Response to both reviewers

In this feedback to reviewers we copied in the author response that we provided for each point mentioned by the reviewer and then added our response and how we revised the manuscript. Reviewer feedback is copied in with black text, our final author response and our plan to revise our manuscript (doi:10.5194/amt-2019-402-AC1 and doi:10.5194/amt-2019-402-AC2) is given in indented gray text, and our response of how we actually were able to do this is presented with blue text. Line numbers relating to the clean revised manuscript are given in blue font, followed by the line numbers in the track-changes version of the revised manuscript in red font.

Response to Reviewer #1

1 General comments

The manuscript summarises the findings from a 6-year field deployment of a small, low-cost methane sensor under low Arctic conditions. Given the interest of the community in small, low-cost sensors on the one hand and the measurement of atmospheric methane on the other hand, this work is of high relevance to the readers of Atmospheric Measurement Techniques.

The manuscript is written and structured well. The methods are described in appropriate detail. However, I see some shortcomings in the data analysis and presentation of results, detailed below, that should be addressed before publication.

> Thank you for your very valuable and detailed assessment, which definitely helps us improve our paper.

2 Specific comments

It is unclear which quantity the Authors use when they report the abundance of methane. The sentence 'We report all gas concentrations in mixing ratios by volume (ppm or ppb dry mole fractions)' (ll. 78–79) is contradictory, as it uses three distinct quantities as if they were synonymous. In a mixture of two components A and B, concentration c is defined as c = qA/(VA + VB) , where V is volume and q one of the quantities mass, amount, volume or number concentration [1]. Mixing ratio by volume r is rather uncommon and defined as r = VA/VB [2]. Mole fraction x (IUPAC recommends y for gaseous mixtures, but this is not common in atmospheric science) finally is defined as x = nA/(nA + nB), where n is amount of substance [3]. Given that the WMO scale for methane abundance is a mole fraction scale [4], the reporting of mole fractions would be desirable. While the use of the term 'concentration' for mole fraction is accepted for communication with the general public [4], a publication in a scientific journal should in my eyes favour exact terminology. In any case the Authors must make clear which quantity is reported.

> We apologize for the units confusion. We measured the dry mole fraction of methane (confirmed with our associate Dr. Martin Steinbacher from the Swiss Federal Labs for Materials Science and Technology, Empa), and we have replaced all other terms used to express methane units in the manuscript with "mole fraction" (defined in the first instance as dry mole fraction). We also now define ppm explicitly as μmol mol^{-1} for clarity.
>
> Mole fraction is now used throughout the text, and the explicit clarification that the reference measurements are reported as dry mole fractions is given in Section 2.2 when specifying the LGR instruments (lines 77–80 [79–82]). Moreover, we now use μmol mol^{-1} and nmol mol^{-1} in text and figures.

Two TGS 2600 sensors were deployed at the site, referenced to as #1 and #2. However, in several instances in the manuscript 'TGS 2600' appears without a number when I think TGS 2600 #1 is meant. Also, 'sensor', 'TGS' and 'TGS2600' are used. This should be made more consistent. Results from TGS 2600 #2 are presented exclusively in l. 104, ll. 185–186 and Fig. 9. Explaining the minor role of TGS 2600 #2 around l. 59, l. 104 or l. 162 might prevent confusion of the reader.

> The paper is revised to be more explicit about the minor role of TGS 2600 #2. With only one sensor in place, an acceptable critique would have been that one sensor could be far off in its readings. Hence the second sensor #2 was only there for redundancy, and using this sensor actually showed that sensor #1 on which our analysis is primarily based, could be backed up by another (randomly selected) sensor of the same type.

We added the following text in Section 2.2 (lines 60–63 [61–64]):
Sensor #1 is the primary sensor used in this study, whereas sensor #2 was only used as a replicate to simplify assessing potential problems with sensor #1. Because no such problems occurred, we will focus only on the results obtained with sensor #1 except for Section 2.2, where we used both sensors to assess their performance at weekly time resolution.

ll. 74–77 How often and to which scale were the reference analysers calibrated?

Note that this is not a GAW reference station, but a seasonal research camp where we were able to check the LGR every summer season. We'll add this information and the numbers of the NOAA reference gas bottles we used. Before we had NOAA reference gases available it turned out that the accuracy of LGR was better than the specification of the available reference gases for calibration at the research station.

We added the following text to lines 92–96 [95–99]:
Both FMA and the later FGGA analysers were used for eddy covariance applications, and thus the instruments were not calibrated as frequently as is done in applications for the Global Atmosphere Watch network (WMO, 2001). Both sensors were more accurate than the available calibration CH_4 gases at TFS. In 2015 for the first time it was possible to use a NOAA reference gas cylinder (#CB09827) to fine-tune the FGGA. This was typically done in the early summer season when field personnel arrived at TFS (late May).

l. 153 '[...] relative humidity (which is a ratio and not a physical variable of atmospheric water content)' – I think I have an idea of what the authors mean, but I find the wording not quite right. Would the authors say that the refractive index of a material is not a physical variable because it is a ratio? In general, I miss some thoughts about the temperature dependence of the quantity used for expressing humidity.

We eliminated this statement about the ratio (line 172 [180–181]). A statement addressing the thoughts on temperature depencence of humidity was added in lines 301–304 [325–328]):
Based on physical considerations one might expect that specific humidity or water vapor mixing ratio instead of absolute humidity could lead to further improvements, because absolute humidity still depends on temperature. However, our tests have not indicated a relevant gain of information or accuracy of prediction, but future work should also try to find a better physical correction model than the purely empirical one used here based on manufacturer information.

The Figaro TGS 2600 has a heated sensing element, so the relative humidity at the sensing surface is different from the relative humidity in the environment. The temperature dependence of relative humidity makes this quantity a less than ideal choice for this type of correction. Unfortunately, both alternative quantities chosen by the authors to express water vapour content depend on temperature as well. Mixing ratio (by mass) or specific humidity would be temperature-independent alternatives [6]. Using the ideal gas equation, the terms in Eq. 2 that contain the product Ta·ρ_v would also cancel out the temperature dependence of absolute humidity if Ta was absolute temperature (in K) – but in the manuscript a Celsius temperature is used. Hence, my suggestion to the authors is to try out either mixing ratio by mass or specific humidity as an independent variable in Eq. 2. Using absolute instead of Celsius temperature might be advisable as well.

We see the point, but one challenge is that such low-cost sensors are closer to a black-box system than to a fully controllable physical sensor. In addition, the manufacturer does not reveal all details of the sensor operation. In a conversation with Dr. Nick Martin, National Physics Laboratories, London, we learned that his lab has also explored the validity of using water concentration instead of relative humidity – in their laboratory tests they found an almost linear relationship between the two. We tried absolute humidity, vapor pressure, and relative humidity in our calculations, but absolute humidity always performed slightly better than the other two. We considered specific humidity and mixing ratio to be too close to absolute humidity and thus did not test this so far. Further tests could be run, but our approach is to adopt the same method as was used by previous authors (Collier-Oxandale et al., 2018; lines 258–260 in the discussion paper version).

We also understand the point that if we mix units of K or °C in the various equations (due to how the temperature dependence is expressed), we might introduce a constant bias (or at best just make the units messy). We can empirically test the question of whether temperature in K or °C gives the better results in our prediction models, and select the one that performs better for the revised version.

This is the parameter table from the linear model where temperature was entered in K instead of °C:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.5259	0.1021	44.31	<0.0001
Rs_R0	-0.4626	0.0247	-18.73	<0.0001
I(1/Rs_R0)	-0.5538	0.1189	-4.66	<0.0001
TA_AVG_M1B1_1.65_1_K	-0.0112	0.0004	-26.86	<0.0001
RHO_V_CALC_M1B1_1.65_1	-148.7784	5.7656	-25.80	<0.0001
Rs_R0:TA_AVG_M1B1_1.65_1_K	0.0019	0.0001	18.15	<0.0001
I(1/Rs_R0):TA_AVG_M1B1_1.65_1_K	0.0037	0.0005	7.96	<0.0001
Rs_R0:RHO_V_CALC_M1B1_1.65_1	26.9994	4.9153	5.49	<0.0001
I(1/Rs_R0):RHO_V_CALC_M1B1_1.65_1	-72.1729	15.4509	-4.67	<0.0001
Rs_R0:TA_AVG_M1B1_1.65_1_K:RHO_V_CALC_M1B1_1.65_1	0.1939	0.0213	9.10	<0.0001
I(1/Rs_R0):TA_AVG_M1B1_1.65_1_K:RHO_V_CALC_M1B1_1.65_1	0.4949	0.0442	11.21	<0.0001

And this is the model output where temperature was entered in K instead of °C **and** humidity was entered as mixing ratio instead of absolute humidity:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.6552	0.1005	46.30	<0.0001
Rs_R0	-0.4985	0.0239	-20.83	<0.0001
I(1/Rs_R0)	-0.7269	0.1176	-6.18	<0.0001
TA_AVG_M1B1_1.65_1_K	-0.0117	0.0004	-28.59	<0.0001
MR_CALC_M1B1_1.65_1	-113.5592	4.3851	-25.90	<0.0001
Rs_R0:TA_AVG_M1B1_1.65_1_K	0.0021	0.0001	20.24	<0.0001
I(1/Rs_R0):TA_AVG_M1B1_1.65_1_K	0.0043	0.0005	9.46	<0.0001
Rs_R0:MR_CALC_M1B1_1.65_1	42.1642	3.8241	11.03	<0.0001
I(1/Rs_R0):MR_CALC_M1B1_1.65_1	-71.0242	11.8599	-5.99	<0.0001
Rs_R0:TA_AVG_M1B1_1.65_1_K:MR_CALC_M1B1_1.65_1	0.0709	0.0164	4.31	<0.0001
I(1/Rs_R0):TA_AVG_M1B1_1.65_1_K:MR_CALC_M1B1_1.65_1	0.4354	0.0339	12.84	<0.0001

For comparison, the same output from the model that we use in Eq. (2) is copied in below:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.4611	0.0172	84.97	<0.0001
Rs_R0	0.0597	0.0046	13.03	<0.0001
I(1/Rs_R0)	0.4499	0.0162	27.82	<0.0001
TA_AVG_M1B1_1.65_1	-0.0112	0.0004	-26.86	<0.0001
RHO_V_CALC_M1B1_1.65_1	-148.7784	5.7656	-25.80	<0.0001
Rs_R0:TA_AVG_M1B1_1.65_1	0.0019	0.0001	18.15	<0.0001
I(1/Rs_R0):TA_AVG_M1B1_1.65_1	0.0037	0.0005	7.96	<0.0001
Rs_R0:RHO_V_CALC_M1B1_1.65_1	79.9518	1.8763	42.61	<0.0001
I(1/Rs_R0):RHO_V_CALC_M1B1_1.65_1	63.0215	4.5666	13.80	<0.0001
Rs_R0:TA_AVG_M1B1_1.65_1:RHO_V_CALC_M1B1_1.65_1	0.1939	0.0213	9.10	<0.0001
I(1/Rs_R0):TA_AVG_M1B1_1.65_1:RHO_V_CALC_M1B1_1.65_1	0.4949	0.0442	11.21	<0.0001

As expected the model performance in all aspects did not change when the unit of the temperature measurements was changed from °C to K, but of course the parameter estimates changed. There was, however, no indication that the empirical interaction term in Eq. (2) cancels out because of the change in units. The last three lines in each table copied in above show these interaction terms. The parameters are 0.1939 ± 0.0213 and 0.4949 ± 0.0442 for the models using ρ_v, and 0.0709 ± 0.0164 and 0.4354 ± 0.0339 for the model using mixing ratio instead of ρ_v (note that these parameter estimates do not change by changing the units used for temperature).

The overall model performances in a similar display as the one used in Table 1 in the manuscript looks like this:

	Eq. (2)		T_a in K, not °C		MR not ρ_v; T_a in K, not °C	
	Calibration	Validation	Calibration	Validation	Calibration	Validation
Overall						
R^2	0.447	0.207	0.447	0.207	0.446	0.208
RMSE (μmol mol^{-1})	0.026	0.041	0.026	0.041	0.026	0.041
Warm conditions ($T_a \geq 0°C$)						
R^2	0.518	0.288	0.518	0.288	0.519	0.289
RMSE (μmol mol^{-1})	0.026	0.032	0.026	0.032	0.026	0.032
Cold conditions ($T_a < 0°C$)						
R^2	0.345	0.034	0.345	0.034	0.343	0.033
RMSE (μmol mol^{-1})	0.027	0.052	0.027	0.052	0.027	0.052

Thus, using mixing ratio results in a marginally lower R^2 overall during the calibration period; although at temperatures above freezing mixing ratio is marginally performing better than ρ_v, the R^2 at temperatures below freezing is lower. Overall, this exercise confirms that the key physical processes

affecting the TGS 2600 are not well understood and hence we decided to keep using ρ_v. However, we added the following text to inform readers about this issue (lines 301–304 [325–328]):

Based on physical considerations one might expect that specific humidity or water vapor mixing ratio instead of absolute humidity could lead to further improvements, because absolute humidity still depends on temperature. However, our tests have not indicated a relevant gain of information or accuracy of prediction, but future work should also try to find a better physical correction model than the purely empirical one used here based on (incomplete) manufacturer information.

ll. 161–162 Using the entire dataset for estimating the parameters of Eq. 2 is a comprehensive test of how well the model can describe the dataset, but is of limited relevance for field deployments where calibrations are performed during limited periods of time and the main interest is in the uncertainty of independent measurements. For this reason, splitting the dataset into a calibration and a validation part yields important insights. The caption of Table 1 explains that the authors have in fact performed analyses of a split data set. This fact should also be mentioned in the main text around the lines given.

This point was also addressed by Reviewer 2. We will thus recalculate the ANN with the same selection for training vs. validation as we used for the linear model.

We did the reanalysis and this solved the issue with non-comparable presentations of goodness of fits due to different choices in calibration vs. validation periods. When done with the same selected periods the R^2 values of the ANN are lower during the calibration periods than our linear model (see updated Table 1), but are slightly higher during the validation period, except for warm conditions with temperatures above freezing.

l. 167 For the reasons given before, the results presented in the columns 'Linear Model - Calibration' and 'Linear Model - Validation' in Table 1 should be discussed here, even more so because the results for the validation period are substantially worse than for the period used for calibration.

This will be done.

This is where the ANN has a more even performance between calibration and validation period, although its performance in the calibration period was lower than was the performance of the linear model. Positively phrased one could say that the linear model makes more of the data that are available, but ANN learns more from the available data to make a prediction with validation data. Depending on what the aim is, both have benefits, and thus we added the following text (lines 205–209 [214–218]):
When testing the linear model approach (Eq. 2) more rigorously by splitting the available data into a calibration period (years 2014–2016) and a validation period (years 2012–2013 and 2017–2018), some limitations can be seen, in particular under cold conditions where none of the approaches performed very well in the validation period. The ANN had a more balanced performance between calibration and validation period, although it performed slightly less well under warm conditions ($T_a \geq 0\,°C$).

ll. 193–194 Is there any conclusion that can be drawn from this finding of a -1:1 relationship?

The conclusion should be that both variability and stability can be improved at the same time because there is no tradeoff visible in Fig. 11. We will add text about this in the revised manuscript version.

We added the following text on lines 246–247 [255–256]:
This indicates that both variability and stability can be improved at the same time because there is no tradeoff visible in Fig. 11.

Sect. 3.4 The discussion in Sect. 3.1 leans heavily on the coefficient of determination R2. In Table 1, each R2 for the ANN approach is higher than the corresponding R2 for the Linear Model. Considering just the validation period, the ANN approach outperforms the Linear Model by a factor of 3–10 by this measure. Similarly, in Fig. 6 and 8 the ANN approach outperforms the Linear Model (comparing R2 of 'ANN' and 'c/v', 'all' is irrelevant in this respect); in Fig. 5 and 7 they perform nearly equally well. None of these comparisons is made here. Instead, the authors state that the root mean square error (RMSE) does not improve substantially with the ANN approach. While I generally appreciate the reporting of RMSE together with R2, its interpretation here is questionable.

The reviewer is right, R^2 in an ANN is not the same as R^2 in a linear model and thus comparing the numbers may be misleading. We will revise this and provide a goodness-of-fit statistic that is comparable between a linear model and ANN. RMSE is of course the better comparison because it has the same meaning in cases where either linear or nonlinear fits are used and evaluated.

We homogenized the computation of the R^2 values for both methods, which however was not the primary issue here; the biggest change was related to the different selection of calibration vs. validation periods. How the values given in Table 1 are directly comparable and relate to the same selection of calibration vs. validation period. The text on lines 159–160 [166–168] was updated accordingly.

On the one hand, RMSE is reported in Table 1 with one significant digit only, potentially masking up to ∼30% differences for an RMSE of 0.03 ppm (0.025 ppm vs. 0.0349 ppm). On the other hand, the RMSE should be seen in the context of the variability of the data, specifically the root mean square difference between the reference measurements and their mean value over the whole dataset, which is not stated. Overall, the discussion in this section appears negatively biased with regard to the ANN approach. This also manifests in the last paragraph of this section, where 'understand[ing] the physical response of TGS sensors' is prioritised over 'technically nicer fits to data', a stark contrast to the lack of a physical interpretation of the terms in the empirical model (Eq. 2). Section 3.4 must be revised to reach the level of neutrality expected from a scientific publication.

Thank you for these suggestions; we will remove the wording used in the conclusion about neural networks, and rethink how we can best compare the two models both with unbiased estimators of their fit to the data and with their usefulness in interpreting the data in mind. We will also adjust the significant digits of the measures we use (e.g., RMSE) to make them more comparable and useful.

We now show one additional digit in both R^2 and RMSE values. We also added the extra digit in the R^2 lines to allow the reviewer to compare the minor differences between some values that would otherwise be hidden in the rounding to the second digit. In Figures 5–8 we added RMSE below R^2.

Section 3.4 was revised to be more neutral and reflect the conditions that when the same calibration and validation periods are used with both approaches, then the apparent outperformance of ANN is less obvious. We agree that for many other applications there is a high potential of using an ANN but not really in this context where we try to separate the influence of T_a and ρ_v on the sensor signal, while at the same time of course T_a and ρ_v influence the production of methane in anaerobic soils and thus modify the CH_4 mole fractions to be measured in ambient air.

We thus rewrote Section 3.4 accordingly (see lines 274–289 [284–312]).

Please take note that in the previous submission we had used ca. five years of calibration period vs. only 1.5 year of validation for the ANN, whereas we used the same setting as in the revised version for the linear model (3 calibration years vs 4 validation years). Please be aware that while we have noticed good performance of ANNs in other applications, we do have some reservations for applications as the one presented here.

ll. 229–231 To make such an argument, the reader must be informed about the amplitude of all input variables, especially SC.

In light of the comments above we will reword or delete this statement.

We deleted this statement since with reduced coverage of winter conditions by the calibration period (to make the linear model and ANN approaches comparable) the ANN performance deteriorated substantially.

l. 237 I might be mistaken, but as far as I understand the term homoscedasticity it would in this case mean that the variance of the deviation in CH4 abundance is the same for every temperature bin. The authors do not report variances, but both interquartile range and 95% confidence interval suggest that the variance is higher at low temperatures than at high temperatures, i.e. heteroscedasticity.

What we wanted to say is that the variances (shown with the 95% CI) are quite constant over a wide range of temperatures and absolute humidity. We of course agree that if one focuses at the difference between one bin and another there are many comparisons where the statement is not 100% correct, even when the Bonferroni correction for multiple testing is considered. In comparisons that span a difference of more than 10–15 °C, the within-bin variances are of course not always homoscedastic. We will therefore reword our description of this point to avoid this confusion.

The new wording (lines 291–293 [314–317]) is:
The inter-quartile ranges and 95% confidence intervals of each air temperature (Fig. 12a) or absolute humidity bin (Fig. 12b) are very similar over a wide range of temperature and humidity, but tend to become more variable in bins with few data (i.e., lowest and highest temperatures, and highest absolute humidities in Fig. 12).

ll. 253–257 ‘laboratory conditions simplify the real world too much’ – What could be the simplification that makes laboratory calibrations problematic? The input variables used in the empirical model (Eq. 2) can – practical difficulties taken aside – be controlled in the lab. Any other variable that might prevent transfer of lab results to field conditions is not included in the empirical model, so the problem would not be a simplification of the lab environment but a model deficiency. The last sentence of the paragraph seems to go in this direction (‘relevant factors’), but is unclear. Please explain better or leave out.

> This statement also caught the attention of Reviewer #2, and given both reviewer comments we will revise the text. As the reviewer says, there are several ways in which laboratory and real world conditions apply to sensor calibrations. For example, in our revision we will emphasize that testing in the temperature and moisture conditions that exist in the range from –40 °C to 0 °C is not easy in a laboratory environment. In addition, we have discussed this general point with Dr. Nick Martin, National Physics Laboratories, London, and will adopt some of his perspective and advice in our revised text.
>
> We revised the text to read (lines 313–320 [338–351]):
> *While this is theoretically correct, it remains difficult to carry out laboratory treatments from –41 °C to 27 °C as would be required for our Arctic site. The data we present indicate that most likely it is absolute humidity (or specific humidity or mixing ratio), not relative humidity, that should be used for such calibrations, which in principle should provide the best quality results if the relevant factors are known and can be included in the calibration set-up. It would be desirable that manufacturers carry out both laboratory tests and field trials and provide the necessary correction functions together with sensors. However, due to the expense and time it takes to carry out long tests, be it in the laboratory or in the field, the present development goes in the direction of collocation studies en route to certification of sensors (N. Martin, NPL, UK, pers. comm.), similar to what we have done in the Arctic.*

ll. 258–261 Suggesting to move the first sentence to l. 150 and to remove the other one (repetition).

> This will be done.
>
> The second sentence was deleted [344–348], and the first was moved to lines 168–170 [177–178].

Fig. 5–8 The graphs are squeezed in horizontal direction, making comparisons between the lines difficult. A shorter period, e.g. 14 days, would give more insight.

> We agree that the graphs are squeezed, but we are concerned that expanding them horizontally would invite comparisons over shorter time periods that are not the focus of the study. For example, on lines 179–181 we wrote “The TGS 2600 is not expected to provide short-term accuracy comparable to high-quality instrumentation (see also Lewis et al., 2018). However, Eugster and Kling (2012) argued that such measurements still may provide additional insights as compared to the passive samplers described by Godbout et al. (2006a, b), integrating over longer time frames”. Given that, we tried to draw the attention of the reader to the fact that such a low-cost sensor is not expected to provide high accuracy over the short-term. Zooming in would thus highlight the opposite of what we consider meaningful in this study. However, we will publish the original data along with our paper (we will however select a different repository than the original mentioned under “Data availability”), which will allow the interested reader to download the data and zoom in as desired.
>
> The data set will be made available via DOI:10.6073/pasta/dddeb05b2806e2f5788fadd6fc590ef1, and the fits shown in Figures 5–8 will be made available via DOI:10.3929/ethz-b-000369689. The text under “Data availability” was updated accordingly. Note that the second DOI will only be published after acceptance of the manuscript.

Fig. 5 and 7 The collected in 2012 and 2015 are both part of the calibration period, not the validation period, which is important to know for the reader to correctly interpret ‘TGS 2600 c/v’ and ‘ANN’. I therefore strongly suggest a note in the figure caption.

> This information will be added.
>
> We added the information *This example belongs to the validation period of the TGS 2600 c/v and ANN fits* to the captions of Fig. 5–8. Note that only the ANN in the previous version was using 2012 and 2015 for calibration. In the revised version both approaches show validation period results in all four figures. Thanks for pointing this out.

Fig. 5 Suggesting to replace '(TGS 2600 - Reference)' with '(TGS 2600 all - Reference)' in the caption

This will be done.

Done.

Fig. 9 A plot of the difference of the methane abundance calculated from the measurements of the two sensors would be of high interest for the readers. With such a new panel it is also important to state if the parameters derived for TGS 2600 #1 have been used when applying Eq. 2 to the measurements of TGS 2600 #2. In my opinion the new panel could replace panel (b), as the signal difference seems of little relevance.

We will replace panel (b) that currently shows the relative differences with a new panel showing the absolute differences as suggested by the reviewer. And we will more clearly state that we used the parameters derived for sensor #1 also for sensor #2 (as correctly interpreted by this reviewer).

We replaced panel (b) but then rearranged (b) and (c), so that the absolute difference between the two sensors is now shown in panel (c). And we added the information *The signals from both sensors were converted to CH_4 using Eq. (2) parameterized with data from TGS sensor #1* to the caption.

Fig. 10 If the main text in ll. 189–190 is correct, 'and the reference' is missing at the end of the first sentence of the figure caption.

Yes, this will be added.

Done.

3 Technical corrections

l. 1 Suggesting to remove "weak" to avoid misunderstanding. Alternatively, it could be written in parentheses like it the conclusions.
Done.

l. 8 Insert a space between value and unit of temperature. This correction is necessary wherever '°C' is used [5].
Done.

l. 76 replace**d**
Done.

l. 140 typeset 'Ta' as Ta
We assume this relates to former line 149, there we found an occurrence of Ta that should be T_a which we corrected.

l. 305 'cross-sensitivities [...] **are** of no concern'
Done.

Fig. 2 There seems to be a non-displayable glyph at the beginning of the label for the vertical axis, possibly a Δ. This is also the case in Fig. 3, 5 and 12.
We could not reproduce this problem in Acrobat or Preview on Mac. Could it be an issue with the PDF reader you're using? If you can specify with which reader you use to see this problem, then we will double-check at print stage to make sure the PDF has all fonts correctly embedded. Please also specify on which operating system you see this problem (Windows, Linux, iOS, Android, MacOS, ...).

'CH4' specifies a substance, not a quantity. Use 'x_{CH4}' or another appropriate quantity symbol. The same applies to Fig. 3 through 9 and Fig. 12.
We now use χ_{CH_4} in all figures as suggested by this reviewer. Please use Acrobat Reader for double-checking, because if Δ is not correctly seen in your tool, then most likely the same might be the problem with χ. But we'll solve this issue at print stage if we can reproduce the problem. Normally this is a mismatch of PDF file type versions. Our figures use PDF version 1.4, but this could be changed if needed (although AMT also uses PDF version 1.4 for their file versions, which is archivable PDF of type PDF/A-1 and thus should be correctly displayed in all readers, hence our question for the details to reproduce the problem observed by this reviewer).

These will be incorporated in the revised manuscript, thank you for highlighting these items.

4 References

[1] IUPAC Gold Book, term ‘concentration’, https://goldbook.iupac.org/terms/view/C01222
[2] IUPAC Gold Book, term ’mixing ratio’, https://goldbook.iupac.org/terms/view/M03948
[3] IUPAC Gold Book, term ’amount fraction’, https://goldbook.iupac.org/terms/view/A00296
[4] GAW Report No. 242 19th WMO/IAEA Meeting on Carbon Dioxide, Other Greenhouse Gases and Related Measurement Techniques (GGMT-2017), https://library.wmo.int/doc_num.php?explnum_id=5456
[5] SI Brochure: The International System of Units (SI), https://www.bipm.org/en/publications/si-brochure/
[6] WMO Guide to Meteorological Instruments and Methods of Observation, https://library.wmo.int/doc_num.php?explnum_id=4147

Response to Reviewer #2

This manuscript presents the results from a field deployment of a pair of low-cost metaloxide sensors. The sensors were co-located with a reference instrument, allowing the researchers to train various calibration models to predict methane concentrations. These calibration models relied on the signals from the low-cost sensors as well as other sensors (i.e., temperature and humidity). Researchers then assessed the performance of and potential for these sensors using the predicted signals.

This manuscript is especially relevant to the field of low-cost sensor research and readers of Atmospheric Measurement Techniques for two reasons: (1) it provides an example of a long-term (multi-year) field deployment of low-cost metal-oxide sensors, and (2) it provides an example of VOC sensors deployed to predict ambient methane levels - two areas that would benefit from further study. Furthermore, the deployment of the sensors in a remote area with little potential for the presence of confounding pollutants provides useful information on the potential ability of this sensor to be used for methane detection. Though a few revisions (listed below) are recommended prior to publication.

> Thank you for your very supportive assessment. Your comments are very valuable for us to improve clarity and add an interesting aspect about the sensitivity of the TGS 2600 to CO from wildfires.

1. Please clarify throughout whether the results for the linear model being discussed in the text are based on the model that was fitted to the complete data set or the model which was fitted to the shorter training data set.

> This will be done.

> We added the following text to clarify (lines 186–189 [195–198]):
> *Unless explicitly mentioned, we analysed CH_4 mole fractions computed with Eq. (2) using the parameters obtained from all data measured by TGS 2600 sensor #1. Only in the direct comparison with the ANN (Section 3.1) did we determine an additional parameter set using the same calibration period as the ANN used, so that a direct comparison of performance in validation was possible.*

Additionally, the training and testing periods defined for the linear model (in Table 1) and for the ANN (in the end of Section 2.4) appear to be different. Could the authors comment on the rationale for this choice and whether the use of these different periods might affect the comparability of the results for these two models presented in Table 1?

> We will recalculate the ANN to match the same data selection as we used for the linear model.

> Done. See also our response to reviewer #1 on this issue.

2. In Section 2.3, please provide information on any additional processing of the sensor data that may have occurred (e.g., filtering outliers, or removing sensor “warm-up” periods), or state that the data did not undergo additional filtering or processing.

> This will be done.

> We added the following text (lines 103–110 [106–113]):
> *Before analyses the data were processed in the following way: (1) outliers were removed; (2) relative humidities $> 105\%$ (accuracy of capacitive humidity sensors) were deleted; (3) reference CH_4 mole fractions obtained from the FGGA (since 2016) were filtered based on hard boundaries of housekeeping variables available for quality control. For the latter we used the following hard boundaries for filtering: (a) sample cell pressure had to be in the range 130–143 Torr; (b) the instrument-specific*

ringdown time of the laser for CH_4 measurements had to be in the range 13–17 ms. The accepted reference CH_4 mole fractions were thus all measured in the narrow range of cell pressures between 139.7 and 140.3 Torr and laser ringdown times between 14.02 and 14.94 ms, which indicates best performance of the analyser. Before 2016 (FMA instrument) these house-keeping variables were not recorded.

3. Suggest moving the description of the motivation and development of the model for heat loss to an earlier point in the manuscript (e.g., after the description of the linear model in Section 3.0). This would assist the reader in their interpretation of the results in Table 1. Though the discussion of how this approach could be improved should remain in Section 3.5.

This can be done.

We moved former lines 268–278 to Section 3 (now lines 192–202 [201–211]) with an additional introductory sentence: *If ambient temperature influences the signal of the TGS 2600 in such a way as expected from the technical documentation (Figaro, 2005b,a), then wind speed could be a third factor influencing the conversion from TGS 2600 sensor voltages to CH_4 mole fractions.*

In the gap from where we moved the text away we inserted (lines 327–328 [358–368]):
To address this additional factor, we used the heat loss model given in Eq. (3).

4. Could the authors provide additional information or discuss how the parameters of the model were selected (Eq. 2), for example, did this model yield substantial improvements over a simpler linear model?

We used the stepAIC function of the MASS package in R. We will add the details to the text.

We added the text (lines 184–186 [193–195]):
The linear model in Eq. (2) was derived from a suite of candidate models including interactions among predictors and including quadratic terms of each variable, and then stepwise elimination using the stepAIC function in the MASS package of R was employed to find the model with the lowest AIC (Akaike's Information Criterion).

5. Suggest expanding on the point made in Section 3.5 (Lines 253-254) to explain in what ways laboratory conditions over-simplify real-world conditions. This observation has been demonstrated in other studies [1, 2] and it could be valuable to highlight the challenges that may be associated with laboratory calibrations of sensors for this particular application.

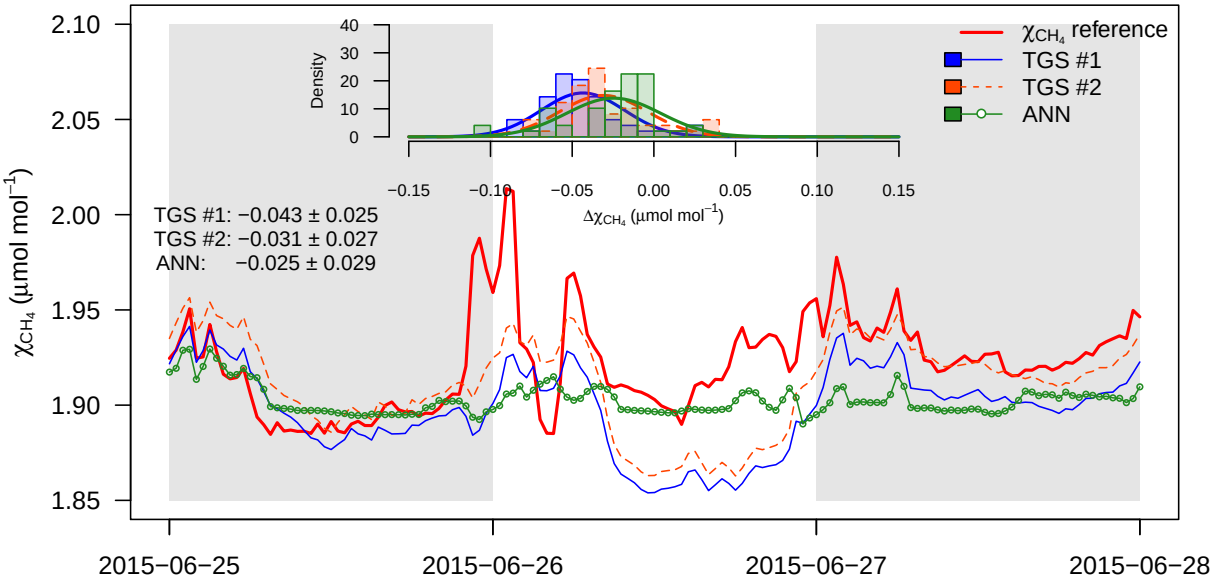
This aspect also caught the attention of Reviewer #1. We will thus revise the text, in part with advice from Dr. Nick Martin, National Physics Laboratories, London on this topic. In addition we will emphasize that testing in the temperature and moisture conditions in the range from –40 °C to 0 °C is not easy in a laboratory environment.

Please see our response to the same point of Reviewer #1.

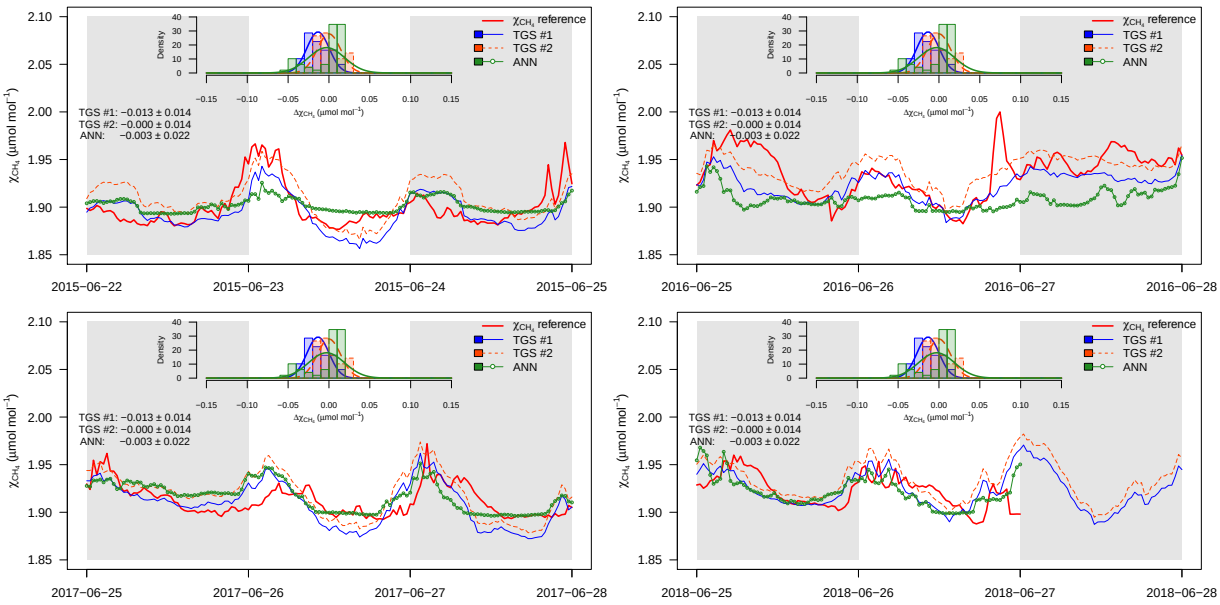
6. Could the authors provide additional detail on the potential or likelihood for confounding pollutants, in particular carbon monoxide (Section 2.2)? For example, are there any towns nearby where emissions from wintertime heating may be a concern, or did any major wildfires occur in the area throughout the deployment period?

We will add to the text that the study site is in such a remote site that winter time heating negligible, certainly not what one would expect from an urbanized area. However, wildfire influences in summer time (from fires in forests far to the south of our site and on the south side of the Brooks Range) may produce high CO levels that would lead to apparent high CH4 mole fractions. We will add an extra analysis of an episode where smoke from a wild fire south of the Brooks Range mountains was present at Toolik Field Station according to our own records, and compare that period with the period immediately before the smoke arrived. We can also compare from similar weeks in the year before and year after when the fire occurred.

We have written records from the smell of smoke from wildfires south of the Brooks Range that affected the measurements on 26 June 2015. Thus, we carried out a case study analysis for that day and show the day of interest (white background) with the day before and the day after (gray background):



As an insert we plotted the histograms of the difference between the two TGS-derived CH_4 mole fractions and the ANN fit. Then, we carried out the same analyses for four other cases: (1) three days before the wildfire event, and (2–4) the same date one, two and three years later:



The general impression is that the TGS-derived CH_4 mole fractions are roughly 0.03 μmol mol^{-1} reduced by the increased carbon monoxide mole fractions and other air pollutions during the second part of the day. The mean differences ± standard deviations are given in the left part of each panel. This offset does not differ between TGS #1 and TGS #2, but it should be considered a coincidence that using Eq. (2) that was parameterized with data from TGS #1 performs better with TGS #2 in all comparison case studies. The variability of the difference also appears to be slightly larger (±0.025 to ±0.027 on 26 June 2015 vs. ±0.014 in all four comparisons).

Thus, in the text we added the following information (lines 221–228 [230–237]):
Because of the absence of local sources of carbon monoxide and other air pollutants to which the TGS 2600 sensor is also sensitive besides CH_4, we investigated a special case when smoke and haze from wildfires south of the Books Range polluted the air in the TFS area on 26 June 2015 and compared the performance of both TGS sensors during that day with conditions three days before that event, and on the same date in the following three years. The net effect of increased air pollutants was an apparent small decrease of the CH_4 mole fractions calculated via Eq. (2) by approximately –0.03 μmol mol^{-1}. At the same time the variability of the residuals increased from typically ±0.014

to ±0.027 µmol mol^{-1} (24-hour averages). Thus, the influence of the wildfire smoke was of the same order of magnitude as the difference between TGS-derived CH_4 mole fractions and the reference instrument on most other days of the year (see Figs. 5–8).

7. Is there any concern that the temperature/humidity sensor described in Section 2.2 might itself experience any issues with drift or aging over such a long field deployment?

Any sensor might be subject to drifting and aging. What we can do in our revisions is to compare our dedicated temperature and relative humidity sensor with the reference sensor of the long-term weather station at the same site.

The HMP45AC sensor that we used for the long-term temperature and relative humidity measurements has a sensor head that can be exchanged against a calibrated one to avoid long-term drifts in the measurements. We exchanged the head every ca. 3 years to minimise long-term drift errors of our temperature and relative humidity measurements. We added this information to the corresponding text (lines 87–89 [91–92]:
The factory-calibrated HMP45AC sensor head was exchanged against a newly calibrated one every ca. 3 years to minimize long-term drift effects in temperature and relative humidity measurements (J. Laundre, pers. comm.).

We initially thought that a comparison with the temperature and relative humidity sensors of the long-term weather station at TFS (roughly 0.5 km to the north-east) would be the way to go, and hence our Final Author Response (in gray above). But when comparing the data we realized that the microclimatic difference between TFS close to Toolik Lake and our inland site at some distance from the lake shore and at higher elevation in the wet tundra show important seasonal trends that rather look like real differences, and thus do not really allow to test for drifts. Thus, having replaced the calibrated sensor head every ca. 3 years is probably the best we could do to minimize concerns about sensor drift.

8. Line 38: add an 's', "assessment of low-cost sensor[s]" 9. Line 66: delete 'e.g.,', "in an area like e.g., the arctic" 10. Line 246-247: change the color of the red text to black 11. Line 254: delete 'it', "as it would be required"

These minor changes will be applied as suggested.

Done.

References

[1] Castell, N., Dauge, F., Schneider, P., Vogt, M., Lerner, U., Fishbain, B., . . . Bartonova, A. (2017). Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates? Environment International, 99, 293-302.
[2] Piedrahita, R., Xiang, Y., Masson, N., Ortega, J., Collier, A., Jiang, Y., . . . Shang, L. (2014). The next generation of low-cost personal air quality sensors for quantitative exposure monitoring. Atmospheric Measurement Techniques Discussions, 7(2), 2425- 2457.

Thanks for these references which were incorporated in the text.

Long-term reliability of the Figaro TGS 2600 solid-state methane sensor under low Arctic conditions at Toolik ~~lake~~Lake, Alaska

Werner Eugster[1], James Laundre[2], Jon Eugster[3,4], and George W. Kling[5]

[1]ETH Zurich, Department of Environmental Systems Science, Institute of Agricultural Sciences, Universitätstrasse 2, 8092 Zurich, Switzerland
[2]The Ecosystem Center, Marine Biology Laboratory, Woods Hole, MA 02543, USA
[3]University of Zurich, Institute of Mathematics, Winterthurerstrasse 190, 8057 Zurich, Switzerland
[4]now at: School of Mathematics, The University of Edinburgh
[5]University of Michigan, Department of Ecology & Evolutionary Biology, Ann Arbor, MI 48109-1085, USA

Correspondence: Werner Eugster (eugsterw@ethz.ch)

Abstract. The TGS 2600 was the first low-cost solid state sensor that shows a ~~weak~~ response to ambient levels of CH_4 (e.g., range ≈1.8–2.7 ~~ppm~~μmol mol^{-1}). Here we present an empirical function to correct the TGS 2600 signal for temperature and (absolute) humidity effects and address the long-term reliability of two identical sensors deployed from 2012 to 2018. We assess the performance of the sensors at 30-minute resolution and aggregated to weekly medians. Over the entire period the agreement between TGS-derived and reference CH_4 ~~concentrations~~ mole fractions measured by a high-precision Los Gatos Research instrument was R^2 = 0.42, with better results during summer (R^2 = 0.65 in summer 2012). Using absolute instead of relative humidity for the correction of the TGS 2600 sensor signals reduced the typical deviation from the reference to less than ±0.1 ~~ppm~~ μmol mol^{-1} over the full range of temperatures from –41 °C to 27 °C. At weekly resolution the two sensors showed a downward drift of signal voltages indicating that after 10–13 years a TGS 2600 may have reached its end of life. While the true trend in CH_4 ~~concentrations~~ mole fractions measured by the high-quality reference instrument was 10.1 ~~ppb~~ nmol mol^{-1} yr^{-1} (2012–2018), part of the downward trend in sensor signal (ca. 40–60%) may be due to the increase in CH_4 ~~concentration~~mole fraction, because the sensor voltage decreases with increasing CH_4 ~~concentration~~mole fraction. Weekly median diel cycles tend to agree surprisingly well between the TGS 2600 and reference measurements during the snow-free season, but in winter the agreement is lower. We suggest developing separate functions for deducing CH_4 ~~concentrations~~ mole fractions from TGS 2600 measurements under cold and warm conditions. We conclude that the TGS 2600 sensor can provide data of research-grade quality if it is adequately calibrated and placed in a suitable environment where cross-sensitivities to gases other than CH_4 ~~is~~ are of no concern.

Keywords: Toolik ~~lake~~Lake, Northern Alaska, Methane, Trace gas sensor, SnO2 sensor, Leak detector.

1 Introduction

Low-cost trace gas sensors open new deployment opportunities for environmental observations. Still, their long-term performance in real-world applications is largely unknown, and thus, scientific research with such low-cost sensors is challenged with

a high risk of failure and questionable data quality. Hence low-cost sensors are only considered as a complementary source of information on air quality ~~(e.g. Lewis et al., 2018)~~(e.g. Lewis et al., 2018; Castell et al., 2017). Here we report on a ~~six-year~~ seven-year (2012–2018) deployment of two low-cost Figaro TGS 2600 methane (CH_4) sensors during summer and winter conditions in the relatively harsh low Arctic climate of northern Alaska to explore the long-term stability and reliability of CH_4 ~~concentration~~ mole fraction estimates. The sensors were previously deployed over Toolik ~~lake~~ Lake during the ice-free season 2011 (Eugster and Kling, 2012), where similar values between TGS-derived and reference CH_4 ~~concentrations~~ mole fractions were only found if measurements were integrated over at least 6 hours, or if they were aggregated to mean diel cycles over the season. Other studies have deployed the same sensor type in complex rural and urban environments along the Colorado Front Range (Collier-Oxandale et al., 2018), in an oil and gas production region (Greeley, Colorado; Casey et al., 2019), and in urban South Los Angeles (Shamasunder et al., 2018). An application on an unmanned aerial vehicle, however, did not successfully detect CH_4 hotspots (Falabella et al., 2018). These are all pioneering studies, but mostly restricted to a few days to months of measurements. Thus, our study is the first long-term comparison of high-precision measurements to those from CH_4 sensitive, low-cost sensors under challenging climatic conditions.

Typically, new sensors are first calibrated under controlled conditions in a laboratory environment. Extensive calibration tests with a similar low-cost sensor (Figaro TGS2611-E00) from the same manufacturer as our TGS 2600 have been carried out by van den Bossche et al. (2017). Despite the care taken in their calibration effort, the residual CH_4 ~~concentration~~ mole fraction after calibration was still on the order of ± 1.7 ~~ppm~~μmol mol^{-1}, which is acceptable for chamber flux measurements, e.g. over water (as e.g., done by Duc et al., 2019), but not sufficient to measure ambient atmospheric ~~concentrations~~mole fractions, which are of the same order of magnitude as the calibration uncertainty. The issue of important differences between laboratory assessments of low-cost ~~sensor~~ sensors and their real-world performance is well known and typically relates to different data correction and calibration approaches in real-world than laboratory applications (Lewis et al., 2018). Hence we decided to use outdoor measurements obtained over a wide range of temperatures and relative humidity – the major cross-sensitivities experienced by such sensors – and derive a calibration function via parameter extraction using this data set. Our goals were thus to (1) establish a statistical calibration function from field measured conditions that can also be used in different contexts to linearize the TGS 2600 sensor signal (which then can still be fine-tuned with a two-point calibration in a specific application); (2) assess the reliability of the TGS 2600 low-cost sensor under winter and summer conditions in the Arctic over seven years of continuous deployment; and (3) explore potential improvements for sensor data processing, which includes (3a) wind effects that are neglected in laboratory environments, and (3b) artificial neural networks (ANN) to find out whether results can be improved over standard statistical regression methods for calibration of the sensor.

2 Material and methods

2.1 Study site

Field measurements were carried out at the Toolik wet sedge site (TWE, 68°37′27.62″ N, 149°36′08.10″ W, 728.14 m elevation, WGS 84 datum) where seasonal eddy covariance flux measurements were carried out during the summer seasons of

2010–2015 and partially during winters starting in 2014 until 15 June 2016, with the continuation as a meteorological station until present. The site is a wetland that is a local source of CH_4 with a flux rate that is roughly one order of magnitude stronger than adjacent Toolik ~~lake~~Lake, where the Eugster and Kling (2012) study was performed. The site is a wet graminoid tundra dominated by sedge species, namely cotton grass (*Eriophorum angustifolium*) and *Carex aquatilis* (Walker and Everett, 1991).

2.2 Instrumentation and measurements

Two Figaro TGS 2600 sensors ~~(Figaro, 2005b, a)~~ (Figaro, 2005a, b) that were already deployed over Toolik ~~lake~~ Lake (TOL) during the ice-free season 2011 (Eugster and Kling, 2012) were installed at the TWE site in late June 2012 (Fig. 1). Sensor #1 is the primary sensor used in this study, whereas sensor #2 was only used as a replicate to simplify assessing potential problems with sensor #1. Because no such problems occurred, we will focus only on the results obtained with sensor #1 except for Section 3.2, where we used both sensors to assess their performance at weekly time resolution. The TGS 2600 is a high-sensitivity solid-state sensor for the detection of air contaminants (Figaro, 2005a). It is sensitive to methane at low ~~concentrations~~mole fractions, but also to hydrogen, carbon monoxide, iso-butane, and ethanol. It is the only low-cost solid-state sensor that we are aware of for which the manufacturer indicates a sensitivity to methane even under ambient (≈2 ~~ppm) methane concentrations~~μmol mol^{-1}) methane mole fractions, whereas most other sensors are only sensitive at ~~concentrations~~ mole fractions that exceed ambient levels by at least one or two orders of magnitude. This high sensitivity to low methane ~~concentrations~~ mole fractions comes at the expense that no specific molecular filter prevents the other components to reach the sensor surface. Thus, our considerations made here assume that deployment is made in an area like ~~e.g.,~~ the Arctic where levels of carbon monoxide, iso-butane, ethanol and hydrogen are rather constant and do not vary as strongly as does methane, so that the sensor signal can be interpreted as a first approximation as a methane ~~concentration~~ mole fraction signal. For additional details on the TGS 2600 sensor the reader is referred to Eugster and Kling (2012).

The TWE site receives line power from the Toolik Field Station (TFS) power generator. During the snow and ice-free summer season (typically late June to mid-August) measurements are almost interruption free, but during the cold season (typically September to late May) longer power interruptions limit the winter data coverage. Nevertheless, this is the first study that provides low-cost sensor methane ~~concentration~~ mole fraction measurements over a temperature range from Arctic winter temperatures of –41 °C and to a relatively balmy 27 °C during short periods of the Arctic summer. Reference CH_4 ~~concentrations~~ dry mole fractions were measured by a Fast Methane Analyzer (FMA, Los Gatos Research, Inc., San Jose, CA, USA; years 2012–2016) which was ~~replace~~ replaced by a Fast Greenhouse Gas Analyzer for combined CH_4/CO_2/H_2O ~~concentration~~ dry mole fraction measurements (FGGA, Los Gatos Research, Inc., San Jose, CA, U.S.A.; since 2016). Until 18 June 2016 the CH_4 ~~concentrations~~ mole fractions were calculated as 1-minute averages from the raw eddy covariance flux data files. We report all gas ~~concentrations in mixing ratios by volume (ppm or ppb dry mole fractions)~~mole fractions in μmol mol^{-1} or nmol mol^{-1}. The FMA and FGGA sampling rate was set to 20 Hz, and the flow rate of sample air was ca. 20 L min^{-1}. After the termination of eddy covariance flux measurements, the FGGA measurements were continued with the instrument's internal pump (flow rate ca. 0.65 L min^{-1}) with 1 Hz raw data sampling. In addition to digital recording, the CH_4 signal was converted to an analog voltage that was recorded on a CR23X data logger (Campbell Scientific Inc. [CSI],

Logan, UT, USA). The same data logger also recorded air temperature and relative humidity (HMP45AC, CSI), wind speed and wind direction (034B Windset, MetOne, Grants Pass, OR, USA), plus ancillary meteorological and soil variables not used in this study. The factory-calibrated HMP45AC sensor head was exchanged against a newly calibrated one every ca. 3 years to minimize long-term drift effects in temperature and relative humidity measurements (J. Laundre, pers. comm.). Sensors were measured every 5 seconds and 1-minute averages were stored on the logger. These data were then screened for outliers and instrumental errors and failures, and 30-minute averages were calculated for the present analysis.

Both FMA and the later FGGA analysers were used for eddy covariance applications, and thus the instruments were not calibrated as frequently as is done in applications for the Global Atmosphere Watch network (WMO, 2001). Both sensors were more accurate than the available calibration CH_4 gases at TFS. In 2015 for the first time it was possible to use a NOAA reference gas cylinder (#CB09827) to fine-tune the FGGA. This was typically done in the early summer season when field personnel arrived at TFS (late May).

Because the TGS 2600 sensors only show a weak response to CH_4, but are highly sensitive to temperature and humidity, a LinPicco A05 Basic sensor (IST Innovative Sensor Technology, Wattwil, Switzerland) was added next to the TGS 2600 (see Fig. 1). The A05 is a capacitive humidity module that also has a Pt1000 platinum 1 kΩ thermistor on board to measure ambient temperature. The relative humidity output by the A05 is a linearized voltage in the range 0–5 V, and the Pt1000 thermistor was measured in three-wire half bridge mode using an excitation voltage of 4.897 V.

2.3 Calculations

Before analyses the data were processed in the following way: (1) outliers were removed; (2) relative humidities $>105\%$ (accuracy of capacitive humidity sensors) were deleted; (3) reference CH_4 mole fractions obtained from the FGGA (since 2016) were filtered based on hard boundaries of house-keeping variables available for quality control. For the latter we used the following hard boundaries for filtering: (a) sample cell pressure had to be in the range 130–143 Torr; (b) the instrument-specific ringdown time of the laser for CH_4 measurements had to be in the range 13–17 ms. The accepted reference CH_4 mole fractions were thus all measured in the narrow range of cell pressures between 139.7 and 140.3 Torr and laser ringdown times between 14.02 and 14.94 ms, which indicates best performance of the analyser. Before 2016 (FMA instrument) these house-keeping variables were not recorded.

The basic principle of operation of the TGS 2600 sensor was described in detail by Eugster and Kling (2012). The methane sensing mechanisms of different active materials used in solid state sensors was described by Aghagoli and Ardyanian (2018). The TGS 2600 uses a SnO_2 micro crystal surface (Figaro, 2005b). Whereas the manufacturer defines the sensor signal as R_s/R_0, the ratio of the electrical resistance R_s of the heated sensor material surface normalized over its resistance R_0 in the air under absence of CH_4, Hu et al. (2016) define the sensor signal as the ratio between R_s and R_g, the resistance of the surface in pure gas of interest (here CH_4). In all cases, considerations of technical sensor information is made for high ~~concentrations~~ mole fractions of CH_4 (e.g., 200 ~~ppm~~ μmol mol^{-1}) for a SnO_2 surface according to Hu et al. (2016), not for ambient ~~concentrations~~ mole fractions in the typical range 1.7–4 ~~ppm~~ μmol mol^{-1} (or less). Hence, some adaptations are always necessary because present-day sensors are not yet designed for such low ~~concentrations~~mole fractions. In order to

simplify calculations compared to what we presented in Eugster and Kling (2012) – which closely followed the technical information provided by the manufacturer ~~(Figaro, 2005b, a)~~ (Figaro, 2005a, b) – we define the sensor signal as $S_c = R_s/R_0$, but with R_0 arbitrarily set to the resistance observed when the sensor delivers V_0 = 0.8 V output at V_c = 5.0 V supply voltage. The highest voltages measured at TWE were 0.7501 and 0.7683 V from sensors #1 and #2, respectively (which theoretically corresponds to the lowest CH_4 ~~concentrations~~mole fractions). With these assumptions the sensor signal S_c can easily be approximated as a function of the inverse of the measured TGS signal voltage V_s,

$$S_c = \frac{R_S}{R_0} \approx 0.952381 \cdot \frac{1}{V_s} - 0.1904762 \,. \tag{1}$$

The full derivation is

$$\begin{aligned} \frac{R_s}{R_0} &= \frac{\frac{V_c \cdot R_L}{V_s} - R_L}{\frac{V_c \cdot R_L}{V_0} - R_L} = \frac{\frac{V_c \cdot R_L - V_s \cdot R_L}{V_s}}{\frac{V_c \cdot R_L - V_s \cdot R_L}{V_0}} = \\ &= \frac{V_0 \left(V_c - V_s\right) R_L}{V_s \left(V_c - V_0\right) R_L} = \frac{V_0 \left(V_c - V_s\right)}{V_s \left(V_c - V_0\right)} = \\ &= \frac{V_0}{V_c - V_0} \left(\frac{V_c}{V_s} - 1\right) = \frac{1}{V_s} \cdot \frac{V_c \cdot V_0}{V_c - V_0} - \frac{V_0}{V_c - V_0} \,. \end{aligned}$$

Here R_L is the load resistor over which V_s is measured (see Figaro (2005a) or Eugster and Kling (2012), for more details), but which can be eliminated in this algebraic simplification.

To compute absolute humidity, we used the Magnus equation to estimate saturation vapor pressure e_{sat} (in hPa) at ambient temperature T_a (in °C),

$$e_{\text{sat}} = 6.107 \cdot 10^{a \cdot T_a / (b + T_a)} \,,$$

with coefficients a = 7.5 and b = 235.0 for $T_a \geq 0$ °C, and a = 9.5 and b = 265.5 for $T_a < 0$ °C.

Actual vapor pressure e (hPa) was then determined as

$$e = e_{\text{sat}} \cdot \frac{rH}{100\%} \,,$$

with relative humidity rH in percent, and converted to absolute humidity ρ_v (kg m^{-3}) with

$$\rho_v = \frac{e}{T_a + 273.15} \cdot \frac{p}{p - e} \cdot \frac{100}{R_v} \approx 0.217 \cdot \frac{e}{T_a + 273.15} \,,$$

with p atmospheric pressure (hPa), and R_v the gas constant for water vapor (461.53 J kg^{-1} K^{-1}).

2.4 Statistical analyses

Statistical analyses were performed with R version 3.5.2 (R Core Team, 2018). Trend analyses were performed for both trend in CH_4 ~~concentration~~ mole fraction and drift of TGS 2600 measurements using the Mann-Kendall trend test implemented in

the rkt package that is based on Marchetto et al. (2013). The annual linear trend (or drift) was calculated using the robust Theil-Sen estimator ~~Akritas et al. (1995)~~ (Akritas et al., 1995) using weekly median values, and the significance of the trend (or drift) was assessed using Kendall's Tau parameter. All trend and drift estimates were significant at $p < 0.05$. The highest two-sided p-value of the presented results was $p = 0.000054$ and thus no detailed information on p-values is given when statistical significance of trends or drift is mentioned in the following.

For assessing the quality of the proposed calculation of CH_4 ~~concentrations~~ mole fractions from TGS 2600 sensors we inspected weekly aggregated data using four key indicators:

Bias – the mean of the difference of each 30-min averaged pair of CH_4 ~~concentrations in ppm~~mole fractions in μmol mol^{-1}, $CH_{4,TGS} - CH_{4,ref}$;

Stability – the bias expressed as a percent deviation from the reference CH_4 ~~concentration~~mole fraction, $(CH_{4,TGS} - CH_{4,ref})/CH_{4,ref} \cdot 100\%$;

Variability – the mean relative deviation of the 95% confidence interval (CI) observed with the TGS 2600 sensor from the corresponding 95% CI of the CH_4 reference measurements (in percent), $(CI_{95\%,TGS} / CI_{95\%,ref} - 1) \cdot 100\%$;

Correlation of median diel cycles – Pearman's product-moment correlation coefficient between hourly-aggregated median diel cycles of CH_4 measured by the TGS 2600 and reference instruments.

In addition to conventional linear model (least square method) fits we used an ~~artificial neural network~~ ANN approach. This was performed in Python 3.7.1 using MLPRegressor from sklearn.neural_network version 0.20.2 (Pedregosa et al., 2011). We used a network with four hidden layers of size 500, 100, 50 and 5, respectively, and an adaptive learning rate. Learning was done with the data obtained during the calibration period ~~from first measurements until 31 March 2017 and the remaining data (1 April 2017 to 31 December 2018)~~ 2014–2016, whereas the remaining years 2012–2013 and 2017–2018 were used for validation.

3 Results and discussion

CH_4 ~~concentrations~~ mole fractions estimated from TGS 2600 measurements during the cold seasons differed strongly from the reference measurements when the Eugster and Kling (2012) approach was used (not shown); that approach translated the information from the technical specifications of the TGS 2600 sensor ~~(Figaro, 2005b, a)~~ (Figaro, 2005a, b) to outdoor applications. With temperatures above freezing the agreement with the CH_4 reference measurements was within ± 0.1 ~~ppm~~ μmol mol^{-1} (Fig. 2), but not so during cold conditions (~~Ta~~ T_a < 0 °C). The differences between TGS estimates and CH_4 reference were largest with the Eugster and Kling (2012) approach when relative humidity was between 50 and 90% (Fig. 3a). When converting relative humidity to absolute humidity, the results became satisfactory for higher absolute humidity values > 0.004 kg m^{-3} (Fig. 3b). Using absolute humidity in place of relative humidity for the correction of the TGS 2600 was already attempted by Collier-Oxandale et al. (2018), which contrasts with the manufacturer's suggestion (Figaro, 2005a). Because absolute humidity > 0.004 kg m^{-3} is only possible at temperatures > 0 °C it appears quite obvious that temperature and humidity corrections of solid state sensors most likely do not relate to relative humidity~~(which is a ratio and not a physical variable of atmospheric~~

~~water content)~~, but to either actual vapor pressure (in hPa) or absolute humidity (in kg m^{-3}). In all tested models absolute humidity performed marginally better than vapor pressure or mixing ratio (measured by R^2 ~~,~~; not shown), hence we suggest the following model and parameterization to estimate CH_4 ~~concentrations in ppm~~ mole fractions in μmol mol^{-1} from TGS 2600 signal voltage measurements:

$$\begin{aligned} CH_4 \quad = \quad & 1.425 + 0.12\,S_c + 0.375/S_c - 0.0065\,T_a + \\ & +53.3\,\rho_v + 0.0022\,S_c \cdot T_a - 0.0017\,T_a/S_c + \\ & +4.9\,S_c \cdot \rho_v - 67.4\,\rho_v/Sc - 0.39\,S_c \cdot T_a \cdot \rho_v \\ & +1.15\,T_a \cdot \rho_v/S_c\,, \end{aligned} \tag{2}$$

with S_c the dimensionless sensor signal (see Eq. 1), T_a ambient air temperature in °C, and ρ_v absolute humidity in kg m^{-3}. The parameter estimates were derived from the entire dataset 2012–2018 for TGS sensor #1 (Table 1, "entire period"). For other sensors the result from Eq. (2) can be considered as a linearized signal that can be fine-tuned with a sensor-specific two-point calibration as suggested in Section 3.4 of Eugster and Kling (2012).

The linear model in Eq. (2) was derived from a suite of candidate models including interactions among predictors and including quadratic terms of each variable, and then stepwise elimination using the stepAIC function in the MASS package of R was employed to find the model with the lowest AIC (Akaike's Information Criterion). Unless explicitly mentioned, we analysed CH_4 mole fractions computed with Eq. (2) using the parameters obtained from all data measured by TGS 2600 sensor #1. Only in the direct comparison with the ANN (Section 3.1) did we determine an additional parameter set using the same calibration period as the ANN used (Section 2.4), so that a direct comparison of performance in validation was possible.

If ambient temperature influences the signal of the TGS 2600 in such a way as expected from the technical documentation (Figaro, 2005a, b), then wind speed could be a third factor influencing the conversion from TGS 2600 sensor voltages to CH_4 mole fractions. To investigate this additional factor, we produced a heat loss model, assuming that the sensor correction is related to the cooling of the heated surface of the solid state sensor, which has a nominal surface temperature T_s = 400 °C (Falabella et al., 2018) that is the typical operation temperature of SnO_2-Ni_2O_3 sensors (Hu et al., 2016). Our candidate model for heat loss (HL in W) was

$$HL \sim \xi \cdot \overline{u}^2 \cdot (T_s - T_a) \cdot (\rho_d \cdot C_d + \rho_v \cdot C_v)\,, \tag{3}$$

with $\overline{u}$ mean horizontal wind speed (m s^{-1}), T_s and T_a sensor surface and ambient air temperature (K), respectively, ρ_d density of dry air (kg m^{-3}), ρ_v absolute humidity (kg m^{-3}), and C_d and C_v heat capacity of dry air and water vapor, respectively (J kg^{-1} K^{-1}). The scaling coefficient ξ is a best fit model parameter (units: s m). The assumption made here was that the wind speed governs the eddy diffusivity of heat transported along the temperature gradient between the sensor surface and ambient air, and the moisture correction is only associated with the fact that water vapor has a higher heat capacity (1859 J kg^{-1} K^{-1}) than dry air (1005.5 J kg^{-1} K^{-1}), and hence the heat capacity of moist air increases accordingly with ρ_v.

3.1 Performance of the TGS 2600 sensor at 30-minute resolution

Using Eq. (2) yields satisfying agreement with 30-minute averaged data under both typical low Arctic summer and winter conditions (Fig. 4) with an overall R^2 of ~~0.42~~ 0.424 (Table 1). When testing the linear model approach (Eq. 2) more rigorously by splitting the available data into a calibration period (years 2014–2016) and a validation period (years 2012–2013 and 2017–2018), some limitations can be seen, in particular under cold conditions where none of the approaches performed very well in the validation period. The ANN had a more balanced performance between calibration and validation period, although it performed slightly less well under warm conditions ($T_a \geq 0$°C).

A detailed inspection of four representative seven-week time periods at full 30-minute resolution is shown in Figures 5–8. Typical summer conditions at the beginning of this study (Fig. 5) and towards the end of the analyzed period (Fig. 6) indicate that the short-term agreement (R^2 = ~~0.65~~0.653, Fig. 5) is better when the TGS sensor was still relatively new than at age seven (R^2 = ~~0.38~~0.381, Fig. 6), but the variability decreased (improved) from –42% to –9% with no relevant difference in bias and stability (0.01~~ppm~~ μmol mol^{-1} and 0.4% vs. 0.00 ~~ppm~~ μmol mol^{-1} and 0.0%, respectively). In winter the timing of most events is correctly captured (Fig. 7) with a R^2 of ~~0.45~~0.445, but the dynamics are not satisfactorily captured by the TGS sensor, indicated by a 59% underestimation of the 95% CI during this mid-winter period. The transition from warm to cold season (Fig. 8) shows a mixture of days, where the regular diel cycle, which is typical for the warm season, is still adequately captured, but the dynamics of periods with air temperature <0 °C (see Fig. 4) when CH_4 ~~concentrations~~ mole fractions tend to be highest as in winter (Fig. 7), are not adequately captured. Still, with a R^2 of ~~0.51~~ 0.512 (Fig. 8) more than 50% of the variance observed in the 30-minute averaged CH_4 reference measurements is captured by the low-cost TGS 2600 sensor.

Because of the absence of local sources of carbon monoxide and other air pollutants to which the TGS 2600 sensor is also sensitive besides CH_4, we investigated a special case when smoke and haze from wildfires south of the Books Range polluted the air in the TFS area on 26 June 2015 and compared the performance of both TGS sensors during that day with conditions three days before that event, and on the same date in the following three years. The net effect of increased air pollutants was an apparent small decrease of the CH_4 mole fractions calculated via Eq. (2) by approximately –0.03 μmol mol^{-1}. At the same time the variability of the residuals increased from typically $\pm$0.014 to $\pm$0.027 μmol mol^{-1} (24-hour averages). Thus, the influence of the wildfire smoke was of the same order of magnitude as the difference between TGS-derived CH_4 mole fractions and the reference instrument on most other days of the year (see Figs. 5–8).

3.2 Performance of weekly aggregated data

The TGS 2600 is not expected to provide short-term accuracy comparable to high-quality instrumentation (see also Lewis et al., 2018). However, Eugster and Kling (2012) argued that such measurements still may provide additional insights as compared to the passive samplers described by Godbout et al. (2006a, b), integrating over longer time frames. Thus, here we inspected the performance of weekly aggregated estimates derived from the TGS 2600 in order to inspect drift of the two sensors and their performance over the seven-year deployment period. Note, that in Eq. (2) we did not include a drift correction. Figure 9 shows weekly medians of sensor signals, the agreement with the reference signal, and the difference between the ~~raw signals from~~

both CH_4 mole fractions obtained from both TGS 2600 sensors mounted at the same position (Fig. 1), and the agreement with the reference signal. The two TGS 2600 sensors (#1 and #2) showed a trend in their signals of –18.8 mV yr^1 and –15.5 mV yr^{-1}, respectively (Fig. 9a,b). Thus, with typical signals on the order of 200–700 mV (Fig. 9a) the lowest (winter) readings may no longer be measurable after 10–13 years of continuous operation, indicating the end of life of a TGS 2600.

Figure 10 shows the weekly median bias, variability, and the correlation between the weekly aggregated median diel cycle of CH_4 at hourly resolution between the TGS #1 measurements and the reference. Despite the trend of the sensor signal shown in Figure 9a,b both the bias and variability primarily show a seasonal pattern with a slightly negative bias (around –0.02 ppmμmol mol^{-1}) during peak growing season and a corresponding positive deviation in mid-winter when temperatures can be well below –30 °C (Fig. 10a). The variability (Fig. 10b) shows the inverse pattern of the bias. If bias is expressed as the relative bias (i.e., stability), the stability vs. variability plot (Fig. 11) shows points lying uniformly around the line of a –1:1 relationship (R^2 = 0.67). This indicates that both variability and stability can be improved at the same time because there is no tradeoff visible in Fig. 11.

3.3 Linear trend and drift estimates

All linear trend estimates were statistically significant (see Section 2.4). However, our measurements started with warm-season measurements only (2012–2014) that were successively expanded to include cold season measurements. Thus, all interpretation of the trends and drifts presented here should be considered with caution given the long gaps in data due to technical challenges operating such equipment under adverse winter conditions. The CH_4 concentration mole fraction trend observed with the high-quality reference measurements was 10.1 ppb nmol mol^{-1} yr^{-1}. This is 2.5 times the trend observed from 2005–2011 by NOAA (28.6 $\pm$ 0.9 ppb nmol mol^{-1} or 4.09 ppb nmol mol^{-1} yr^{-1}; Hartmann et al. (2014); their Table 2.1), but of the same order of magnitude reported by Nisbet et al. (2014) for 2013 (last year covered by that study) for latitudes north of the Tropic of Cancer. Thus, this trend may be real and hence all trends seen in low-cost sensor signals are not necessarily solely an artefact of such sensors. It however remains a challenge to deduce the true trend in CH_4 concentrations mole fractions over longer time periods using such a low cost sensor because of drifting signals. Thus, we inspected the drift of the TGS 2600 derived concentration mole fraction with respect to the (true) CH_4 trend observed with the high-quality reference instrument. These drifts appear to be smaller than the true trend, but are still considerable: the bias of TGS-derived CH_4 concentrations mole fractions drifted by 4–6 ppb nmol mol^{-1} yr^{-1} (40–60% of actual trend) and variability drifted by –0.24% yr^{-1}. They provide encouraging results suggesting that with occasional (infrequent) calibration against a high-quality standard, e.g. using a traveling standard operating during a few good days with adequate coverage of the near-surface diel cycle of CH_4, TGS 2600 measurements might be suitable for the monitoring of CH_4 concentrations mole fractions also in other areas. As shown in Figure 10c the correlation of median diel cycles between TGS estimates and CH_4 reference is one of the weak points in the current performance of the TGS 2600 sensors. Also, we observed a significant negative trend of the correlation coefficient of –0.051 yr^{-1} (Fig. 10c). However, the key finding is that the typical diel cycle during the warm season (air temperature > 0 °C) disappears during winter conditions (Figs 4, 7) and thus separate transfer functions for warm and cold temperatures might be

a solution for future studies (Table 1). Our Eq. (2) is thus on purpose derived from the entire dataset to provide a starting point for more elaborate fine-tuning in projects where this is desired.

3.4 Potential of using artificial neural networks

Casey et al. (2019) found that artificial neural networks (ANNs) outperformed linear models in mitigating curvature and linear trends in trace gas measurements when used with the same set of input variables during a three-month comparison period. To inspect the potential of ANNs at our Arctic long-term dataset, we added the ANN results to Figures 5–8. In summer (Figures 5, 6) we did not find a substantial ~~advantage of ANN over~~ difference between an ANN and the linear approach of Eq. (2) in terms of root mean square error (RMSE) ~~of the prediced~~ or R^2 between prediced and measured CH_4 ~~concentrations~~ mole fractions (Table 1). In ~~particular~~winter with temperatures below freezing, the ANN ~~was unable to better reflect the skewed distribution of diel variations of CH_4 concentrations and tended to underestimate daytime minimum concentration without a~~clear improvement of the fit to nocturnal peaks.performed clearly better in the validation than the linear approach, but both approaches remained unsatisfactory ($R^2 < 0.1$), despite the fact that both approaches were similar in the calibration period ($R^2 \approx 0.3$, Table 1).

During the warm period we found cases where the ANN was much better in capturing a specific daily feature as e.g. on 11 July 2012 (Fig. 5a) where the daily minimum was nicely captured by the ANN but the linear model was much too low. Contrastingly, in 2017 (Fig. 6a) periods could be found where the daily dynamics was correctly captured by the ANN but at too low mixing ratios (e.g. 10–18 July 2017, Fig. 6a). It should be noted that at this latitude the sun does not set between 24 May and 20 July, thus nocturnal conditions are clearly different from conditions at lower latitudes such as the ones investigated by Casey et al. (2019). Similarly, the transition from warm to cold season (Fig. 8) ~~does also not provide strong evidence of improvements with an ANN that uses the same input variables as a linear model . However, in winter when the diel cycle of all variables is less dominant than during the warm and transition seasons, the ANN approach is capable of producing a more realistic~~ was challenging both the linear model and ANN approach. We have only used two variables, T_a and humidity, that according to manufacturer specifications (Figaro, 2005a, b) influence the TGS2600 sensor signal. In reality, the same two variables also influence the CH_4 ~~estimate from TGS 2600 signals than the linear model approach (e.g., Fig. 7). This is not very surprising as the prediction range of a linear model decreases linearly with the amplitude of the input variables (e.g., in absolute humidity, which does not vary strongly at sub-zero temperatures), whereas an ANN does not have such linearity constraints.~~ production in water logged ecosystems, and thus contribute to the true CH_4 signal in addition to the cross-sensitivity which we try to correct with Eq. (2).

~~The drawback of ANNs is the difficulty in determining why the fit has become better than an empirical linear model fit. On the other side, this could encourage future work to better understand the physical response of TGS sensors to cold environments. In summary, we strongly recommend to focus on physical relationships, because it is unclear what scientific insights an ANN fit could offer in addition to the technically nicer fits to data~~An ANN that can separate the effect of ambient variations of CH_4 mole fractions from the artefact of cross-sensitivity of the TGS2600 to T_a and humidity may however outperform a linear

model approach in future studies, if more potentially important driving variables are included than only those specified by the manufacturer (Figaro, 2005a, b).

3.5 Suggestions for future work

The ~~residuals of Eq. (2) are homoscedastic both when plotted against~~ inter-quartile ranges and 95% confidence intervals of each air temperature (Fig. 12a) or absolute humidity bin (Fig. 12b) ~~.~~ are very similar over a wide range of temperature and humidity, but tend to become more variable in bins with few data (i.e., lowest and highest temperatures, and highest absolute humidities in Fig. 12). Deviations are generally constrained within ± 0.1 ~~ppm~~ μmol mol^{-1} or better, but with higher variability at both temperature ends where data coverage is poor (gray bars at bottom of Fig. 12a) as temperatures < –30 °C were not frequently covered due to technical problems with the measurement station, or at summer temperatures > 20 °C that are still rather rare at this low Arctic latitude (Hobbie and Kling, 2014). A slightly different picture emerges for low absolute humidity: 56% of measurements are at lower humidities than the saturation humidity at 0 °C (0.0049 kg m^{-3}), thus the rather homgenous variances at low humidity (Fig. 12b) indicate that humidity is not of concern at low temperatures, and future attempts for improvements should rather focus on humidity > 0.01 kg m^{-3} and temperatures > 20 °C that are not normally found in the Arctic.

Based on physical considerations one might expect that specific humidity or water vapor mixing ratio instead of absolute humidity could lead to further improvements, because absolute humidity still depends on temperature. However, our tests have not indicated a relevant gain of information or accuracy of prediction, but future work should also try to find a better physical correction model than the purely empirical one used here based on manufacturer information.

Another approach was taken by van den Bossche et al. (2017) who performed an in-depth laboratory calibration of the very similar but less sensitive Figaro TGS 2611-E00 sensor ~~(Figaro, 2013, the manufacturer only shows a response above 300 ppm CH$_4$)~~ (Figaro, 2013, the manufacturer only shows a response above 300 μmol mol^{-1} CH$_4$) at different temperatures and relative humidity over a CH$_4$ calibration range starting at $\approx$ 2~~ppm ambient concentration~~ μmol mol^{-1} ambient mole fraction up to 10~~ppm~~ μmol mol^{-1} CH$_4$. Despite the effort, the residual ~~concentrations~~ mole fractions remained large (range of ca. –1.5~~ppm~~ μmol mol^{-1} to +1.1~~ppm~~ μmol mol^{-1}) – too large for the application we present here. Our efforts to calibrate our TGS 2600 sensors in a laboratory climate chamber in a similar way was not satisfactory (Eugster, unpublished), hence our approach presented here to determine the sensor behavior from long-term outdoor measurements under real-world conditions. Contrastingly, Kneer et al. (2014) are convinced that "to be of use for advanced applications metal-oxide gas sensors need to be carefully prepared and characterized in laboratory environments prior to deployment". ~~We however do not fully agree because laboratory conditions simplify the real world too much, and it is~~ While this is theoretically correct, it remains difficult to carry out laboratory treatments from –41 °C to 27 °C as ~~it~~ would be required for our Arctic site. The data we present indicate that most likely it is absolute humidity (or specific humidity or mixing ratio), not relative humidity, that should be used for such calibrations, which in principle should provide the best quality results if the relevant factors are known and can be included in the calibration set-up.

~~Using absolute humidity in place of relative humidity for the correction of the TGS 2600 was already attempted by Collier-Oxandale et al.~~ ~~, which contrasts with the manufacturer's suggestion Figaro (2005a)~~It would be desirable that manufacturers carry out both laboratory tests and field trials and provide the necessary correction functions together with sensors. However, ~~as mentioned above, relative humidity is not a physical entity and hence using actual vapor pressure or absolute humidity may indeed be a better approach to understand the TGS 2600 sensor sensitivity to environmental conditions other that CH_4 concentrations.~~ due to the expense and time it takes to carry out long tests, be it in the laboratory or in the field, the present development goes in the direction of collocation studies (Piedrahita et al., 2014) en route to certification of sensors (N. Martin, NPL, UK, pers. comm.), similar to what we have done in the Arctic.

The TGS 2600 sensor's best performance is in applications where passive samplers would be another option (see also Eugster and Kling, 2012). Contrastingly, using the TGS 2600 for short-term measurements (resolution of seconds to minutes) has not yet led to satisfactory results (Kirsch, 2012; Falabella et al., 2018). In our dataset we found that adding wind speed to the ~~calculation~~ empirical linear model slightly improved the model fit during the warm season, but because no reliable continuous winter wind speed measurements were possible at the TWE site we did not include wind speed in our Eq. (2). However, this may be a key component for understanding the variability of TGS 2600 measurements when flying an unmanned aerial vehicle (UAV) where turbulent conditions may change within seconds to minutes. To address this additional factor, we ~~produced a heat loss model, assuming that the sensor correction is related to the cooling of the heated surface of the solid state sensor, which has a nominal surface temperature T_s = 400°C (Falabella et al., 2018) that is the typical operation temperature of SnO_2-Ni_2O_3 sensors (Hu et al., 2016). Our candidate model for heat loss~~ used the heat loss model given in Eq. (~~HL in W) was~~

$$HL \sim \xi \cdot \overline{u}^2 \cdot (T_s - T_a) \cdot (\rho_d \cdot C_d + \rho_v \cdot C_v) \ ,$$

~~with $\overline{u}$ mean horizontal wind speed (m s^{-1}), T_s and T_a sensor surface and ambient air temperature (K), respectively, ρ_d density of dry air (kg m^{-3}), ρ_v absolute humidity (kg m^{-3}), and C_d and C_v heat capacity of dry air and water vapor, respectively (J kg^{-1} K^{-1}). The scaling coefficient ξ is a best fit model parameter (units: s m). The assumption made here was that the wind speed governs the eddy diffusivity of heat transported along the temperature gradient between the sensor surface and ambient air, and the moisture correction is only associated with the fact that water vapor has a higher heat capacity (1859 J kg^{-1} K^{-1}) than dry air (1005.5 J kg^{-1} K^{-1}), and hence the heat capacity of moist air increases accordingly with ρ_v.~~ 3). However, although this approach is more mechanistic than Eq. (2), ~~it's~~ its ability to predict CH_4 ~~concentration~~ mole fraction from TGS 2600 measurements was much worse than that of the empirical linear model and ANN approaches (Table 1). But in order to make further progress on improving the transfer function from TGS 2600 signals to defensible CH_4 ~~concentrations~~ mole fractions it will be essential to increase our understanding of the physical processes that influence such measurements. This is not an easy task since there is a substantial proprietary knowledge that is unrevealed by the manufacturer. Newer, promising developments are underway that work with a mixed potential sensor using tin doped indium oxide and platinum electrodes in combination with yttria-stabilized zirconia electrolyte that show a logarithmic signal range of 0–10 mV for the range of 1–3 ~~ppm~~ μmol mol^{-1} CH_4 of interest for ambient air studies (Sekhar et al., 2016). The basic principle that the active metal-oxide is charged with O_2 (or O^{2-}), which then oxidizes CH_4, seems to be similar to the SnO_2-based TGS 2600; thus there is a

good chance that our findings for the TGS 2600 are also useful for assessing the performance of newer solid-state sensors with different active materials.

4 Conclusions

We present the first long-term deployment of two identical TGS 2600 low-cost sensors that show a ~~(weak)~~ sensitivity to ambient levels of CH_4 (here: range 1.824–2.682 ~~ppm~~ μmol mol^{-1} as measured by a high-quality Los Gatos Research reference instrument). We suggest a new transfer function to correct the TGS 2600 signal for cross-sensitivity to ambient temperature and humidity that also yields satisfactory results under cold climate (Arctic) conditions with temperatures down to –40 °C. This was only possible by using absolute humidity and not relative humidity for the correction. With this correction determined over the entire 2012–2018 data period, the 30-minute average CH_4 ~~concentration~~ mole fraction could be derived from TGS 2600 measurements within $\pm$0.1 ~~ppm~~μmol mol^{-1}. The two completely different regimes of diel CH_4 ~~concentration~~ mole fraction variations during the cold season (typically with a snow cover and frozen surface waters) and the warm season (when plants are active in the low Arctic) suggest that further improvements can be obtained by more specifically developing separate transfer functions for cold and warm conditions. ~~The use of an artificial neural network shows the potential for improvements in transfer functions under very cold conditions.~~

We consider the quality of TGS 2600 derived CH_4 estimates adequate if aggregated over reasonable periods (e.g., days or one week), but caution should be taken with application where short-term response is of key relevance (e.g., within seconds to minutes required by mobile measurements with UAVs). The deterioration of the sensor signal over time indicates that a TGS 2600 operated under ambient conditions as in our deployment at a low Arctic site in northern Alaska (Toolik wet sedge site) has an estimated life time of ca. 10–13 years. Thus, there is a potential beyond preliminary studies if the TGS 2600 sensor is adequately calibrated and placed in a suitable environment where cross-sensitivities to gases other than CH_4 ~~is~~ are of no concern.

Data availability. The data used in this study can be downloaded from the Environmental Data Initiative (EDI) portal via doi:10.6073/pasta/dddeb05b2806e2f5788fadd6fc590ef1. The statistical fits shown in Figs. 5–8 are made available via the ETH Zurich Research Collection, doi:10.3929/ethz-b-000369689.

Author contributions. WE, JL and GWK designed the study, set up the instrumentation, and serviced the site. WE carried out the main analyses. JE helped with the main analyses, set up and carried out the ANN calculations. WE wrote the manuscript and all co-authors worked, commented, and revised various versions.

Competing interests. None.

Disclaimer. The authors are independent from the producers of the instruments and sensors referenced in this article, and thus the authors do not have a commercial interest to promote any of the mentioned products.

Acknowledgements. We thank Jeb Timm, Colin Edgar, and other members of the Toolik Field Station science support staff for field help under difficult conditions. We also thank support staff from CPS for help with power supplies, technicians and students supported on several NSF grants, as well as several students supported by the NSF-REU program for help in the field over the years.

We acknowledge support received from the Arctic LTER grants NSF-DEB-1637459, 1026843, 1754835, NSF-PLR 1504006, and supplemental funding from the NSF-NEON and OPP-AON programs. Gaius R. Shaver (MBL) is acknowledged for initiating the study and supporting our activities in all aspects. ETH is acknowledge for supporting the purchase of the Fast Greenhouse Gas Analyzer that replace the older Fast Methane Analyzer in 2016 (grant 0-43683-11).

We thank the two anonymous reviewers for their careful and helpful assessments.

References

Aghagoli, Z. and Ardyanian, M.: Synthesis and study of the structure, magnetic, optical and methane gas sensing properties of cobalt doped zinc oxide microstructures, Journal of Materials Science: Materials in Electronics, 29, 7130–7141, https://doi.org/10.1007/s10854-018-8701-4, 2018.

Akritas, M. G., Murphy, S. A., and Lavalley, M. P.: The Theil-Sen estimator with doubly censored data and applications to astronomy, Journal of the American Statistical Association, 90, 170–177, https://doi.org/10.1080/01621459.1995.10476499, 1995.

Casey, J. G., Collier-Oxandale, A., and Hannigan, M.: Performance of artificial neural networks and linear models to quantify 4 trace gas species in an oil and gas production region with low-cost sensors, Sensors and Actuators B: Chemical, 283, 504–514, https://doi.org/10.1016/j.snb.2018.12.049, 2019.

Castell, N., Dauge, F. R., Schneider, P., Vogt, M., Lerner, U., Fishbain, B., Broday, D., and Bartonova, A.: Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates?, Environment International, 99, 293–302, https://doi.org/10.1016/j.envint.2016.12.007, 2017.

Collier-Oxandale, A., Casey, J. G., Piedrahita, R., Ortega, J., Halliday, H., Johnston, J., and Hannigan, M. P.: Assessing a low-cost methane sensor quantification system for use in complex rural and urban environments, Atmospheric Measurement Techniques, 11, 3569–3594, https://doi.org/10.5194/amt-11-3569-2018, 2018.

Duc, N. T., Silverstein, S., Wik, M., Crill, P., Bastviken, D., and Varner, R. K.: Greenhouse gas flux studies: An automated online system for gas emission measurements in aquatic environments, Hydrology and Earth System Sciences Discussions, pp. 1–18, https://doi.org/10.5194/hess-2019-83, 2019.

Eugster, W. and Kling, G. W.: Performance of a low-cost methane sensor for ambient concentration measurements in preliminary studies, Atmospheric Measurement Techniques, 5, 1925–1934, https://doi.org/10.5194/amt-5-1925-2012, www.atmos-meas-tech.net/5/1925/2012/, 2012.

Falabella, A. D., Wallin, D. O., and Lund, J. A.: Application of a customizable sensor platform to detection of atmospheric gases by UAS, in: 2018 International Conference on Unmanned Aircraft Systems (ICUAS), IEEE, https://doi.org/10.1109/icuas.2018.8453480, 2018.

Figaro: TGS 2600 – for the detection of air contaminants, Online product data sheet, Rev. 01/05, http://www.figarosensor.com/product/docs/TGS2600B00%20%280913%29.pdf, last Access: 29 September 2019, 2005a.

Figaro: Technical information on usage of TGS sensors for toxic and explosive gas leak detectors, Online product information sheet, Rev. 03/05, https://www.electronicaembajadores.com/datos/pdf2/ss/ssga/tgs.pdf, last Access: 29 September 2019, 2005b.

Figaro: TGS 2611 – for the detection of methane, Online product data sheet, Rev. 10/13, http://www.figarosensor.com/products/docs/TGS%202611C00%281013%29.pdf, last Access: 29 September 2019, 2013.

Godbout, S., Phillips, V. R., and Sneath, R. W.: Passive flux samplers to measure nitrous oxide and methane emissions from agricultural sources, Part 1: Adsorbent selection, Biosystems Engineering, 94, 587–596, https://doi.org/10.1016/j.biosystemseng.2006.04.014, 2006a.

Godbout, S., Phillips, V. R., and Sneath, R. W.: Passive flux samplers to measure nitrous oxide and methane emissions from agricultural sources, Part 2: Desorption improvements, Biosystems Engineering, 95, 1–6, https://doi.org/10.1016/j.biosystemseng.2006.05.007, 2006b.

Hartmann, D. L., Tank, A. M. K., Rusticucci, M., Alexander, L., Broennimann, S., Charabi, Y. A.-R., Dentener, F., Dlugokencky, E., Easterling, D., Kaplan, A., Soden, B., Thorne, P., Wild, M., and Zhai, P.: Observations: Atmosphere and Surface, chap. 2, IPCC AR5, 2014.

Hobbie, J. E. and Kling, G. W., eds.: Alaska's Changing Arctic: Ecological Consequences for Tundra, Streams, and Lakes, Long-Term Ecological Research (LTER) Network Series, Oxford University Press, New York, USA, 331 pp., 2014.

Hu, J., Gao, F., Zhao, Z., Sang, S., Li, P., Zhang, W., Zhou, X., and Chen, Y.: Synthesis and characterization of Cobalt-doped ZnO microstructures for methane gas sensing, Applied Surface Science, 363, 181–188, https://doi.org/10.1016/j.apsusc.2015.12.024, 2016.

Kirsch, O.: Entwicklung, Bau und Einsatz einer autonomen Drohne zur Messung von Gaskonzentrationen in der Luft, Maturarbeit mit beteiligung am wettbewerb schweizer jugend forscht, Kantonsschule Chur, Chur, Switzerland, 17 pp, 2012.

Kneer, J., Eberhardt, A., Walden, P., Pérez, A. O., Wöllenstein, J., and Palzer, S.: Apparatus to characterize gas sensor response under real-world conditions in the lab, Review of Scientific Instruments, 85, 055 006, https://doi.org/10.1063/1.4878717, 2014.

Lewis, A. C., von Schneidemesser, E., and Peltier, R.: Low-cost sensors for the measurement of atmospheric composition: overview of topic and future applications, Tech. Rep. WMO-No. 1215, WMO, Geneva, Switzerland, http://www.wmo.int/pages/prog/arep/gaw/documents/Low_cost_sensors_post_review_final.pdf, last Access: 24 February 2019, 2018.

Marchetto, A., Rogora, M., and Arisci, S.: Trend analysis of atmospheric deposition data: A comparison of statistical approaches, Atmospheric Environment, 64, 95–102, https://doi.org/10.1016/j.atmosenv.2012.08.020, 2013.

Nisbet, E. G., Dlugokencky, E. J., and Bousquet, P.: Methane on the rise – again, Science, 343, 493–495, https://doi.org/10.1126/science.1247828, science Perspectives, 2014.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É.: Scikit-learn: machine learning in Python, Journal of Machine Learning Research, 12, 2825–2830, http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf, 2011.

Piedrahita, R., Xiang, Y., Masson, N., Ortega, J., Collier, A., Jiang, Y., Li, K., Dick, R. P., Lv, Q., Hannigan, M., and Shang, L.: The next generation of low-cost personal air quality sensors for quantitative exposure monitoring, Atmospheric Measurement Techniques, 7, 3325–3336, https://doi.org/10.5194/amt-7-3325-2014, 2014.

R Core Team: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/, 2018.

Sekhar, P. K., Kysar, J., Brosha, E. L., and Kreller, C. R.: Development and testing of an electrochemical methane sensor, Sensors and Actuators B: Chemical, 228, 162–167, https://doi.org/10.1016/j.snb.2015.12.100, 2016.

Shamasunder, B., Collier-Oxandale, A., Blickley, J., Sadd, J., Chan, M., Navarro, S., Hannigan, M., and Wong, N.: Community-Based Health and Exposure Study around Urban Oil Developments in South Los Angeles, International Journal of Environmental Research and Public Health, 15, 138, https://doi.org/10.3390/ijerph15010138, 2018.

van den Bossche, M., Rose, N. T., and De Wekker, S. F. J.: Potential of a low-cost gas sensor for atmospheric methane monitoring, Sensors and Actuators B: Chemical, 238, 501–509, https://doi.org/10.1016/j.snb.2016.07.092, 2017.

Walker, D. A. and Everett, K. R.: Loess Ecosystems of Northern Alaska: Regional Gradient and Toposequence at Prudhoe Bay, Ecological Monographs, 61, 437–464, 1991.

WMO: Global Atmosphere Watch Measurements Guide, Tech. Rep. TD No. 1073, World Meteorological Organisation, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.360.4587&rep=rep1&type=pdf, 87 pp., 2001.

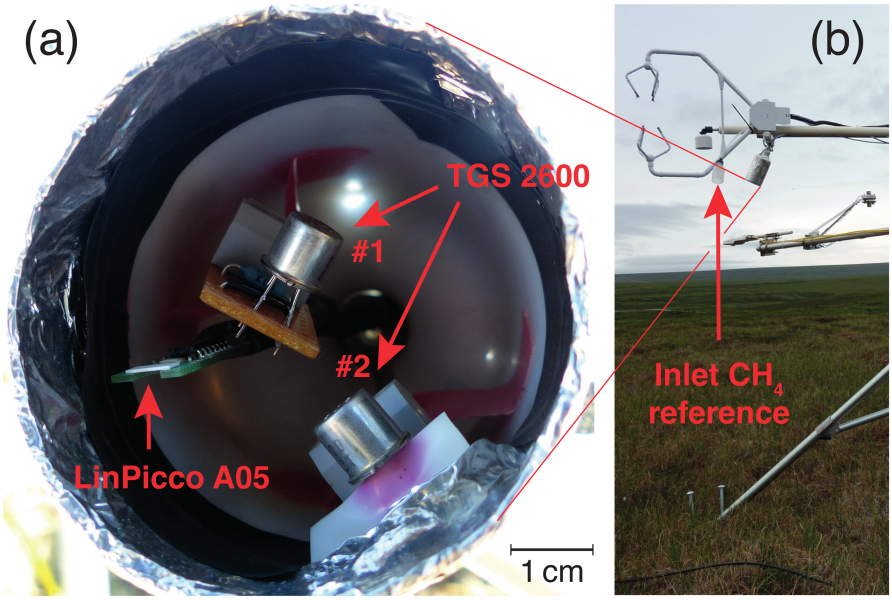


Figure 1. Two TGS 2600 trace gas sensors and the LinPicco A05 temperature and relative humidity sensor (a) inside the weather protection, and (b) the mounting position of the TGS weather protection and reference CH_4 gas inlets at the Toolik wet sedge eddy covariance flux site.

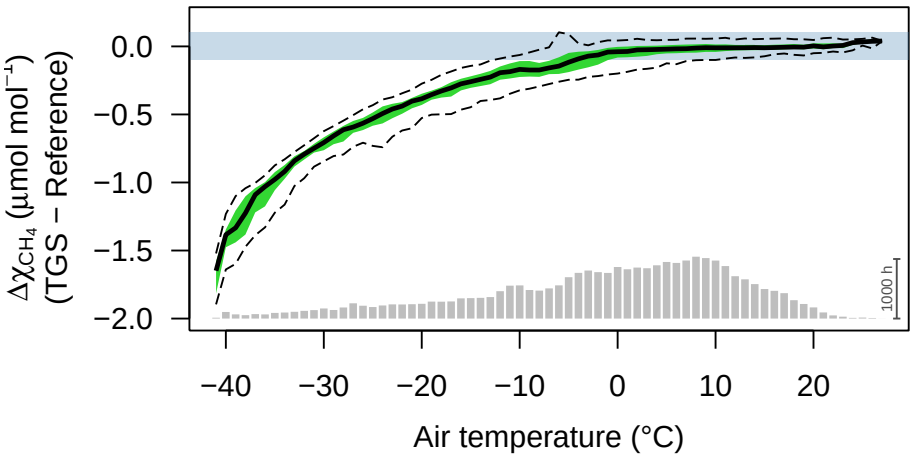


Figure 2. Difference between TGS2600 and reference CH_4 measurements (30-minute averages) as a function of air temperature when using the Eugster and Kling (2012) conversion. Agreement was good when ambient temperature was above freezing. The horizontal color bar shows the ±0.1 ~~ppm~~ μmol mol^{-1} range around a perfect agreement. The green band shows the inter-quartile range of bin-averaged differences (~~Figaro~~ TGS 2600 sensor #1), and dashed lines show the extent of the 95% confidence intervals. Gray bars at the bottom show number of 30-minute averages in each bin. The scale bar (1000 hours) at right specifies their size.

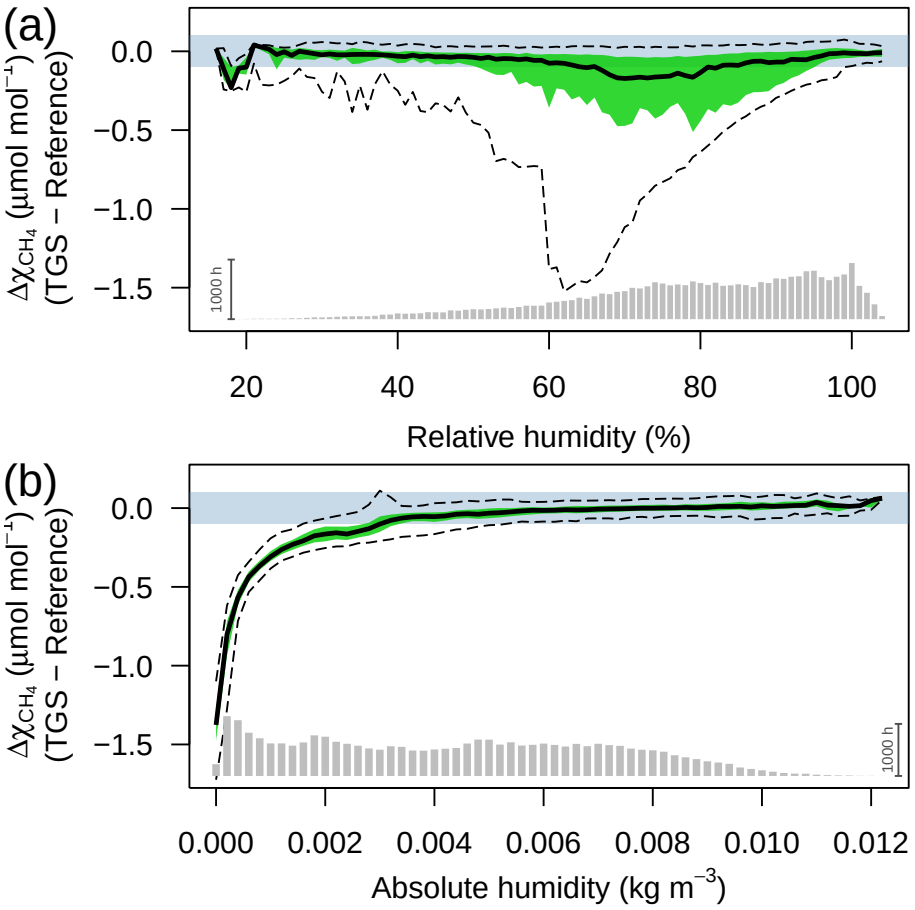


Figure 3. Difference between TGS2600 and reference CH_4 measurements (30-minute averages) as a function of (a) relative humidity (in %) and (b) absolute humidity (in kg m^{-3}). The horizontal color bar shows the ±0.1 ~~ppm~~ µmol mol^{-1} range around a perfect agreement. The green band shows the inter-quartile range of bin-averaged differences (~~Figaro~~ TGS 2600 sensor #1), and dashed lines show the extent of the 95% confidence intervals. Gray bars at the bottom show number of 30-minute averages in each bin. The scale bar (1000 hours) at right specifies their size.

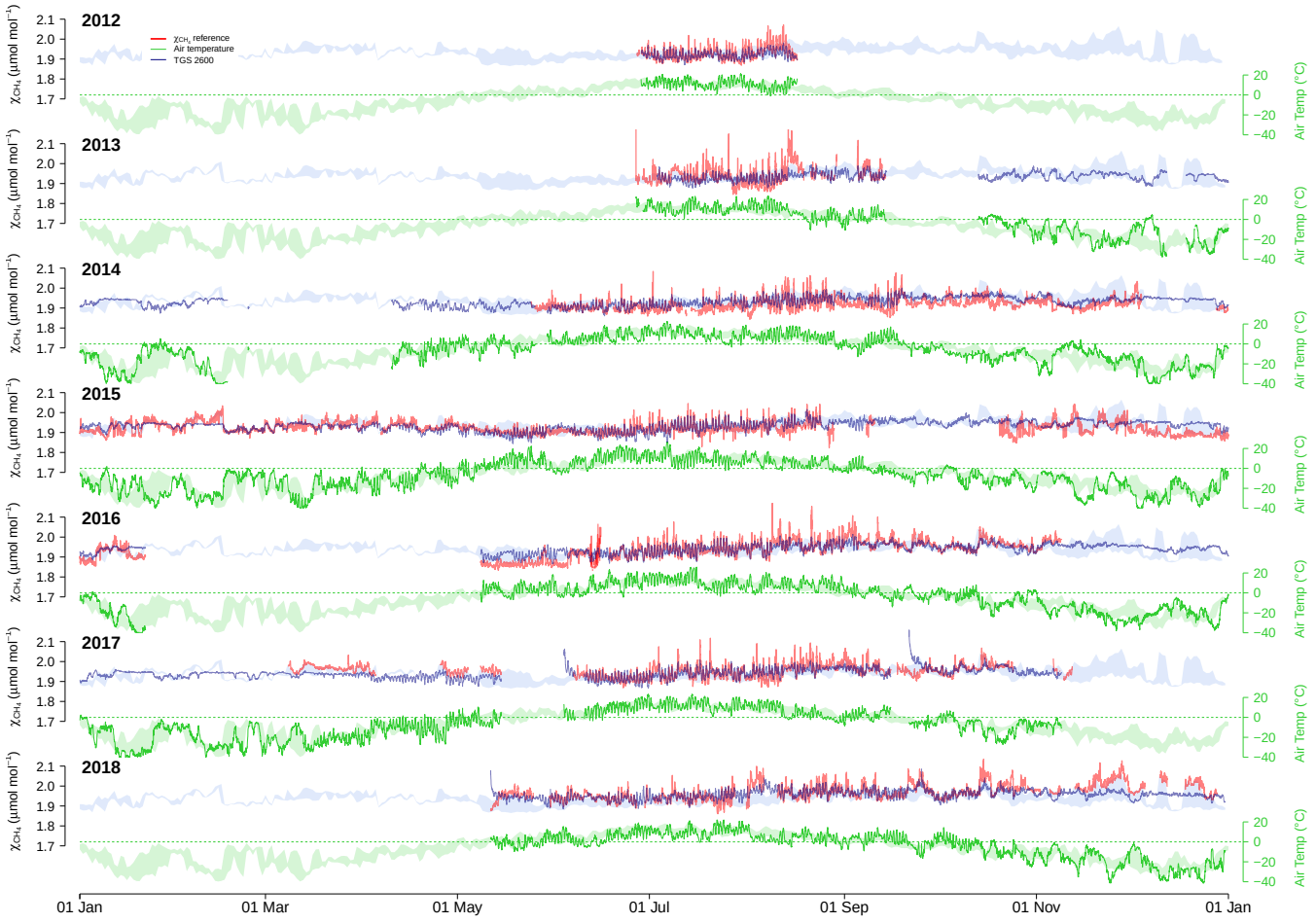


Figure 4. Overview over annual courses of 30-minute averaged air temperature (green) and CH_4 ~~concentrations~~ mole fractions (blue and red). Pale color bands show the daily inter-quartile range (50% of values between first and third quartile) of measurements from all years. Solid lines show actual measurements. Red lines are the reference CH_4 measurements, and blue lines show the CH_4 ~~concentration~~ mole fraction derived from TGS 2600 measurements (sensor #1). Actual measurements show 30-minute mean values.

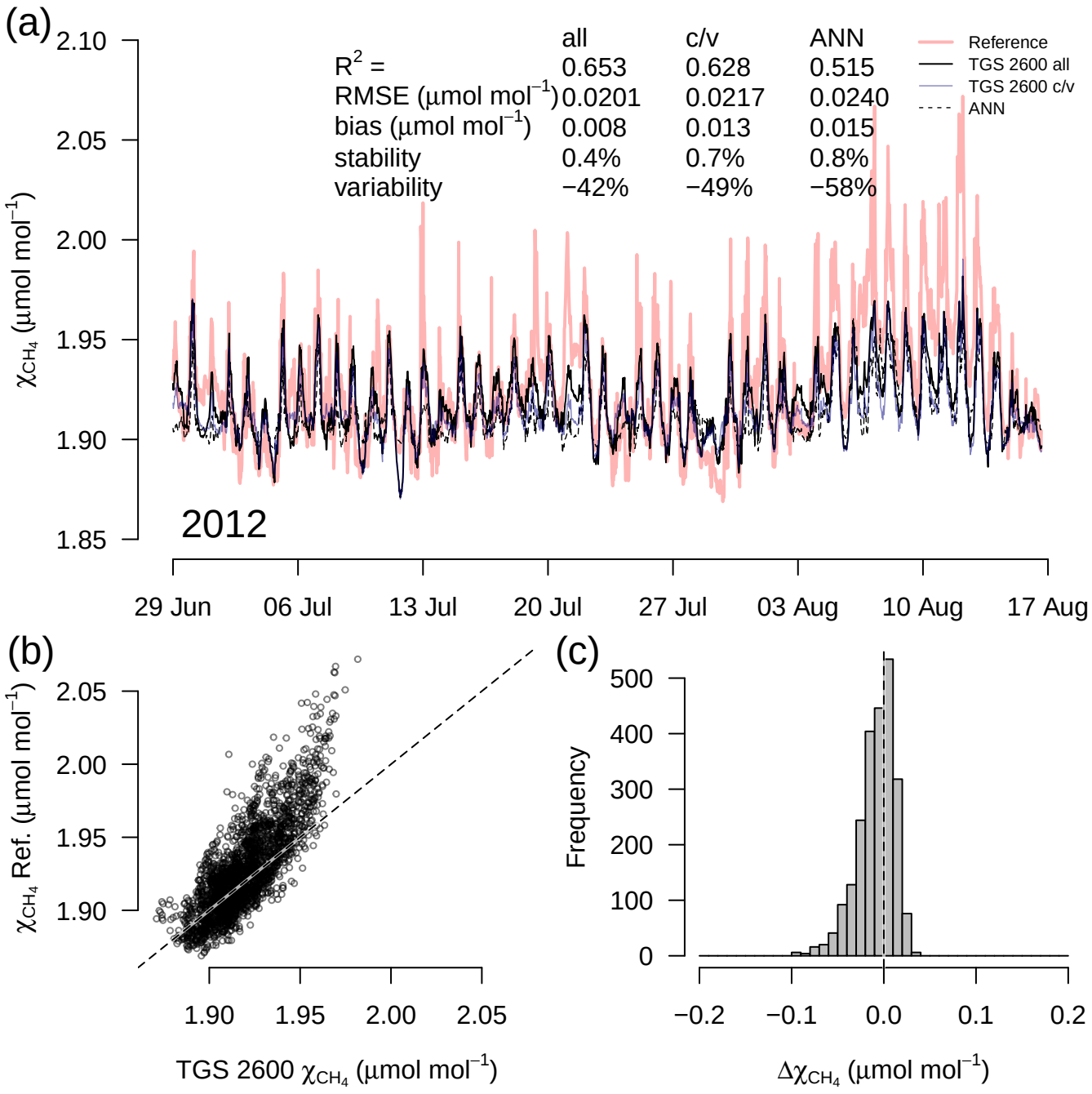


Figure 5. Timeseries (a) of TGS #1 derived CH_4 during a 7-week snow- and ice-free period in the first year of the long-term deployment (2012), (b) correlation with reference ~~concentration~~mole fraction, and (c) residuals (TGS 2600 all – Reference) of 30-minute averaged measurements. Thin solid lines in (a) show the result when all data are used with Eq. (2); reference ~~concentration~~ mole fraction is shown with a red bold line; c/v shows an alternative fit from splitting the available data into a calibration and a validation part; and the dashed line shows the performance of an artificial neural network (ANN) fit. This example belongs to the validation period of the TGS 2600 c/v and ANN fits.

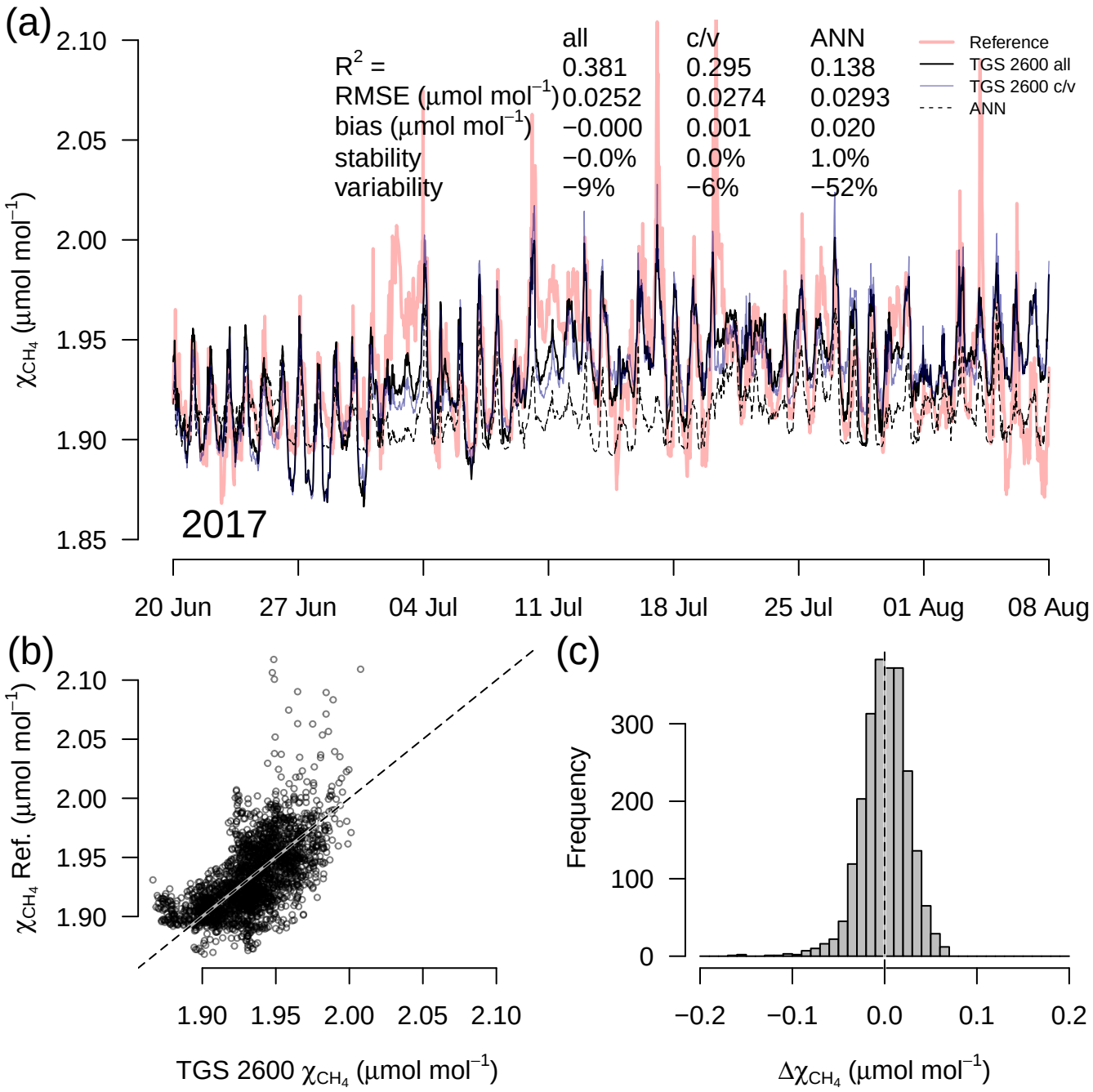


Figure 6. As in Figure 5 but with measurements from a 7-week snow- and ice-free period in 2017 at sensor age of seven years. This example belongs to the validation period of the TGS 2600 c/v and ANN fits.

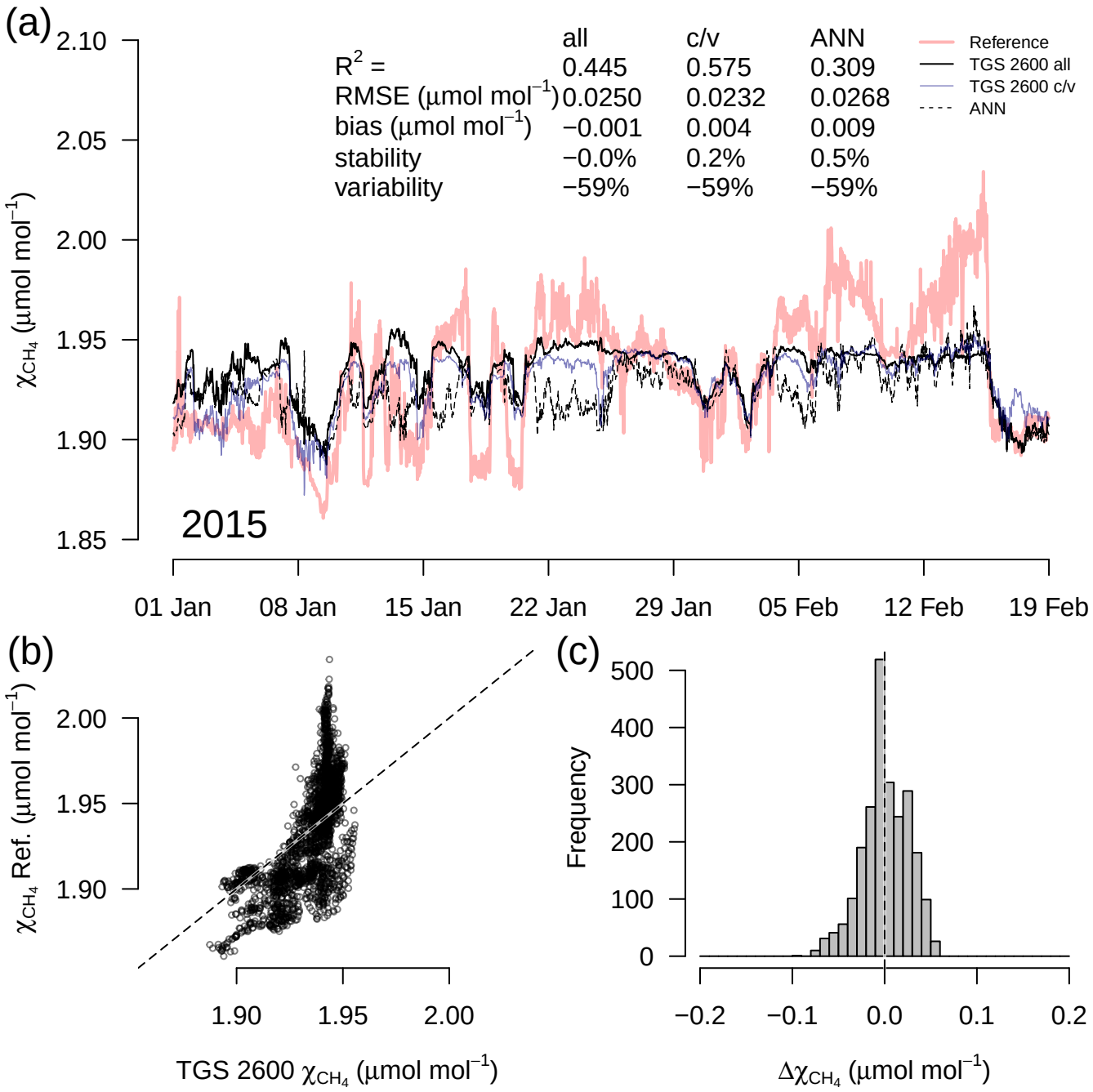


Figure 7. As in Figure 5 but with measurements from a 7-week period in mid winter with temperatures plunging down to –40 °C. High CH_4 ~~concentrations~~ mole fractions coincide with the coldest temperatures (see Fig. 4). This example belongs to the validation period of the TGS 2600 c/v and ANN fits.

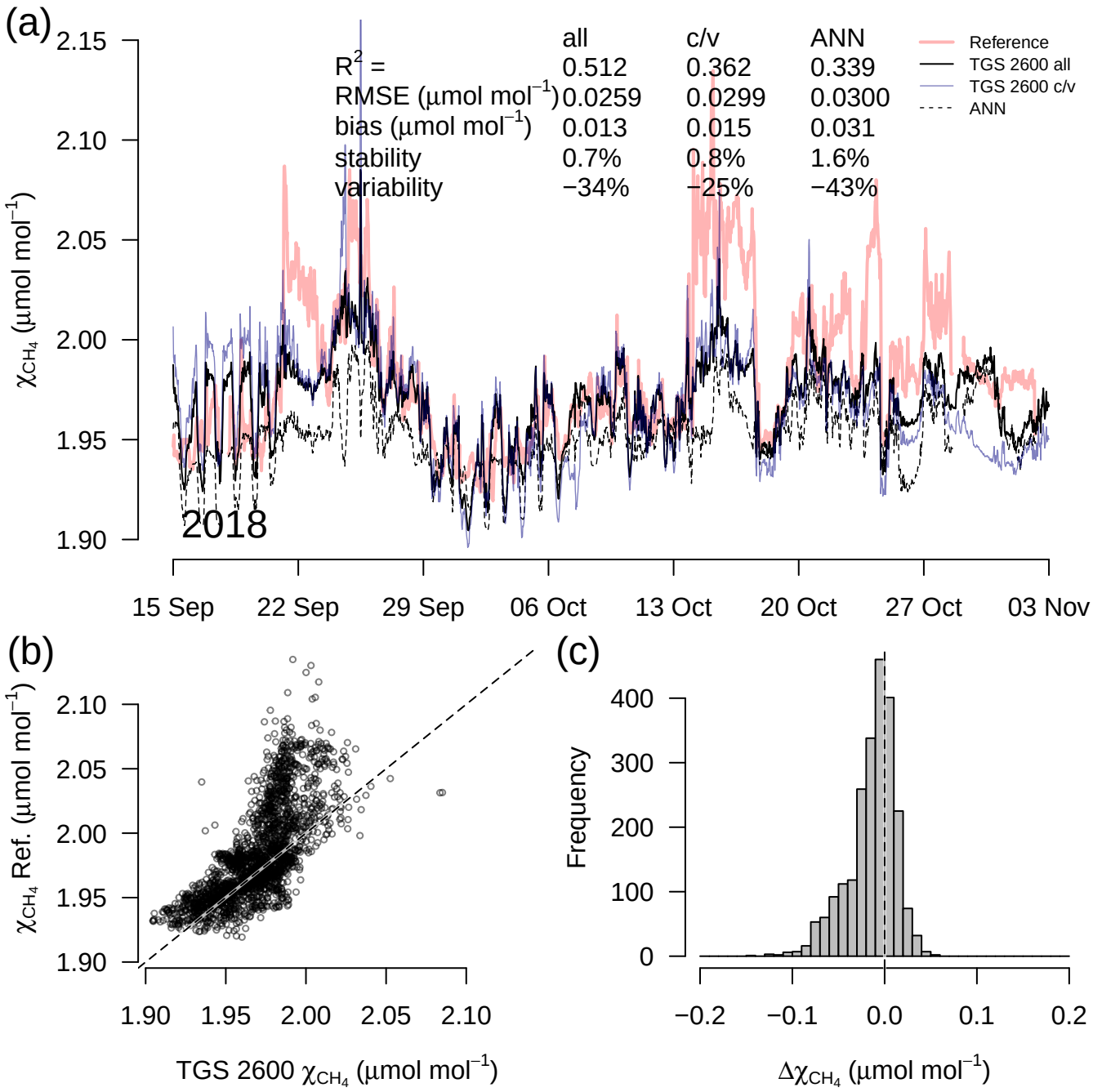


Figure 8. As in Figure 5 but with measurements from a 7-week period during the transition from fall to early winter. This example belongs to the validation period of the TGS 2600 c/v and ANN fits.

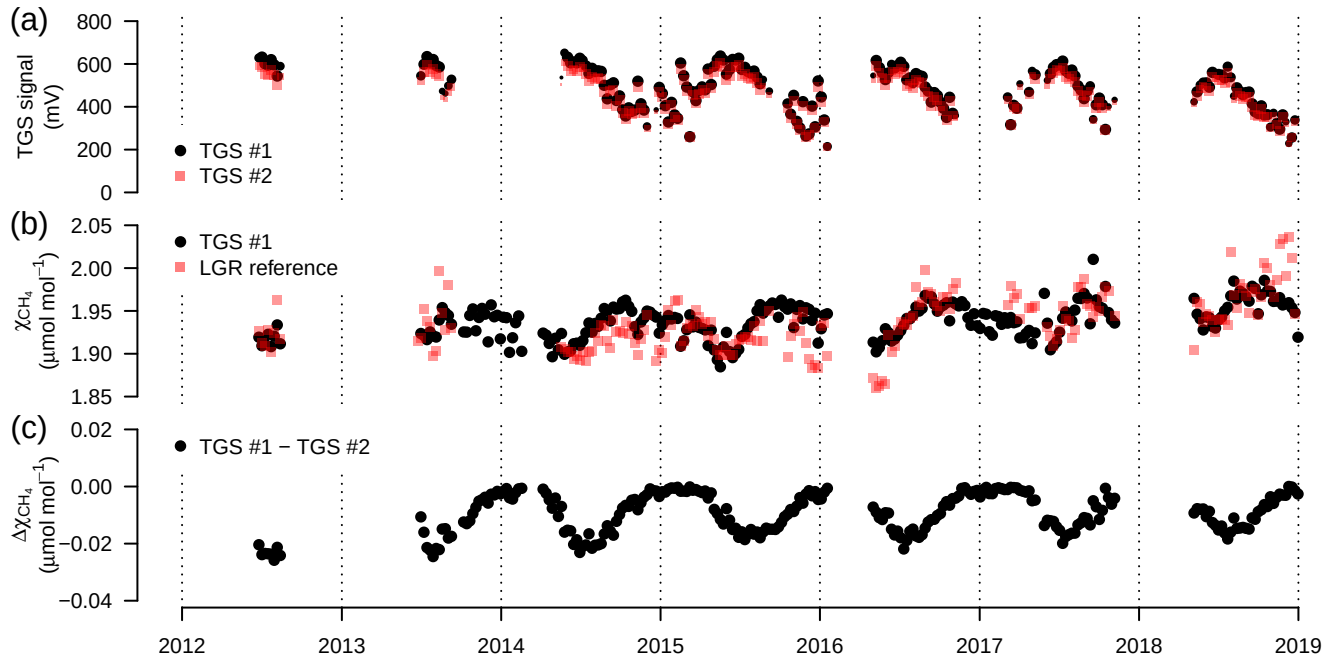


Figure 9. (a) Weekly median sensor signals from both TGS 2600 sensors, (b) ~~relative signal difference between the two TGS 2600 sensors in percent of the signal measured with the primary sensor #1, and (c)~~ CH_4 derived with Eq. (2) for TGS sensor #1 and measured by the Los Gatos Research reference instrument, and (c) absolute difference between the two TGS 2600 sensors. The signals from both sensors were converted to CH_4 using Eq. (2) parameterized with data from TGS sensor #1. Symbol size is proportional to relative data coverage.

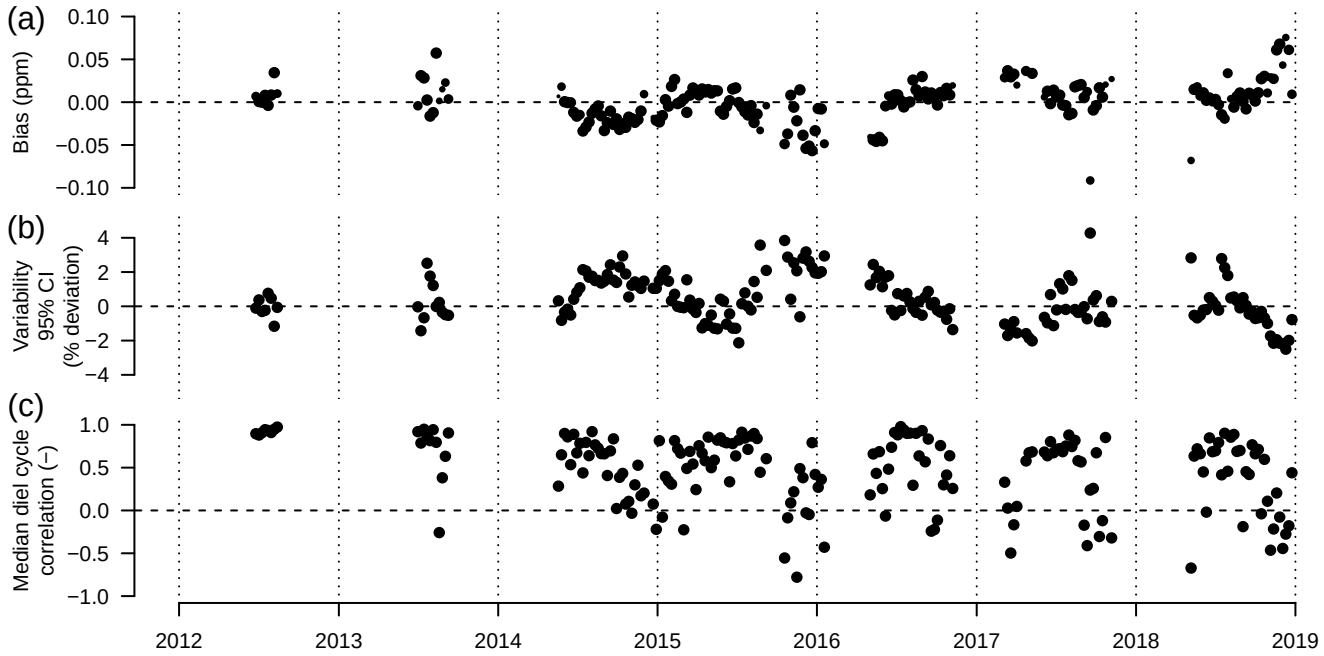


Figure 10. (a) Weekly median bias, (b) median variability, and (c) correlation between weekly median diel cycles of TGS 2600 sensor #~~1.~~1 and the reference. Symbol size is proportional to relative data coverage.

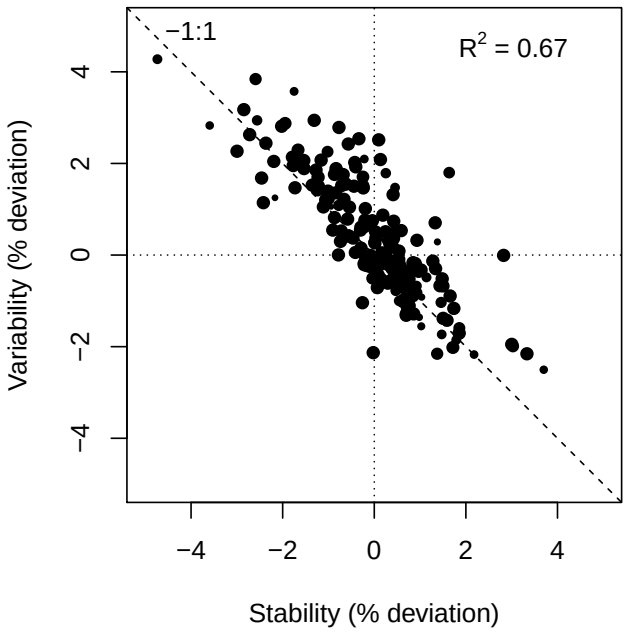


Figure 11. Variability and stability (relative bias) of weekly median TGS 2600 sensor #1 are inversely related and plot along the –1:1 line. Symbol size is proportional to relative data coverage.

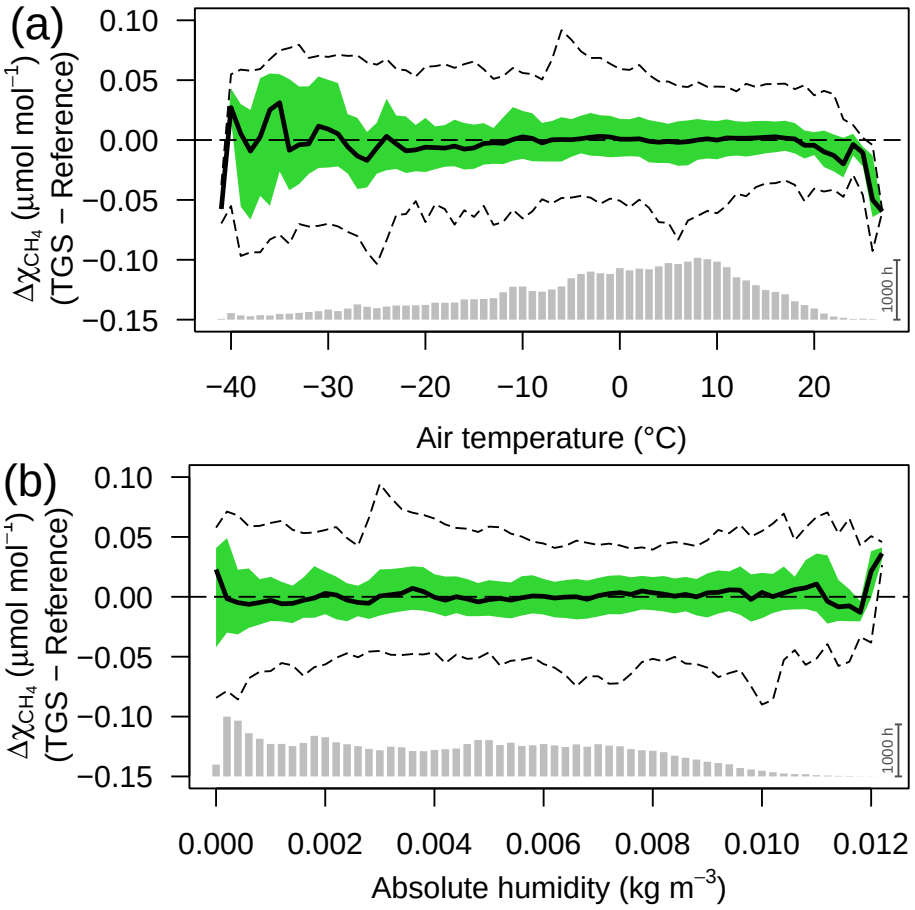


Figure 12. Residuals of 30-minute averaged TGS 2600 vs. reference CH_4 measurements as a function of (a) air temperature, and (b) absolute humidity. Colored areas show the inter-quartile range (50% CI), bold lines show the median, and dashed lines show the bin-averaged 95% confidence interval. Bin size was 1 °C and 0.0002 kg m^{-3}, respectively. Gray bars at the bottom show number of 30-minute averages in each bin; the scale bar (1000 hours) at right specifies their size.

Table 1. Goodness of fit of TGS 2600 (sensor #1) derived CH_4 ~~concentrations~~ mole fractions (30-minute averages) obtained from a linear model using air temperature and absolute humidity (Eq. 2), a heat loss model (Eq. 3), and an artificial neural network (ANN). For the goodness of fit the coefficient of determination (R^2) and the root mean square error of the residuals (RMSE) are reported for the overall model and separately for warm and cold conditions. The parametrization of the linear model given in Eq. (2) used the entire period 2012–2018. For a more rigorous model test, all three approaches were calibrated with the data measured in years 2014–2016, and the remaining data (2012–2013 and 2017–2018) were used for validation.

	Linear Model			Heat Loss Model		Artificial Neural Network	
	Entire period	Calibration	Validation	Calibration	Validation	Calibration	Validation
Overall							
R^2	~~0.42~~ 0.424	~~0.45~~ 0.447	~~0.21~~ 0.207	~~0.17~~ 0.166	~~0.28~~ 0.284	~~0.51~~ 0.311	~~0.54~~ 0.282
RMSE (~~ppm~~μmol mol^{-1})	~~0.03~~ 0.030	~~0.03~~ 0.026	~~0.04~~ 0.041	~~0.03~~ 0.032	~~0.05~~ 0.046	~~0.03~~ 0.030	~~0.03~~ 0.043
Warm conditions ($T_a \geq 0$ °C)							
R^2	~~0.48~~ 0.476	~~0.52~~ 0.518	~~0.29~~ 0.288	~~0.18~~ 0.180	~~0.18~~ 0.181	~~0.56~~ 0.278	~~0.47~~ 0.265
RMSE (~~ppm~~μmol mol^{-1})	~~0.03~~ 0.027	~~0.03~~ 0.026	~~0.03~~ 0.032	~~0.03~~ 0.034	~~0.04~~ 0.039	~~0.03~~ 0.032	~~0.03~~ 0.036
Cold conditions ($T_a < 0$ °C)							
R^2	~~0.32~~ 0.322	~~0.34~~ 0.345	~~0.03~~ 0.034	~~0.16~~ 0.157	~~0.05~~ 0.055	~~0.44~~ 0.314	~~0.37~~ 0.092
RMSE (~~ppm~~μmol mol^{-1})	~~0.03~~ 0.033	~~0.03~~ 0.027	~~0.05~~ 0.052	~~0.03~~ 0.031	~~0.05~~ 0.055	~~0.03~~ 0.028	~~0.03~~ 0.053