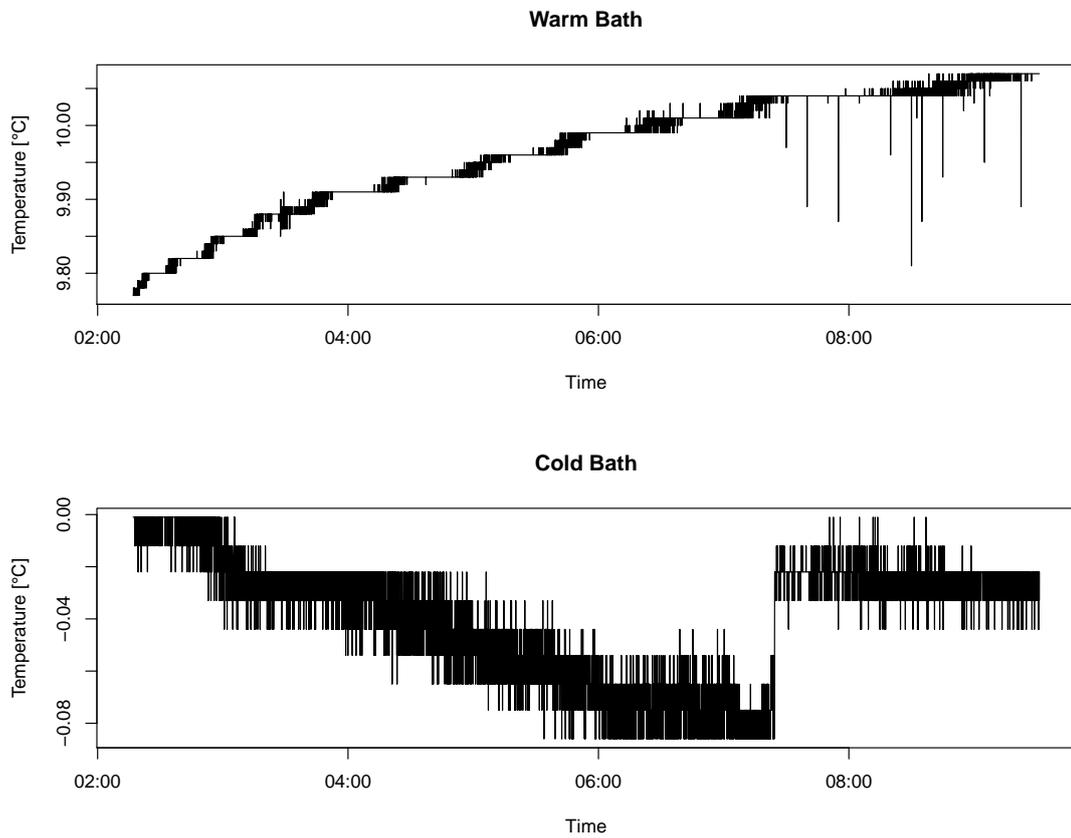
Figure 3: Unfiltered Reference Temperatures (2010)
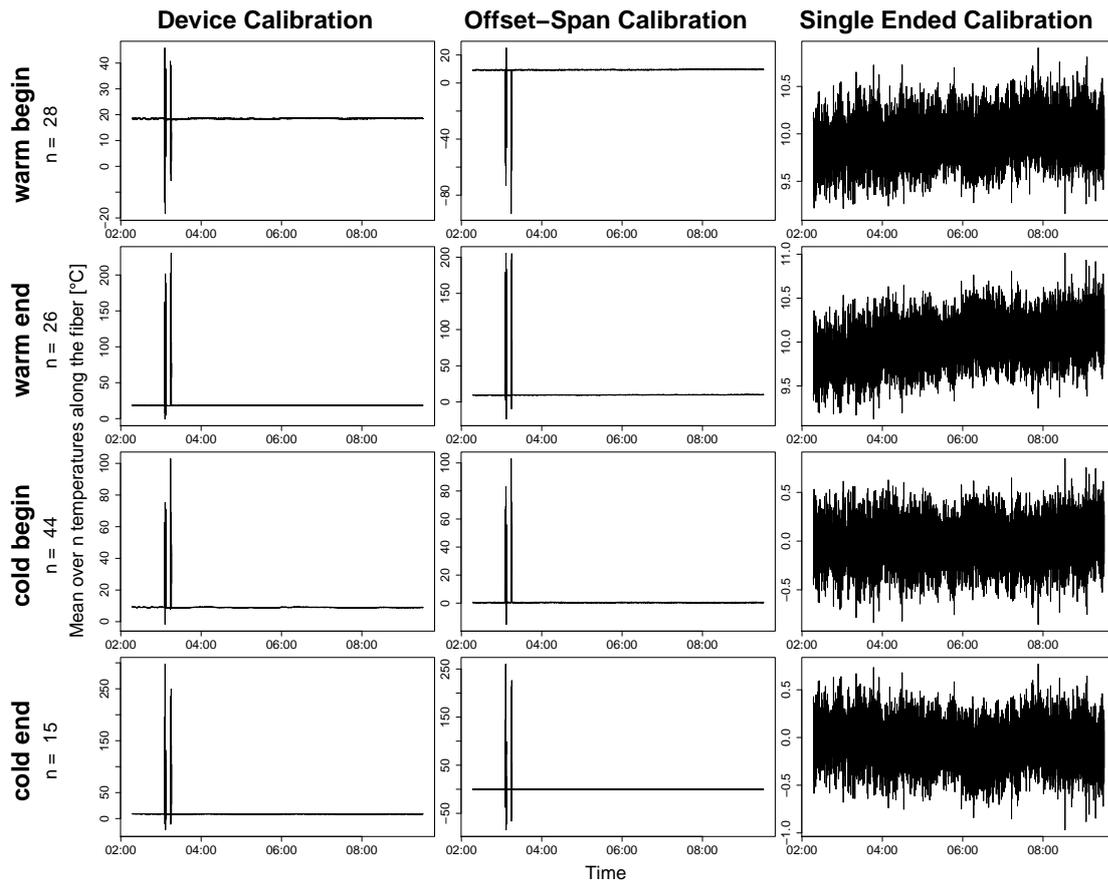
Figure 4: Unfiltered Temperatures (2010)

2.4 Quality Metrics

To describe the different aspects of data quality the terminology as proposed in ISO 5725 (ISO 5725-4:1994) and the international vocabulary of meteorology (De Bièvre, 2012) is used, i.e. trueness as a measure of distance between the mean of measured values and a reference value, precision as an indicator for the deviation of measurement values around their own mean and accuracy as a combination of both (see figure 5).
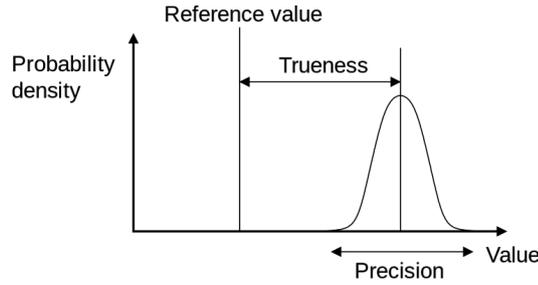
Figure 5: Quality Metric Concepts ; source: `https://en.wikipedia.org/wiki/File:Accuracy_(trueness_and_precision).svg`

2.4.1 Trueness

The formula used to calculate the trueness is

$$T = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - r_i) \tag{1}$$

with T being the trueness, x the measured values and r the reference values. This metric is also commonly known as "bias" in the natural sciences.

2.4.2 Precision

The formula used to calculate the precision is that of the mean absolute deviation (MAD)

$$P = \frac{1}{n} \cdot \sum_{i=1}^{n} (|x_i - \frac{1}{n} \cdot \sum^{n} x|) \tag{2}$$

with P being the precision and x the measured values.

2.4.3 Accuracy

For calculation of the accuracy the formula of the mean absolute error (MAE) was used

$$A = \frac{1}{n} \cdot \sum_{i=1}^{n} (|x_i - r_i|) \tag{3}$$

with A being the accuracy, x the measured values and r the reference values. The MAE was chosen instead of the root mean square error because of the advantages discussed in (Willmott and Matsuura, 2005) and (Chai and Draxler, 2014).

2.5 Evaluation Schemes

The data sets of 2010 and 2011 were evaluated independently from each other as the devices seemed hardly comparable.

2.5.1 Hypothesis One

To test the first hypothesis (section 1.1) the three quality metrics were calculated for every validation bath and calibration technique. An additional summary of those results was created by calculating mean values for every metric's calibration techniques. For the Trueness the mean was calculated from the absolute bias.

As all reference temperatures showed a trend in time except for the cold bath in 2010 (figures 3 and 6), the calculation of the MAD was adjusted to minimize the effects of these trends. This included the data sets of the respective validation baths being split into contiguous blocks along time with an extent of ten minutes each. Absolute deviations were then calculated within the single blocks.

In this context the calculation of the MAD was also adjusted regarding the sharp temperature increase of the cold baths reference temperature in 2010 at about 7:30 (figure 3). Care was taken to not let any data block overlap with that point. Further customization was done by using only the data after 10 o'clock for the 2011 experiment's warm validation baths because the reference temperature showed a weaker trend for that time than during the hours before (figure 6).

The calculation of the bias and the MAE was not affected by the trends in the reference temperatures as the fiber temperatures could be matched with their corresponding reference temperature for every point in time.

In order to identify statistically insignificant differences for the same quality metric and either the same validation bath and different calibration techniques or the same calibration techniques and different validation baths, two-sided t-tests were used with the respective sets of errors, absolute deviations or absolute errors. In the cases of the validation baths being the same the pair-wise variant of the t-test was used.

The t-test's requirement for normal distribution of the tested values was given by the central limit theorem as every data set consisted of thousands of values. The second requirement of independence of these values was achieved for the values along the time dimension by avoiding the effect of the long term trend seen in the reference temperatures as described above. Because backscatter within the fiber at one measurement section also influences adjacent measurement sections (Krause and Blume, 2013) at first independence along the space dimension was not given. This was also true as the temperature measurements of the validation baths in the 2010 experiment showed a distinct "sawtooth" pattern as a consequence of temperature layering (figure 7). To cope with
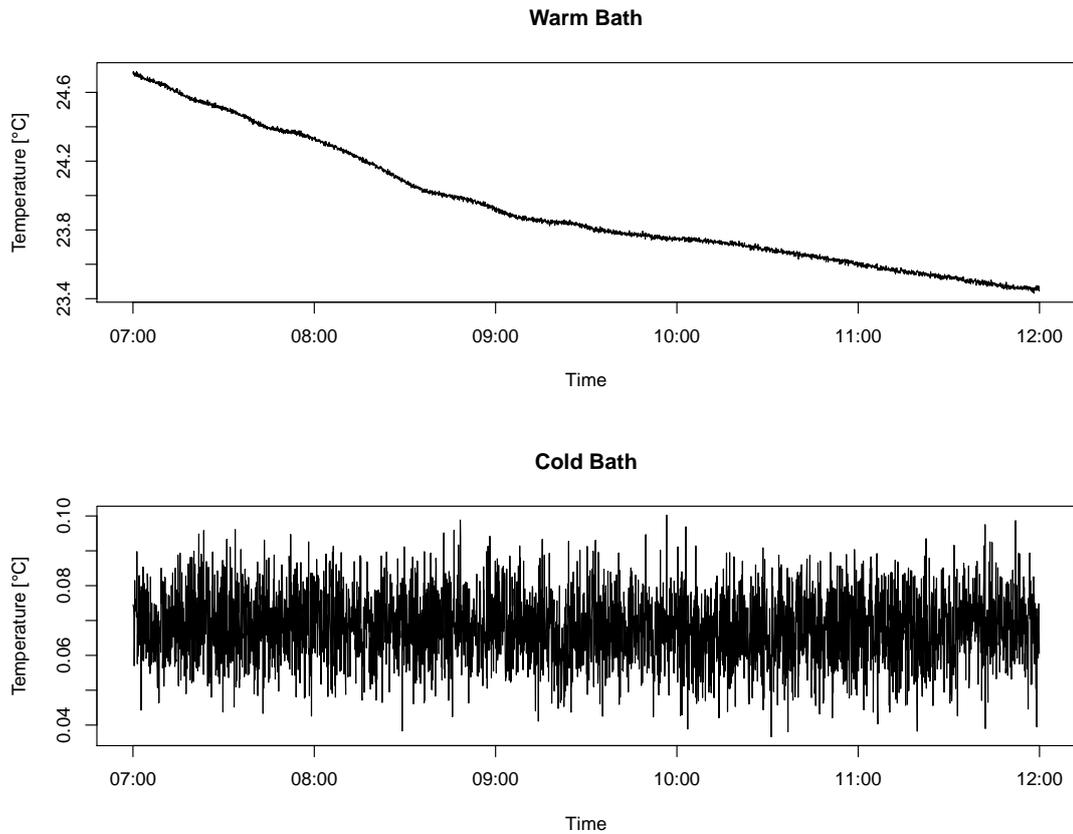
Figure 6: Reference Temperatures (2011)

this inadequacy all data sets were aggregated along the space dimension by calculating means of every three values. The aggregation interval size of three values was chosen because the sawtooth pattern gave the impression of roughly following that period.

The results of the t-tests were evaluated at a significance level of 0.01.

2.5.2 Hypothesis Two

Testing of the second hypothesis was done using aggregation interval sizes ranging from one to twenty-five values along each dimension, resulting in maximum aggregations of 316.25 centimeters along the fiber and 75 seconds for 2010 or 125 seconds for 2011 respectively with the exception of the 2010 experiment's cold bath at the end of the fiber not having a sufficient spatial extent and therefore only reaching a maximum aggregation interval size of fifteen values along the fiber.

To create correspondingly aggregated data ten-thousand samples were taken for every combination of spatial and temporal aggregation interval sizes from every data set of the different validation baths and calibration techniques. The samples were positioned
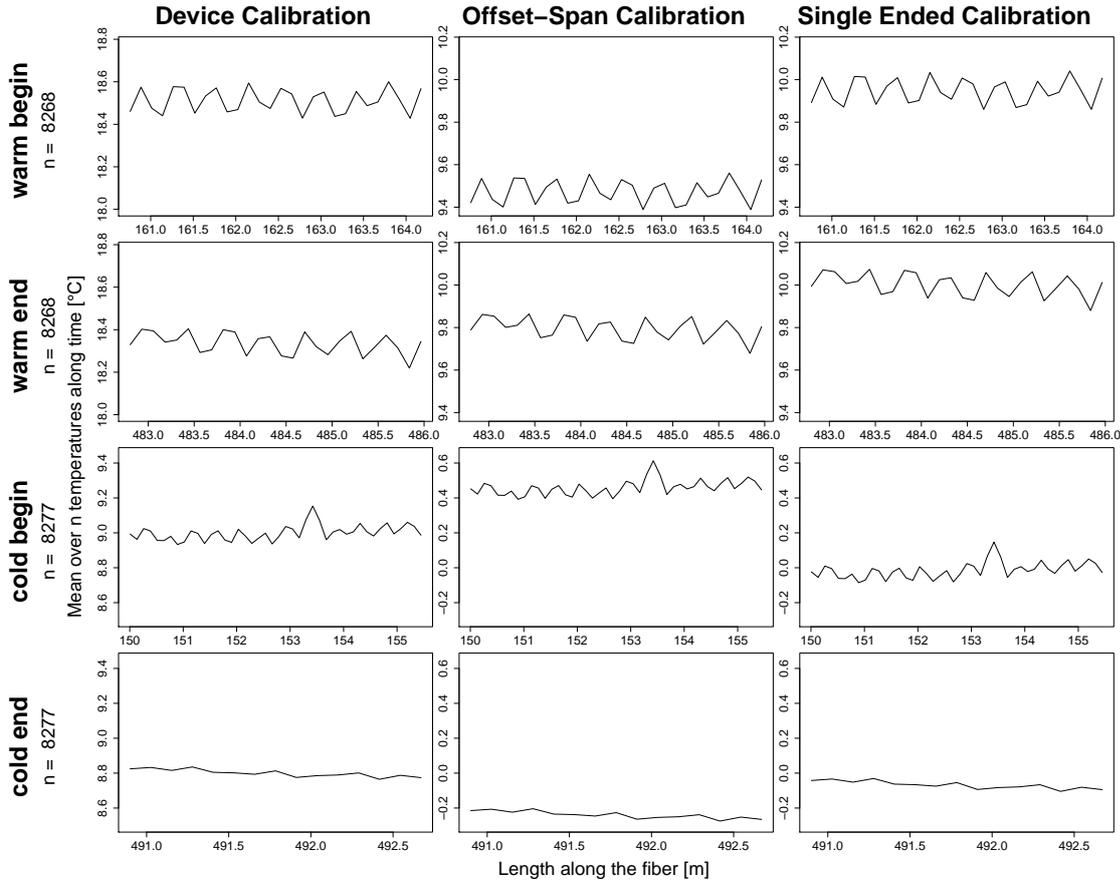
Figure 7: Mean temperatures along the fiber (2010)

randomly within the data sets and consisted of contiguous, two-dimensional data sets of the corresponding interval sizes.

Quality metrics were calculated by taking the arithmetic mean of the samples and comparing that to either the mean of the corresponding reference temperatures in the cases of bias and MAE or the mean of all data points within a range of five minutes before the first and after the last value of the sample in case of the MAD, with the exception of the 2011 experiment's cold bath where the mean of the whole data set was compared to the sample mean.

The errors and deviations thereby calculated for data set were then used to calculate a mean value which would be the corresponding quality metric and a standard deviation to get a measure for the metric's precision.

For the calculation of the MAD of the 2011 experiment's warm bath only the data after 10 o'clock was used similar to the calculation of the overall quality metrics when testing the first hypothesis (section 2.5.1).

To visualize the results, for every quality metric, validation bath and calibration

method two heat maps complemented by contour lines were drawn, showing the relationship between the spatial and temporal extent of the samples and the metric. While the first of the two heat maps was drawn using the quality metric data itself, the second one shows the standard deviation of the corresponding errors or deviations.

3 Results

3.1 2010 Experiment

3.1.1 Data Quality for Different Validation Baths and Calibration Techniques

When comparing the mean bias for different calibration methods, one notices the single ended calibration having the lowest value followed by the offset-span calibration which again shows a lower value than the device calibration (table 2). Note the device calibration having a bias somewhere around eight to nine degrees Kelvin.

For the offset-span and the single ended calibration the biases are both above and below zero in contrast to the device calibration.

Moreover do device and offset-span calibration show a higher absolute bias for the baths at the beginning of the fiber than for the baths at the end while it is the other way around for the single ended calibration.

Table 2: Bias (2010) ; numbers in the same row or column and with the same letters are not significantly different from each other at a p-value of 0.01

	Device	Offset-Span	Single Ended
Warm Begin	8.5385	-0.5008	-0.0263_a
Warm End	8.3685	-0.1721	0.0348
Cold Begin	9.0388	0.4980	0.0242
Cold End	8.8397	-0.1997	-0.0281_a
Mean of Absolutes	8.6964	0.3426	0.0284

For the precision the device calibration shows the lowest MAD and while the offset-span calibration's are only slightly higher, the MAD of the single ended calibration's baths remain above both of them (table 3).

Besides that the MAD of the bath sections at the end of the fiber are generally higher than those of the sections with the same temperature but at the beginning of the fiber.

Table 3: Mean Absolute Deviation (2010) ; numbers in the same row or column and with the same letters are not significantly different from each other at a p-value of 0.01

	Device	Offset-Span	Single Ended
Warm Begin	0.2175	0.2188_a	0.2692
Warm End	0.2535	0.2600	0.2981
Cold Begin	0.2109	0.2174_a	0.2665
Cold End	0.2468	0.2494	0.2830
Mean	0.2322	0.2364	0.2792

As for the calibration techniques' mean MAE, the highest error is made by the device calibration while the offset-span calibration still is a little bigger than the single ended calibration's MAE (table 4).

The Results for the device calibration's MAE are the same as for the device calibration's bias (table 2). This is not the case for the single ended calibration of which the MAE is rather close to it's corresponding MAD (table 3) for all baths.

While the offset-span calibration's MAE are not as close to either their corresponding bias or MAD they show a relatively big difference between bath sections of the same temperature. Furthermore there is no significant difference between bath sections at the same end of the fiber.

Table 4: Mean Absolute Error (2010) ; numbers in the same row or column and with the same letters are not significantly different from each other at a p-value of 0.01

	Device	Offset-Span	Single Ended
Warm Begin	8.5385	0.5088_a	0.2725
Warm End	8.3685	0.2978_b	0.3013
Cold Begin	9.0388	0.5056_a	0.2682
Cold End	8.8397	0.2995_b	0.2845
Mean	8.6964	0.4029	0.2816

3.1.2 Data Quality for Different Aggregation Interval Sizes (AIS)

When calculated from differently aggregated validation data the bias stays mostly the same for every AIS (figure 8) and is as big as the bias calculated from the corresponding data set the way it was done in section 3.1.1. Nevertheless it slightly differs with the AIS along time.
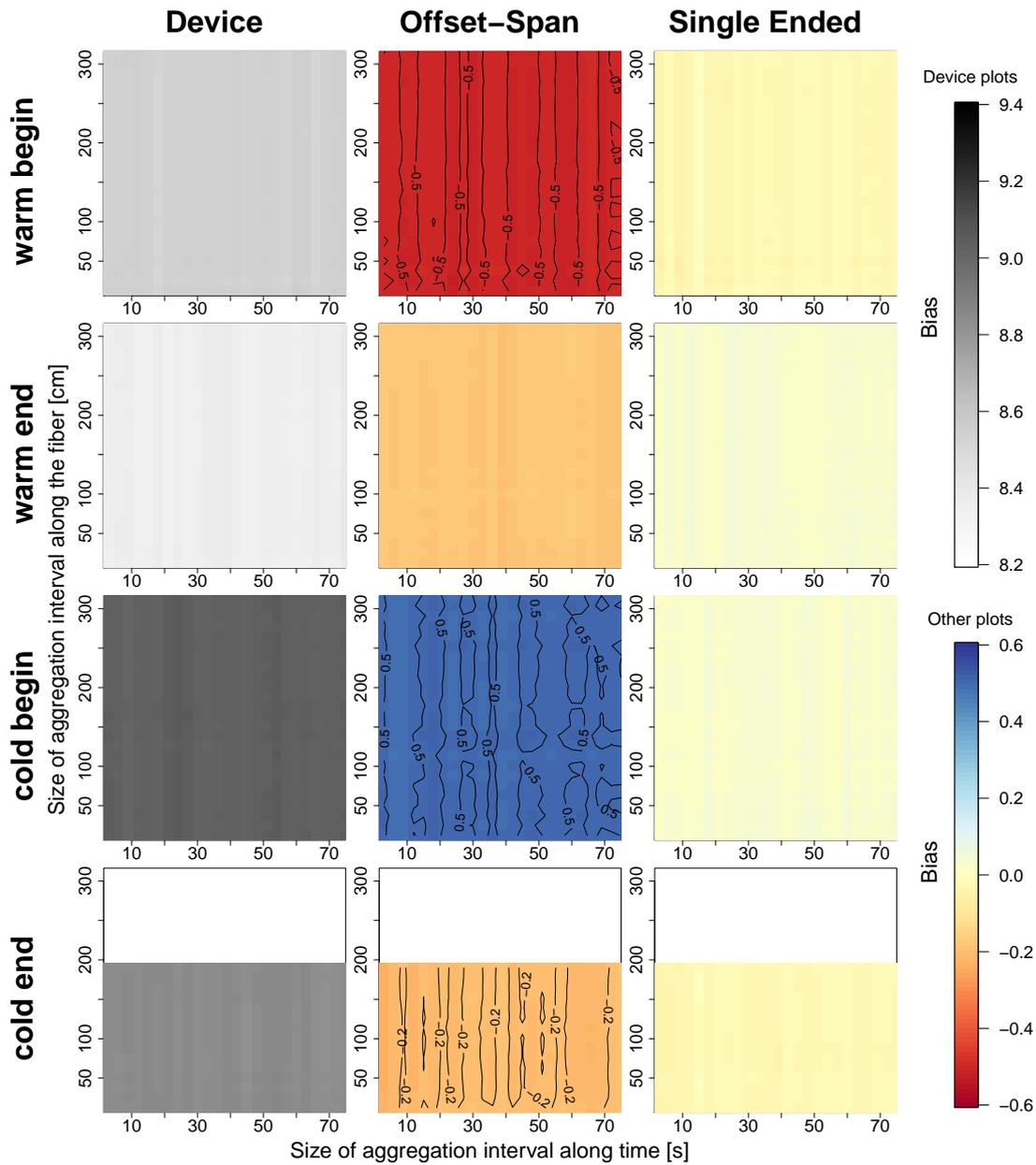
Figure 8: Bias for Different Aggregation Interval Sizes (2010)

The standard deviation of the errors that the bias is calculated from shows a decrease when the AIS is increased (figure 9). However that decrease is stronger while the AIS is still small. When looking at the "background" color, i.e. the overall standard deviation for big AIS, the offset-span calibration shows a smaller standard deviation than the device and the single ended calibration.

Another indicator for the relationship between AIS and the errors' standard deviation could be the intercept of the 0.15 contour line with the x-axis. While for the same validation baths this intercept lies at an only slightly smaller temporal AIS for the offset-span calibration than for the single ended calibration, it meets with a much higher temporal AIS for the device calibration.

Aside from the standard deviations at small AIS there are patterns similar to the bias (figure 8) that show slight differences in the standard deviation for different temporal AIS.

The MAD shows only small differences between the single ended calibration and the other two methods (figure 10).

One indicator would be the temporal AIS at which the 0.1 contour line intercepts the x-axis. While this AIS is only slightly smaller for the offset-span calibration when comparing it to the device calibration within the same validation bath, it is higher for the single ended calibration.

Furthermore the single ended calibration's MAD stays above the 0.05 contour line at bigger AIS than it does for the other calibrations.

To assess differences in the standard deviation of absolute deviations for different calibrations one can look at the position of the 0.04 contour line (figure 11). Even so those lines are hardly comparable for the device calibration, the single ended calibration still shows more values being above that threshold than the offset-span calibration for the whole range of AIS.

For the device calibration the MAE (figure 12) is the same as the bias (figure 8). This is also mostly the case for the absolute values of the offset-span calibration even so the baths at the end of the fiber show a small increase of the MAE for small AIS. The single ended calibration shows the smallest MAE of the three calibrations and it's 0.1 contour lines intercept the x-axis at slightly smaller temporal AIS for the cold validation baths than they do for the warm ones.

When comparing the different calibrations' standard deviation of absolute errors for different AIS (figure 13), the device calibration shows the highest values for big AIS. This can be seen when looking at the 0.1 contour line which does mostly not appear for the device calibration while it nearly intersects the x-axis for most offset-span calibrated data and does intersect it at around 30 seconds for the single ended calibration.
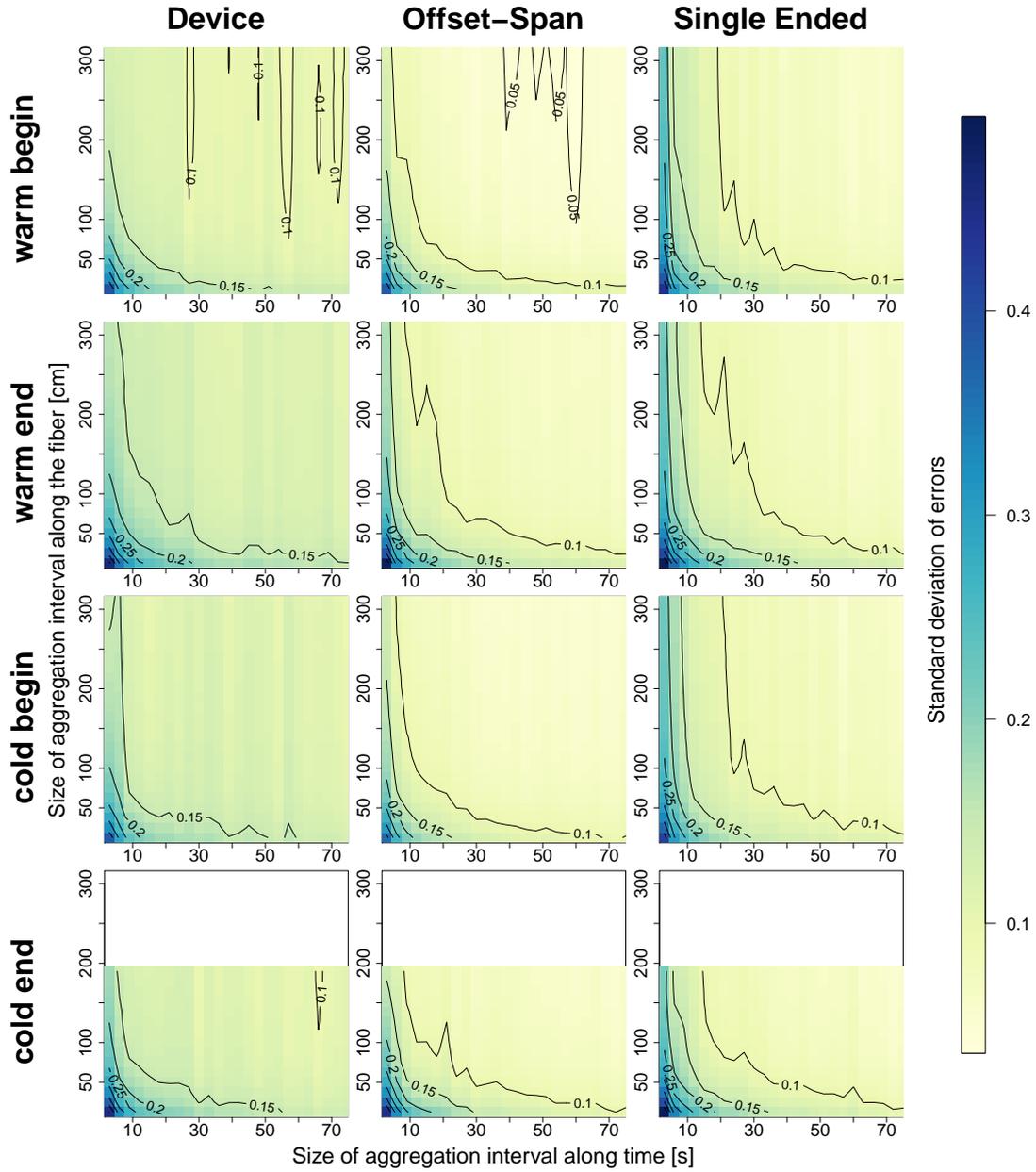
Figure 9: Standard Deviation of Errors for Different Aggregation Interval Sizes (2010)
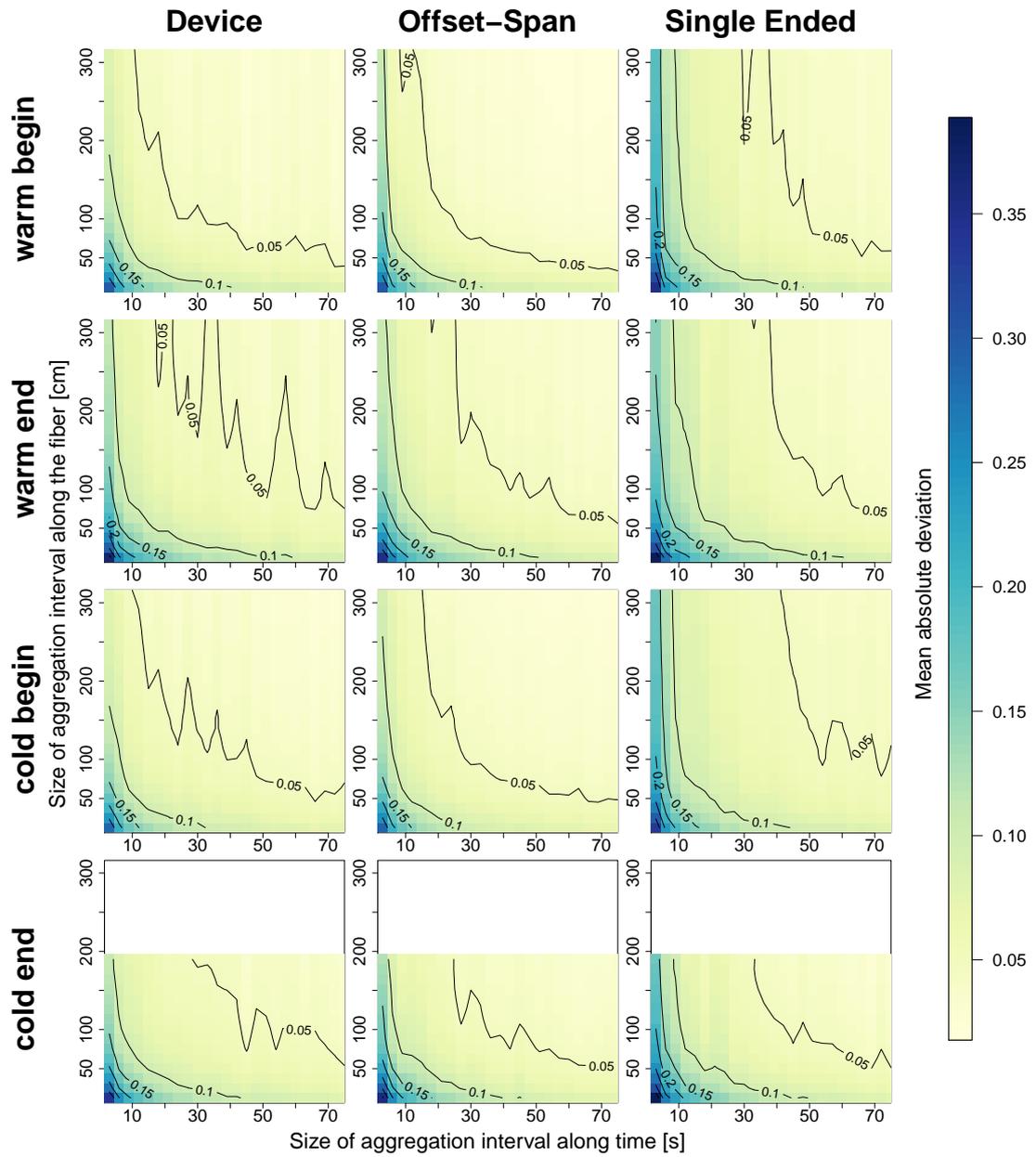
Figure 10: Mean Absolute Deviation for Different Aggregation Interval Sizes (2010)
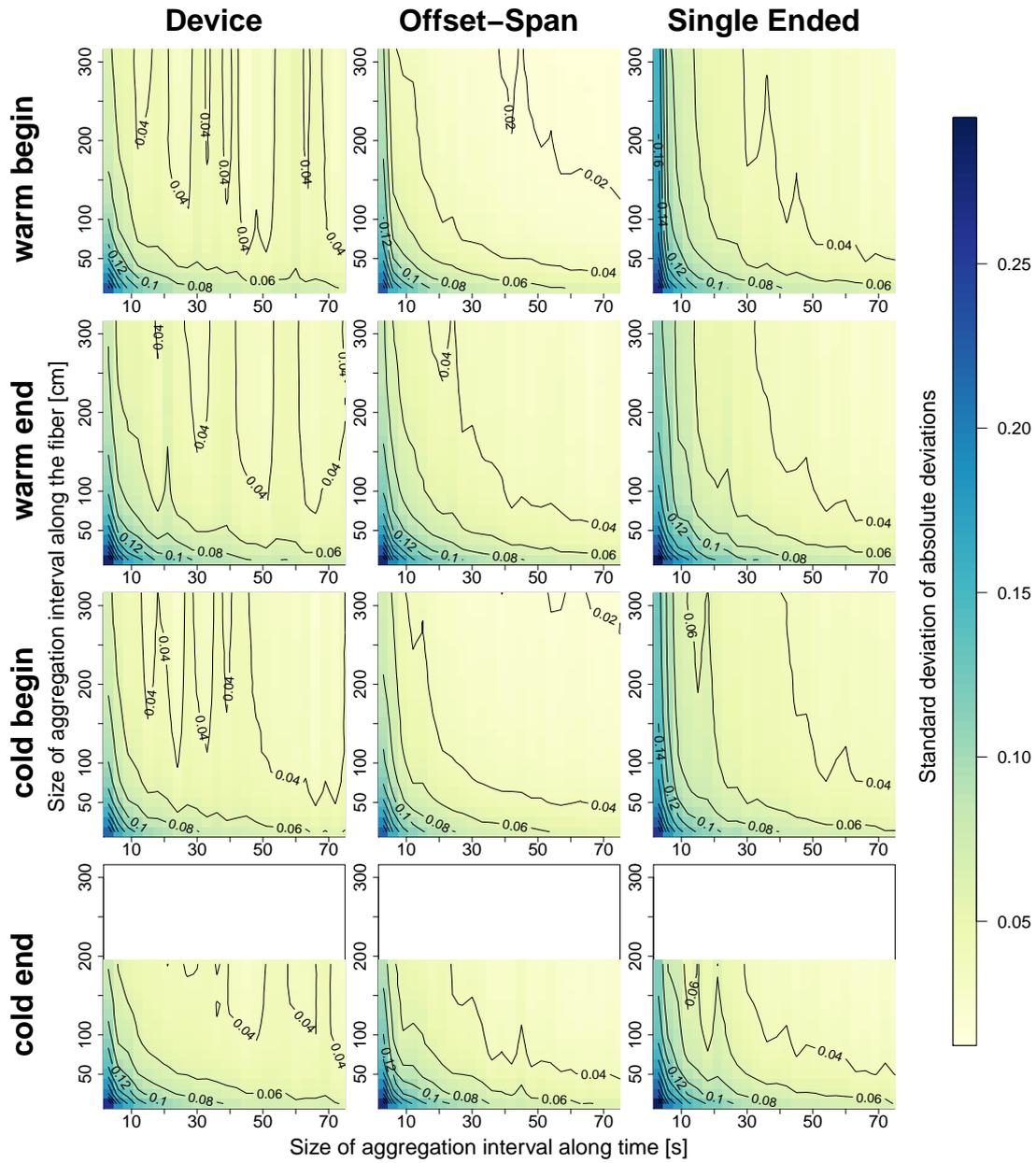
Figure 11: Standard Deviation of Absolute Deviations for Different Aggregation Interval Sizes (2010)
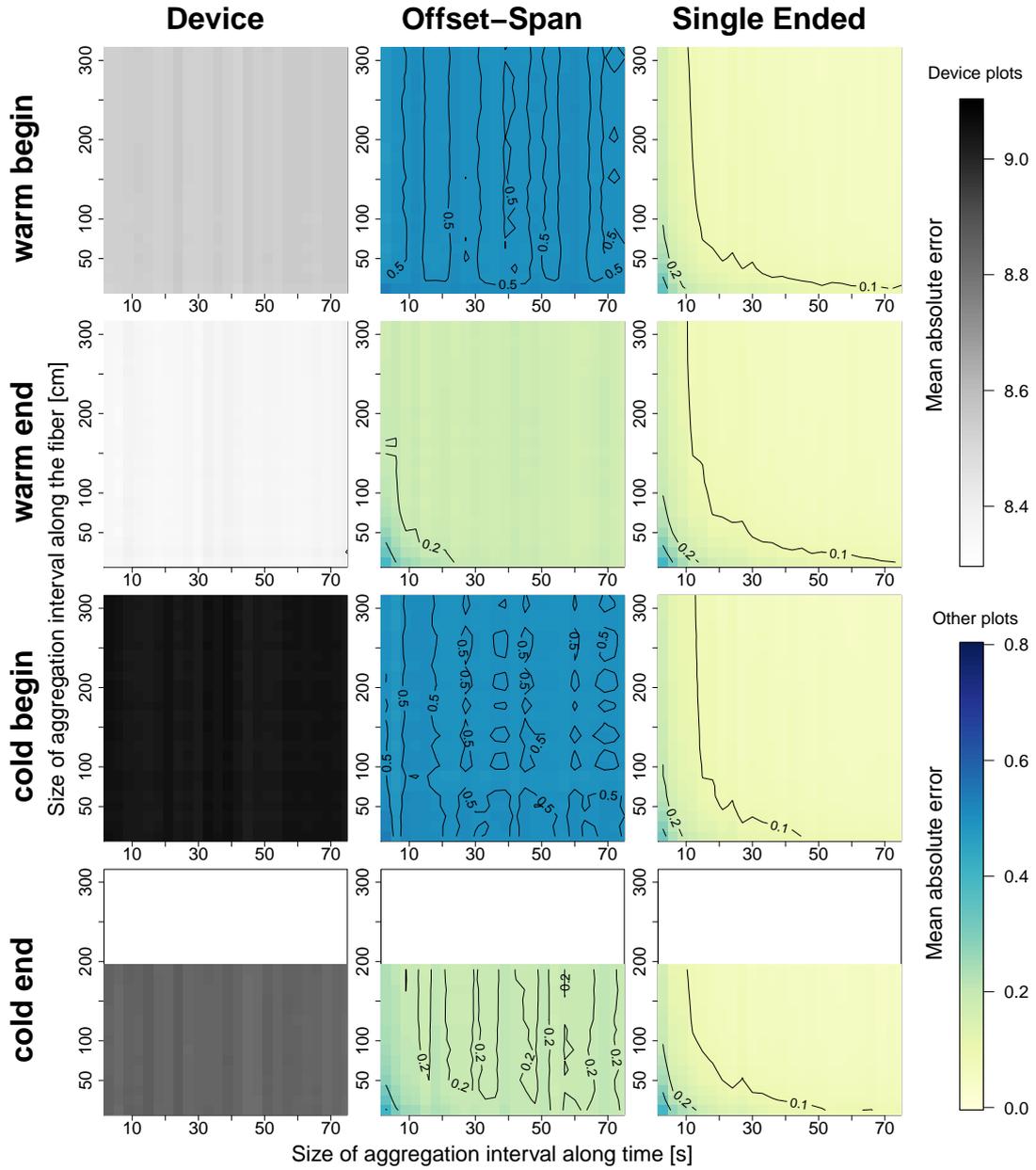
Figure 12: Mean Absolute Error for Different Aggregation Interval Sizes (2010)
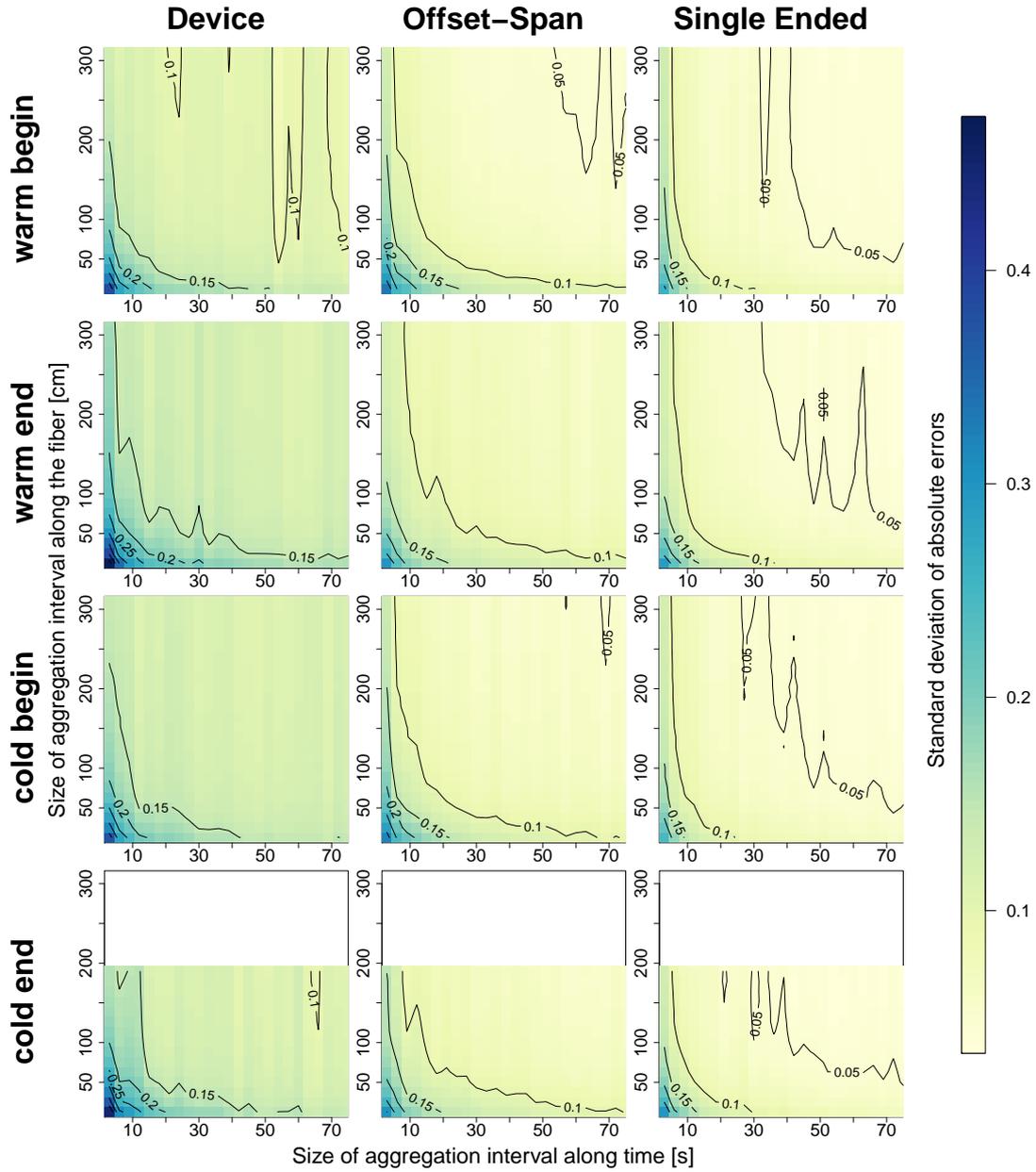
Figure 13: Standard Deviation of Absolute Errors for Different Aggregation Interval Sizes (2010)

3.2 2011 Experiment

3.2.1 Data Quality for Different Validation Baths and Calibration Techniques

The mean absolute bias is highest for the device calibration while the single ended calibration's mean absolute bias is the lowest (table 5). When looking at the sign of the single biases it shows that the device calibration is above zero for every validation bath while the other two calibration techniques produce biases of both signs.

Table 5: Bias (2011) ; numbers in the same row or column and with the same letters are not significantly different from each other at a p-value of 0.01

	Device	Offset-Span	Single Ended
Warm Begin	0.3715	-0.0678	-0.0101_a
Warm End	0.2747	-0.0966	0.0116
Cold Begin	0.4392	0.0680	0.0085
Cold End	0.3522	-0.0871	-0.0087_a
Mean of Absolutes	0.3594	0.0799	0.0097

While the mean MAD is slightly smaller for the single ended calibration than for the other two methods the biggest differences between MAD values present themselves for different validation baths (table 6). Here the single baths show similar values across the different calibration techniques while the differences between the bath sections at the beginning of the fiber and those at the end are relatively big.

Having a comparably big difference to the same validation bath's MAD for the device and offset-span calibration, the MAD for single ended calibration's cold bath section at the end of the fiber doesn't completely match with the pattern of the other values.

Table 6: Mean Absolute Deviation (2011) ; numbers in the same row or column and with the same letters are not significantly different from each other at a p-value of 0.01

	Device	Offset-Span	Single Ended
Warm Begin	0.1547	0.1506	0.1627
Warm End	0.2451	0.2415	0.2434
Cold Begin	0.1470_a	0.1372	0.1471_a
Cold End	0.2431	0.2372	0.2068
Mean	0.1975	0.1916	0.1900

The mean MAE for the device calibration is bigger than for the offset-span calibration which is again only a little bit bigger than that of the single ended calibration (table 7). When comparing the single validation bath sections for the offset-span and the single ended calibration, one can see the baths at the beginning of the fiber having pretty similar MAEs while those at the end show bigger differences.

While the MAE for the single ended calibration is very similar to the same calibration's MAD there is less similarity between the offset-span calibration's MAE and MAD. For the device calibration the validation bath sections at the beginning of the fiber show a close correspondence for their MAE and bias but less so for the bath sections at the end of the fiber where the MAE is bigger than the bias.

Table 7: Mean Absolute Error (2011)

	Device	Offset-Span	Single Ended
Warm Begin	0.3761	0.1565	0.1631
Warm End	0.3539	0.2716	0.2382
Cold Begin	0.4402	0.1481	0.1469
Cold End	0.3894	0.2484	0.2070
Mean	0.3899	0.2062	0.1888

3.2.2 Data Quality for Different Aggregation Interval Sizes (AIS)

The bias does not change much for different AIS but the changes are stronger while changing the temporal AIS rather than the spatial one (figure 14). Furthermore, the overall biases for the different data sets are very similar to those calculated in section 3.2.1.
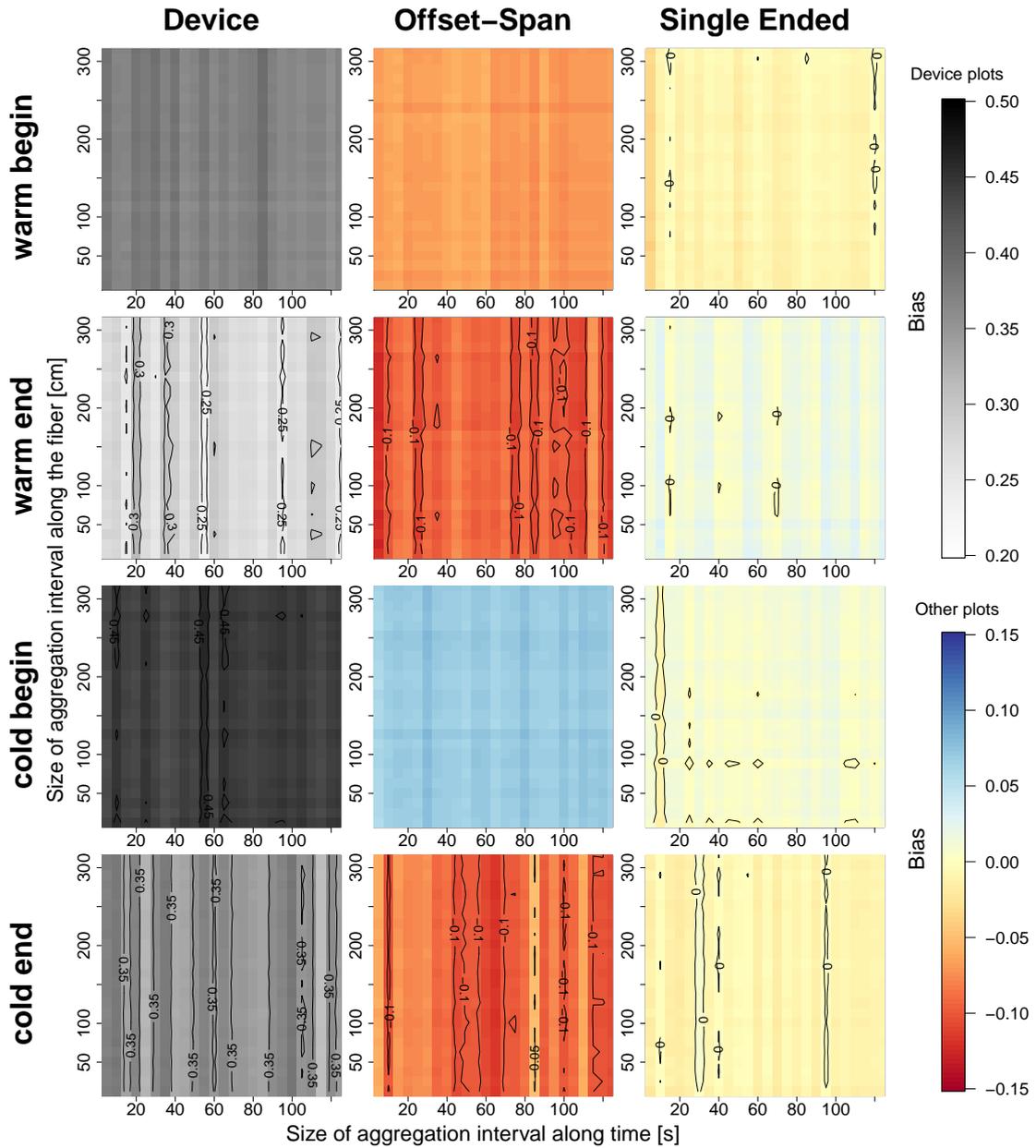
Figure 14: Bias for Different Aggregation Interval Sizes (2011)

For the single ended calibration the standard deviation of errors is quite similar across different validation baths (figure 15). However this is not true for the other two calibration techniques where the bath sections at the end of the fiber differ strongly from those at the beginning. Using the contour line at 0.15 as a marker one can see that for the bath sections at the beginning the standard deviation gets already lower for very small AIS whereas for the bath sections at the end that threshold is only reached a few times and only for high AIS.

When comparing the MAD for different calibration methods and validation bath sections one can look at the temporal AIS where the 0.08 contour line intercepts the x-axis (figure 16). That temporal AIS differs mostly for different validation baths sections and not so much for different calibration techniques. It is quite low for both baths sections at the beginning of the fiber compared to those at the end. One exception is made by the cold bath section at the end for the device and the offset-span calibration where an MAD of 0.8 is not reached at any AIS and the smallest MADs get to around 0.12.

The standard deviation of the data set's absolute deviations behave very similar to the MAD when looking at the 0.06 contour line (figure 17) instead of the 0.08 contour line. The exception here is also made by the cold bath section at the end of the fiber for the device and the offset-span calibration where the smallest values only get to around 0.08.

For the device calibration the MAE shows nearly no changes for different AIS along the fiber but some for different temporal AIS (figure 18). Still the overall MAE for all AIS of the device calibration's different calibration baths is similar to that calculated before for only one small AIS (table 7).

This is not the case for the offset-span and the single ended calibration where the MAE decreases with small increasing AIS and is otherwise pretty similar for the rest of the them. This overall MAE is higher for the offset-span calibrated bath sections at the end of the fiber and a little lower for the single ended calibration's cold bath sections. The other data sets show MAE somewhere between 0.1 and 0.05.

Finally the standard deviation of absolute errors also decreases with the increase of small AIS but is otherwise lowest for the single ended calibration, a little bit higher for the offset-span corrected data and highest for the device calibration (figure 19).

Besides that the MAE for the bath sections at the end of the fiber is generally higher than for those at the beginning. This also applies to the increase in MAE from the single ended to the offset-span and further to the device calibration.
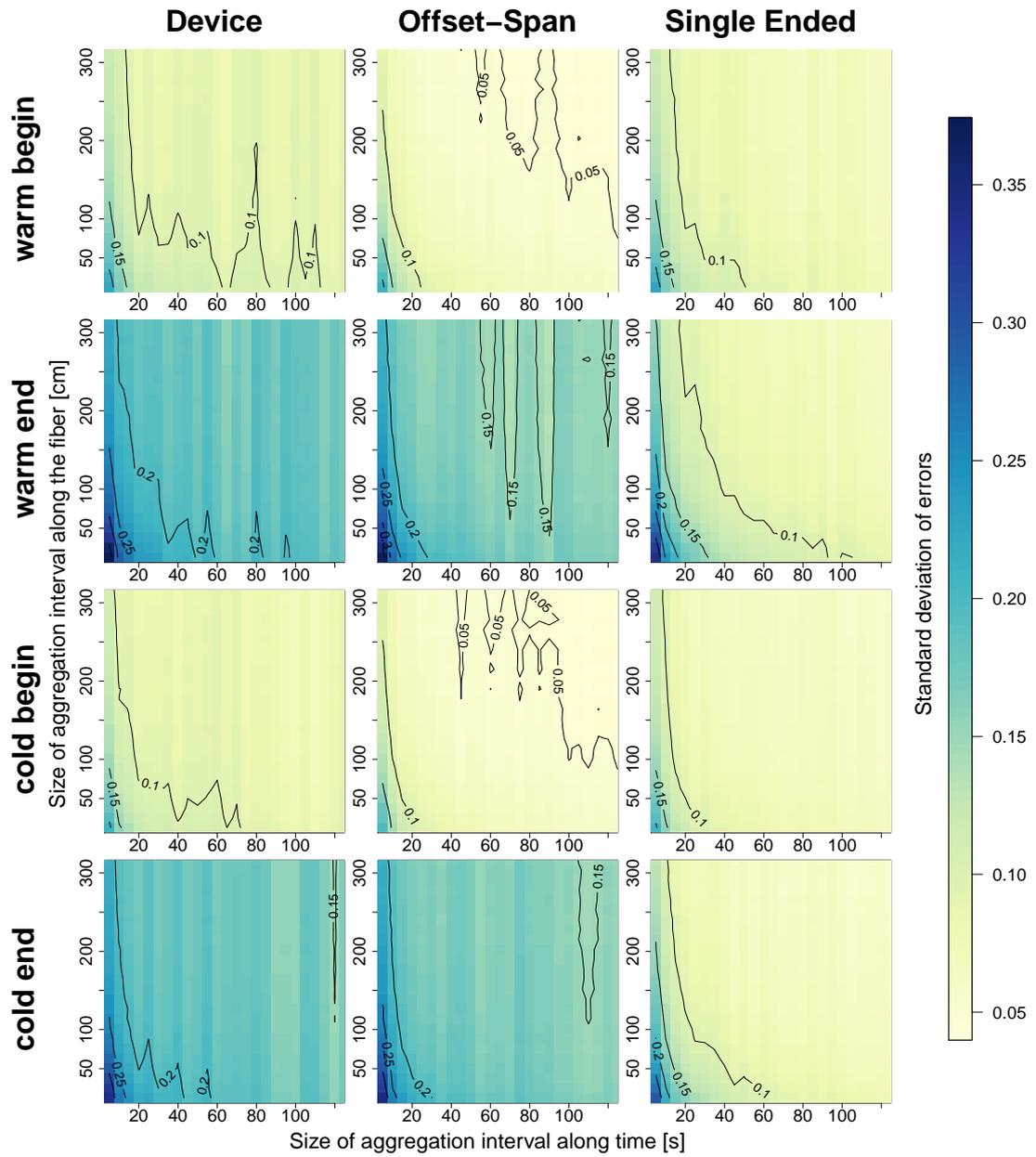
Figure 15: Standard Deviation of Errors for Different Aggregation Interval Sizes (2011)
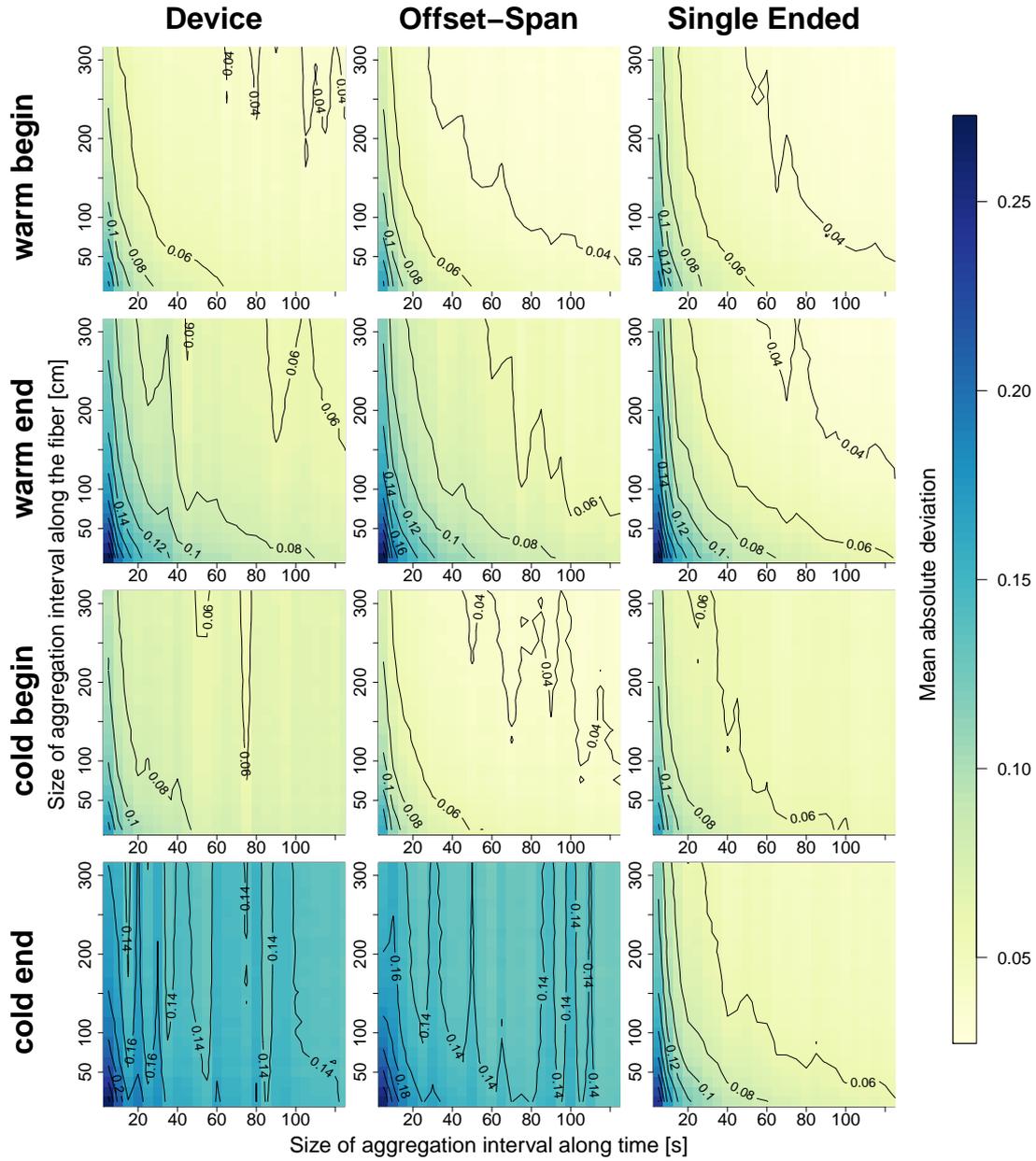
Figure 16: Mean Absolute Deviation for Different Aggregation Interval Sizes (2011)
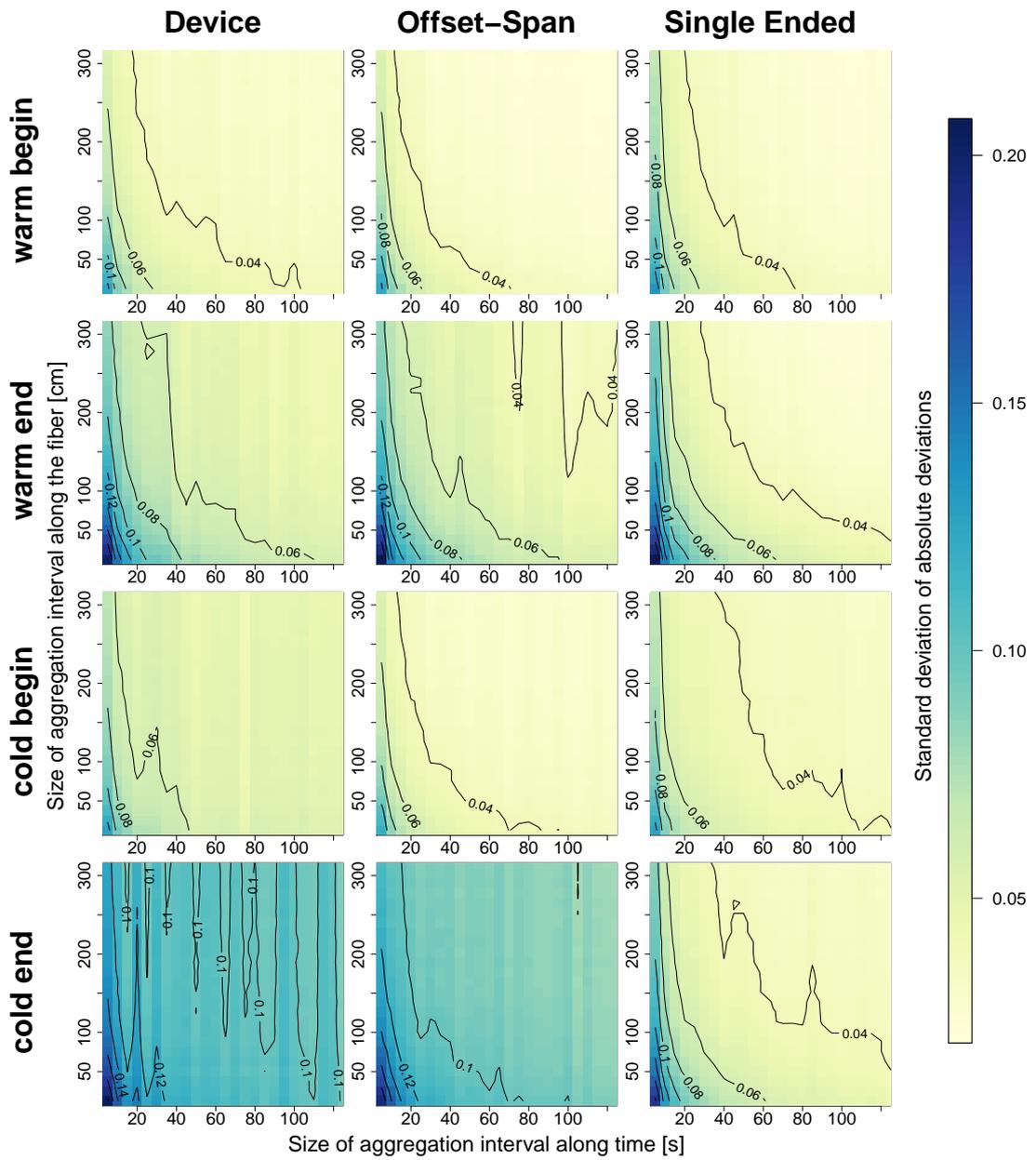
Figure 17: Standard Deviation of Absolute Deviations for Different Aggregation Interval Sizes (2011)
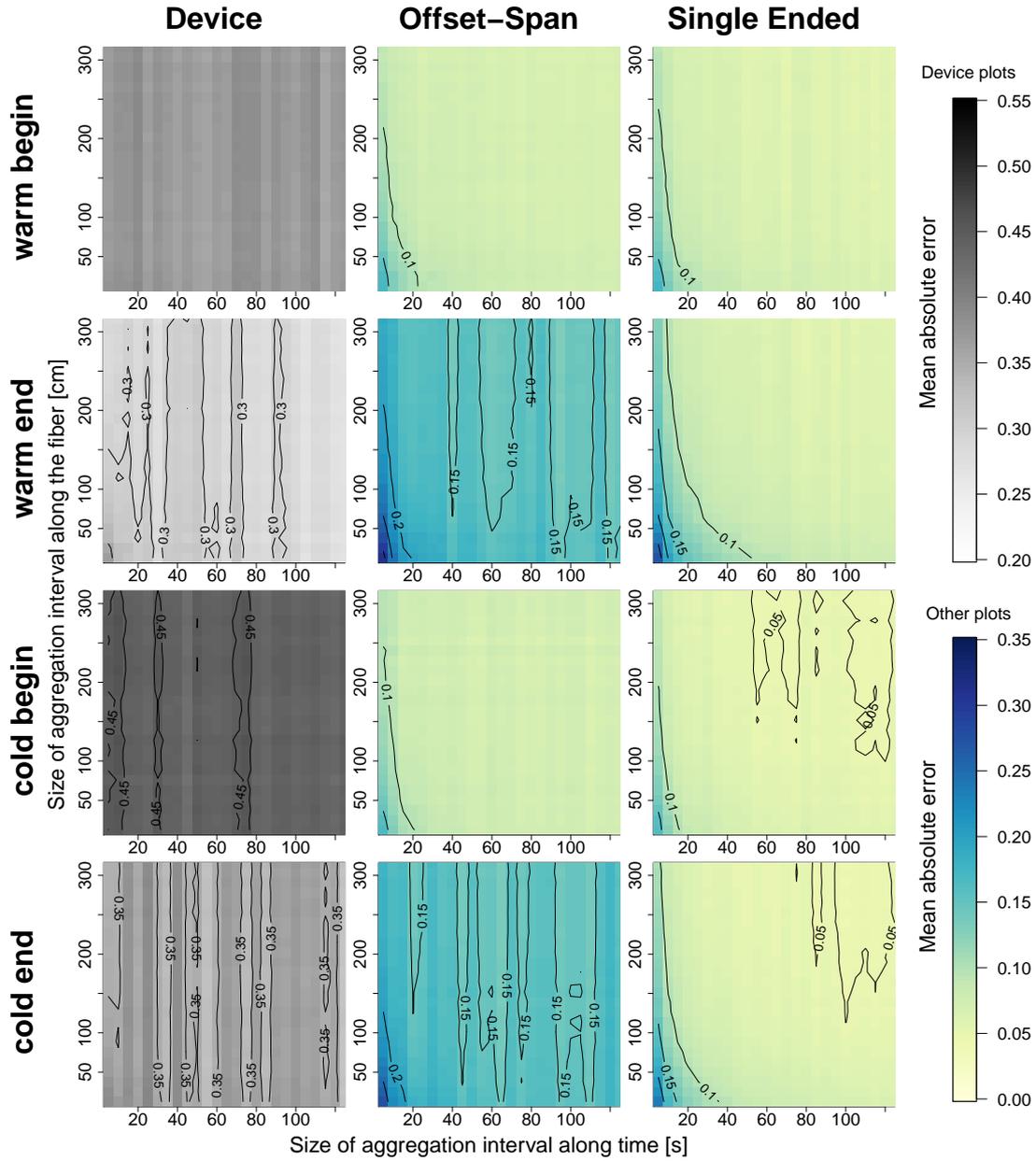
Figure 18: Mean Absolute Error for Different Aggregation Interval Sizes (2011)
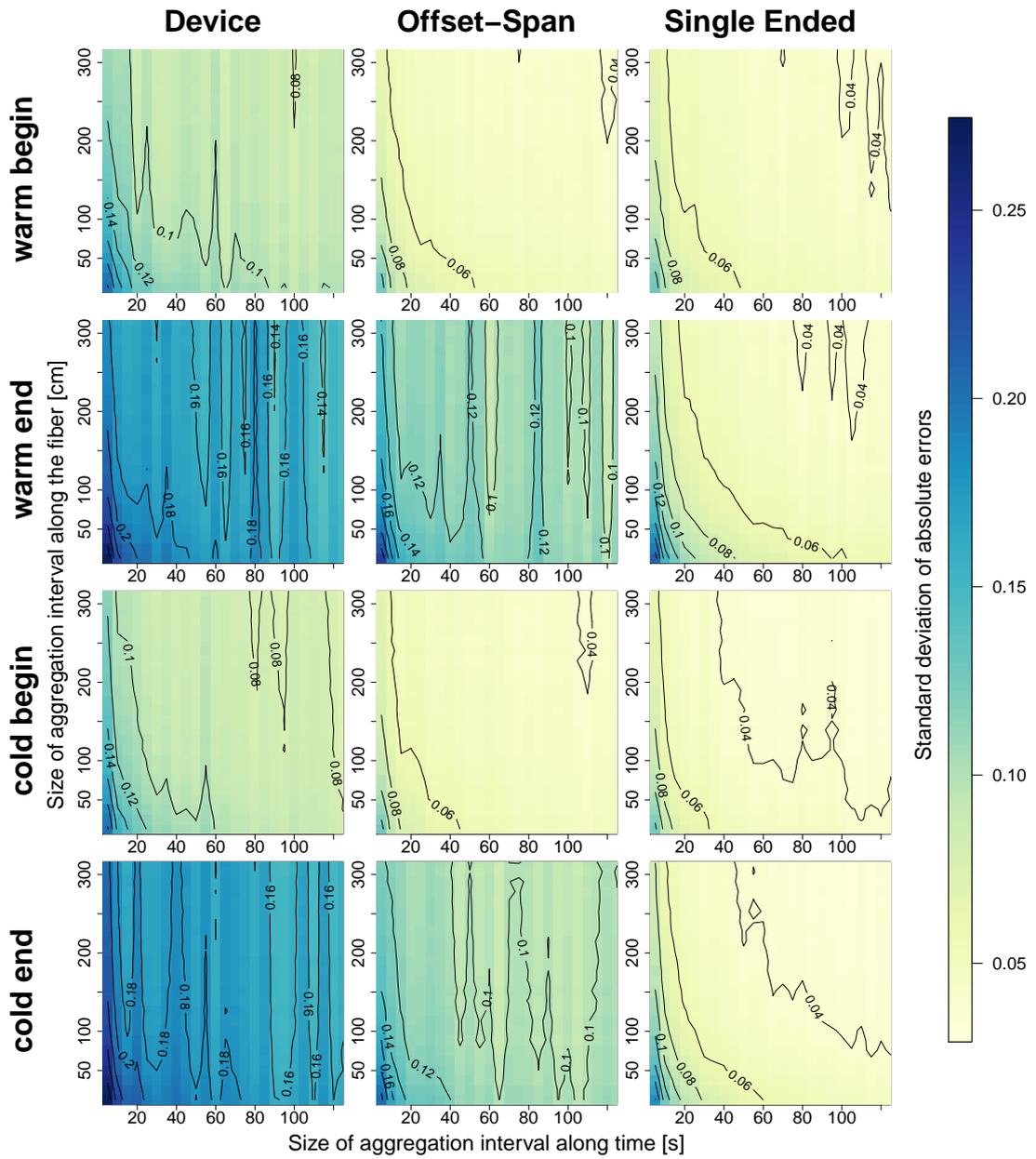
Figure 19: Standard Deviation of Absolute Errors for Different Aggregation Interval Sizes (2011)

4 Discussion

When looking at the results of a calibration one always has to be aware of the limitation that the calibration's results can not be of a better quality than what the deployed reference sensor is capable of measuring at.

In that sense biases below the reference sensor's accuracy are still comparable as they only indicate how close measured values get to the reference value.

The reference sensors' reported specifications are noted in section 2.1.3 but to also get an estimate of their precision the MAD of the 2011 experiment's cold bath reference temperatures (figure 6) was calculated as 0.008 and by looking at the reference temperatures of the experiment in 2010 (figure 3) the precision was assumed to be similar to the temperature resolution which again was estimated to be 0.1 degrees Kelvin.

4.1 2010 Experiment

4.1.1 Hypothesis One

Regarding the bias the first hypothesis holds true. While the big bias of the device calibration can probably be attributed to the device being a prototype the differences between validation bath sections of different temperatures might be caused by an offset between the reference thermometers. On the other hand the difference between the offset-span calibrated bath sections at the beginning of the fiber (table 2) is well above than what could be explained by that alone.

The device calibration and the offset-span calibration showing similar MADs is probably due to the way the offset-span calibration works. As long as the initial offsets of the two calibration bath sections from their corresponding reference temperature aren't too different from each other the deviations of the device calibrated temperatures shouldn't change that much when doing the span correction.

For the single ended calibration to be showing higher MADs than the other two calibrations contradicts the first hypothesis. One reason might be the precision of the reference measurements used for the calibration to be worse than those used internally by the instrument. Another reason might of course also be the single ended calibration itself. Maybe it increases differences between stokes-anti-stokes ratios more than the device calibration.

The phenomenon of the precision getting worse when going farther away from the instrument is well known (Hausner et al., 2011) and caused by the laser pulse loosing energy while traveling through the fiber. Therefore the signal from more remote sections on the fiber is weaker leading to a worse precision of the measurement.

To be the same as the bias, the MAE has that property because the bias is so big that all errors are positive and therefore calculating absolute errors doesn't change anything (table 4). The inverse seems to be true for the single ended calibration where the bias is so small that it plays nearly no role when calculating the MAE. As a mix of both there

is the offset-span calibration of which the bath sections at the beginning of the fiber have a MAE that is similar to their bias and the MAE of the sections at the end of the fiber is close to their MAD. It seems to be the case that the MAE is not approximately a sum of bias and MAD but rather always a little bit bigger or of the same size as the bigger one of them.

Overall the MAE supports the first hypothesis but the question about why the single ended calibration conveys a higher MAD than the other two techniques remains unanswered. Moreover in those cases where the offset-span calibration shows only a small bias it performs very similar to the single ended calibration.

4.1.2 Hypothesis Two

Figure 8 shows that the measurement's bias can't be decreased by aggregating data. The slight changes for different temporal AIS might be due to temperature oscillations in the baths paired with an interplay between the different time constants of the fiber and the reference thermometer.

While the MAD can be decreased a lot by aggregating data (figure 10) this seems to be a little bit less effective for the single ended calibration where higher AIS are needed to reach certain thresholds, e.g. the 0.05 contour line. Nevertheless the second hypothesis is supported by these results even so the first one is not.

The standard deviation of the absolute deviations (figure 11) doesn't seem to be conveying any new information aside from the fact that aggregating data does not only lead to having a lower MAD but also comes with the single absolute deviations deviating less from the MAD.

Similar to the results for the overall data sets in section 3.1.1 the MAE only shows a dependence on the AIS when the bias is small enough and thus the influence of the precision gets stronger (figure 12). This is the case for all validation bath section of the single ended calibration and for those at the end of the fiber for the offset-span technique.

It's interesting that the smaller influence of the precision on the accuracy for the offset-span calibration also leads to the MAE not reaching values as low as it does for the single ended calibration even so the MAEs calculated for the whole data sets are very similar (table 4) for the bath sections at the end of the fiber.

A higher influence of the precision on the accuracy seems to decrease the deviation of the absolute errors around the MAE as can be seen when comparing the different calibrations in figure 13.

4.2 2011 Experiment

4.2.1 Hypothesis One

The Biases of the validation baths for the 2011 experiment definitely support the first hypothesis with the single ended calibration conveying values that are a magnitude smaller than those of the offset-span calibration and two magnitudes smaller than those of the device calibration (table 5).

While calibration techniques perform pretty similar concerning the MAD the offset-span calibration still seems to be a little bit better than the other two. Otherwise the size of the MAD depends mostly on the validation bath section and not on the calibration method. The cause for the low MAD at the single ended calibration's cold validation bath at the end of the fiber couldn't be determined.

For the single ended calibration's MAE (table 7) the bias doesn't seem to play a role because it is too small. This changes a little for the offset-span calibration where the bias is bigger and therefore the MAE also differs more from the MAD. At last for the device calibration where bias and MAD are of the same magnitude the MAE only gets much higher than the bias when the MAD gets close to the bias.

In general the single ended calibration provides the best results, thus endorsing the hypothesis. Yet using calibration techniques other than that of the device only decreases the bias but has little to no impact on the precision. As a consequence enhancing the measurement accuracy by using more elaborate calibration techniques is limited by the measurement precision.

4.2.2 Hypothesis Two

The bias does not support the second hypothesis as it doesn't change much for different AIS. The reason for the seemingly random changes along the spatial AISs are probably the same as for the 2010 experiment (section 4.1.2).

The mean absolute deviation shows a big difference for high AIS between the bath sections at the end of the fiber even so there is no such difference when calculating the MAD for the whole data set (see table 6). It may be possible that the DTS device's internal reference measured non-existent irregularities because such irregularities in the stokes and anti-stokes signals would have shown themselves for the single ended calibration as well.

The standard deviations of errors (figure 15) and absolute deviations (figure 17) don't convey any more information than the corresponding plots already did for the 2010 experiment's data.

Regarding the relationship between bias, MAD and MAE the same patterns emerge for the MAD (figure 18) as they did for the data from the experiment in 2010.

Lastly the second hypothesis can be regarded as confirmed for the single ended and offset-span calibration as those methods convey low biases and therefore enable the aggregation to have a positive impact on the accuracy through the precision.

5 Conclusion

In conclusion, the first hypothesis holds true as long as the bias is not too close to zero as the possibilities of quality enhancement through the usage of elaborate calibration techniques are limited to the trueness and thereby also to the trueness's influence on accuracy.

However in order to enhance data quality even further, the usage of an elaborate calibration technique can be complemented by the aggregation of measurements which has been shown to have a positive impact on precision. A little different to the second hypothesis this effect doesn't stop at a certain aggregation interval size but rather attenuates for bigger aggregation intervals.

The second effect of data aggregation is the decreased deviation of the single measurement values' errors, absolute errors and absolute deviations which increases the quality metrics' reliability for single measurement values.

Because of these properties the second hypothesis only holds true for measurements that do not show too big of a bias.

References

T. Chai and R. R. Draxler. Root mean square error (RMSE) or mean absolute error (MAE)? - Arguments against avoiding RMSE in the literature. *GEOSCIENTIFIC MODEL DEVELOPMENT*, 7(3):1247–1250, 2014. ISSN 1991-959X. doi: {10.5194/gmd-7-1247-2014}.

P. De Bièvre. The 2012 international vocabulary of metrology: "vim". *Accreditation and Quality Assurance*, 17(2):231–232, Apr 2012. ISSN 1432-0517. doi: 10.1007/s00769-012-0885-3. URL https://doi.org/10.1007/s00769-012-0885-3.

M. B. Hausner, F. Suarez, K. E. Glander, N. van de Giesen, J. S. Selker, and S. W. Tyler. Calibrating Single-Ended Fiber-Optic Raman Spectra Distributed Temperature Sensing Data. *SENSORS*, 11(11):10859–10879, NOV 2011. ISSN 1424-8220. doi: {10.3390/s111110859}.

ISO 5725-4:1994. Accuracy (trueness and precision) of measurement methods and results – part 4: Basic methods for the determination of the trueness of a standard measurement method. Standard, International Organization for Standardization, Geneva, CH, Dec. 1994.

S. Krause and T. Blume. Impact of seasonal variability and monitoring mode on the adequacy of fiber- optic distributed temperature sensing at aquifer- river interfaces. *WATER RESOURCES RESEARCH*, 49(5):2408–2423, MAY 2013. ISSN 0043-1397. doi: {10.1002/wrcr.20232}.

S. J. S., T. Luc, H. Hendrik, M. Alfred, L. Wim, van de Giesen Nick, S. Martin, Z. Josef, W. Martijn, and P. M. B. Distributed fiber-optic temperature sensing for hydrologic systems. *Water Resources Research*, 42(12), 2006. doi: 10.1029/2006WR005326. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2006WR005326.

C. K. Thomas, A. M. Kennedy, J. S. Selker, A. Moretti, M. H. Schroth, A. R. Smoot, N. B. Tufillaro, and M. J. Zeeman. High-Resolution Fibre-Optic Temperature Sensing: A New Tool to Study the Two-Dimensional Structure of Atmospheric Surface-Layer Flow. *BOUNDARY-LAYER METEOROLOGY*, 142(2):177–192, FEB 2012. ISSN 0006-8314. doi: {10.1007/s10546-011-9672-7}.

C. Willmott and K. Matsuura. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *CLIMATE RESEARCH*, 30(1):79–82, DEC 19 2005. ISSN 0936-577X. doi: {10.3354/cr030079}.

M. J. Zeeman, J. S. Selker, and C. K. Thomas. Near-surface motion in the nocturnal, stable boundary layer observed with fibre-optic distributed temperature sensing. *Boundary-Layer Meteorology*, 154(2):189–205, Feb 2015. ISSN 1573-1472. doi: 10.1007/s10546-014-9972-9. URL https://doi.org/10.1007/s10546-014-9972-9.

Affidavit

I declare that I wrote this thesis independently and on my own. I clearly marked any language or ideas borrowed from other sources as not my own and documented their sources. The thesis does not contain any work that I have handed in or have had graded as a Prüfungsleistung earlier on.

My name: Leon Steinmeier

Title of my thesis: Uncertainty of temperature estimates from fiber-optic distributed temperature sensing: Comparison of calibration methods

Date: 19th of June, 2018

Signature: