1. Methods

Installation of the sensors in the manhole

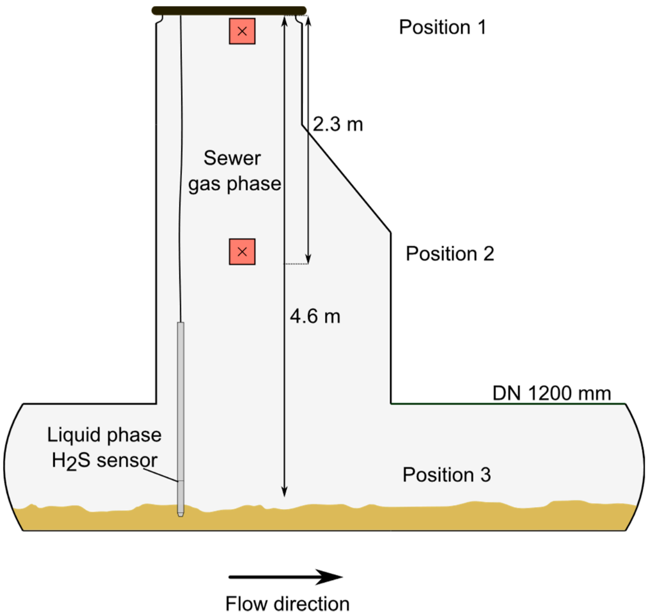

Figure S1. Installation of the sensors in the manhole used for comparing the sensors. The red boxes marks the positions (heights above the wastewater surface) the sensors were installed.

Response Test Set-up

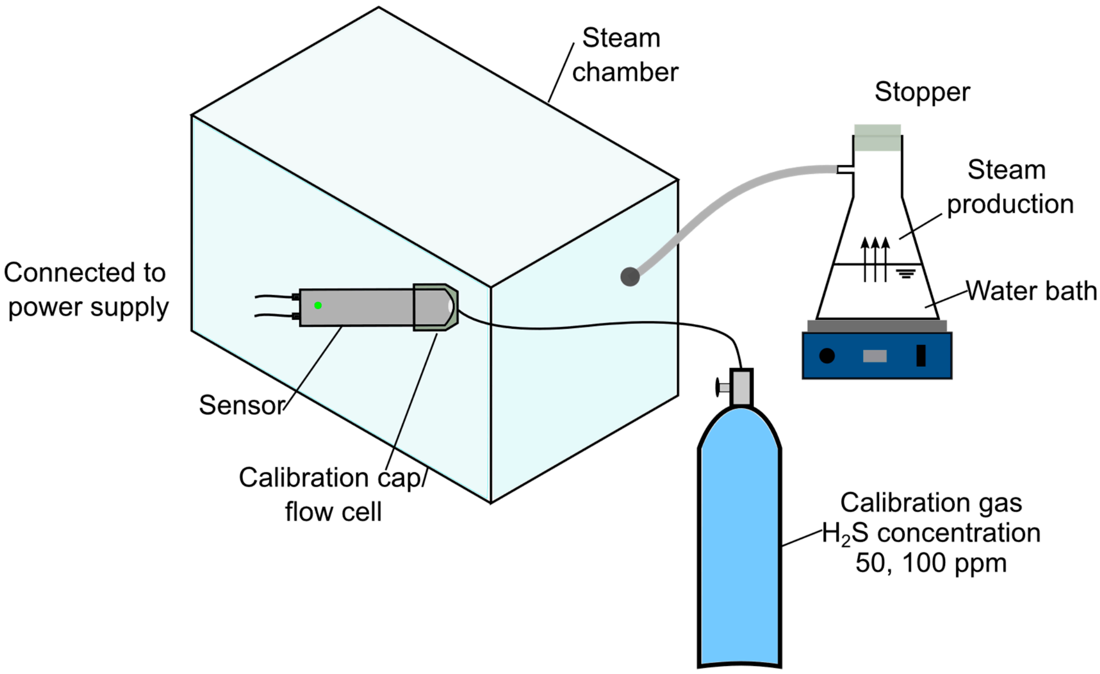

Figure S2. Set-up of the control test. To simulate extreme moisture conditions, we produced steam by preparing a water bath in a vessel that was sealed. The steam produced was directed into the chamber where the sensor was installed. The response test made under ambient conditions was also carried out in the chamber, however, without using the water bath. In instances, H_2S gas of known concentration (50 ppm) was supplied to the flow cell directly covering the sensor's measuring point. Each test was carried out for 5 minutes.

Zero-Offset Correction

Correction of the SulfiLogger™'s zero offset was carried out by computing a linear regression between the reference (OdaLog®) and the sensor measurements. The intercept of the regression was subtracted from the SulfiLogger™ measurements to correct the offset. It was observed that the offset decreased with time, therefore the linear regression and the corresponding correction were computed for each experimental day. Negative values were set to zero as negative H_2S concentrations cannot be measured.

Bland-Altman Plot

Bland and Altman [1] presented a graphical method to easily analyse and compare two methods (in this case sensors) for measuring the same variable, based on quantifying the variance between method differences. This approach is commonly used to compare, for example, an existing method with a new one. The agreement between both methods is quantified by determining limits of agreements ($LoA = \bar{y} \pm 1.96\,\mathrm{s}$), which are based on calculations using the mean ($\bar{y}$) and the standard deviation (s) [2].

The Bland-Altman-Plot requires that the method differences are normally distributed in order to calculate the limits of agreement. This approach might be difficult to implement for practitioners on-site since sensor differences are not always normally distributed and would require data transformation processes previous to their evaluation.

Procedure for computing difference plots

The following procedure is proposed to create the difference plots:

1. An acceptable range for the measurement differences between sensors must be defined beforehand. For this work, this range was set at 0 ± 5 ppm. Moreover, threshold values for assessing the degree of agreement between sensors need to be previously established. The scoring used in this work to evaluate the degree of agreement is provided in Table 2.
2. The measured differences of the H_2S concentration between the reference and the online sensor (y-axis) are plotted over time (x-axis).

$$Measured\ differences = c_{H_2S,\ Reference\ Sensor} - c_{H_2S,\ Online\ Sensor} \tag{1}$$

3. The line of equality is plotted at $y = 0$.
4. The mean bias difference is determined and added to the plot. It should be approximate to zero. The closer the mean bias difference is to zero, the better the agreement between sensors.

2.5. Performance Indicators – Ranking System

When considering adding or buy new sensors to increase monitoring options in sewer systems, several sensors might be evaluated at a time. To ease the sensor comparison, we propose a practical method based on a list with performance indicators and an attached ranking system. The performance indicators are grouped in three main categories: mode of operation, reliability and economic aspect. A more detailed description of the indicators is provided in Table S1.

Table S1. List and description of the performance indicators

Indicator	Description
Mode of Operation	
Ease of calibration	A measure to determine how easily the sensors can be calibrated. This indicator is based on the personal experience of the authors when calibrating the sensors.
Ease of installation	A measure to determine how easily the sensors can be installed in the sewer system. This indicator is based on the personal experience of the authors when installing the sensors at different sites.
Real-time access to data (min or h)	Measures the time between the collection of the data through the sensor and the time when the data can be accessed online. This indicator is based on the personal experience of the authors when accessing the online data.
Reliability	
Accuracy/Uncertainty (ppm)	Closeness of the agreement between the online sensor and the reference sensor (true value) [3]. This indicator is computed based on the response test data.
Mean Bias (ppm)	See definition in section 2.4.1. This indicator can be computed based on the response test or the field data. For this work, we used the response test results.
Repeatability (ppm)	Closeness of the agreement between successive H_2S measurements carried out at a known gas concentration under unchanged conditions [3]. This indicator can be computed based on the response test or the field data. For this work, we used the response test results.
Sensitivity to humid conditions	A relative measure describing the change in the relative humidity and the corresponding change in the measured H_2S concentration [4]. This indicator is computed based on the response test data.
Economic aspect	
Cost (€)	Price of the sensors and the installation accessories. This indicator is determined by the manufacturer´s/vendor´s price.

To provide an overall overview of the sensor performance, a ranking system can be established. In this case, sensors are ranked in each category with a number between one and three, being one the best and three the worst performance level. If two sensors perform equally, they will be ranked at the same level. An example of the sensor ranking is presented in Table S2.

Table S2. Example of the sensor ranking according to the results from the performance indicators.

	Sensor 1	**Sensor 2**	**Sensor 3**
Ease of calibration	1	3	2
Ease of installation	2	2	1
Real time access to data	3	1	2

3.4 Performance Indicators – Ranking System

To further assess the suitability of the sensors for our sewer application, sensors evaluated in this work were ranked according to the performance indicators presented in Table S2. This system provides an overview of where the strengths and weaknesses of sensors are located. It is designed as a guide to identifying categories that are of special importance for each practitioner and/or sewer application. The final decision must be based on the properties required for the studied application, as this is dependant on several factors. For example, high accuracy should be essential when monitoring in sensitive areas or when evaluating mitigation strategies to get below a previously set threshold value. On the other hand, for assessing hotspots for odor and corrosion, easy calibration and installation could be a priority.

The results of the ranking system carried out in this work are presented in Table S3 and are shortly explained below.

Table S3. Sensor ranking based on the performance indicators described in Table S1.

	OdaLog® Logger L2	**Sulfilogger™**	**MyDatasensH2S1000 BLE**
Mode of Operation			
Ease of calibration	2	1	3
Ease of installation	1	2	1
Real time access to data (min or h)	3	2	1
Reliability			
Accuracy/Uncertainty (ppm)	2	1	2
Mean Bias (ppm)	2	1	3
Repeatibility (ppm)	1	3	2
Sensitivity to humid conditions	1	3	2
Economic aspect			
Cost (€)	~1,800	~ 6,000 (incl. installation accessories)	~ 4,200 (incl. maintenance service)

- Ease of calibration: MyDatasensH2S1000 BLE ranked third because it cannot be calibrated by a practitioner on site. It can only be calibrated by the manufacturer. Both OdaLog® and SulfiLogger™, can be calibrated on-site with calibration gas and a flow cell. Moreover, SulfiLogger™ is calibrated then automatically and is, therefore, ranked first.

- Ease of installation: , SulfiLogger™ needs to be installed with an external battery (PowerCom box) and therefore is ranked second in this category.
- Real-time access to data: MyDatasensH2S1000 BLE provided the most reliable and fastest data transmission (1st). Data upload to the server was often delayed by , SulfiLogger™ and needed to be activated manually on site. The OdaLog sensor used in this study does not provide only data transmission and is therefore ranked as last.
- Accuracy: According to the response test results under ambient conditions, Sulfilogger™ had the highest accuracy, followed by OdaLog® and then by MyDatasensH2S1000 BLE.
- Mean Bias: According to the response test results under ambient, Sulfilogger™ had the lowest mean bias, followed by OdaLog® and then by MyDatasensH2S1000 BLE.
- Repeatability: This parameter was computed using the measurements from the control test after the sensor has stabilised (3-5 min). The sensor showing the lowest differences between the repeated measurements is ranked first.
- Sensitivity to humid conditions: According to the response test results under extreme humid conditions, OdaLog® and MyDatasensH2S1000 BLE improved their performance slightly, with the OdaLog® having the lowest mean bias. On the other hand, the accuracy of the Sulfilogger™ decreased under these conditions and is, therefore, ranked third in this category.

3. References

1. Bland, J.M.; Altman, D.G. Measuring Agreement in Method Comparison Studies. Stat Methods Med Res 1999, 26, doi:10.1177/096228029900800204.

2. Giavarina, D. Understanding Bland Altman Analysis. Biochem. Medica 2015, 25, 141–151, doi:10.11613/BM.2015.015.

3. JCGM (Joint Committee for Guides in Metrology) International Vocabulary of Metrology – Basic and General Concepts and Associated Terms (VIM). 2012, 3rd edn.

4. Waldmann, C.; Tamburri, M.; Prien, R.D.; Fietzek, P. Assessment of Sensor Performance. Ocean Sci. 2010, 6, 235–245, doi:10.5194/os-6-235-2010.