# CH4Net: a deep learning model for monitoring methane super-emitters with Sentinel-2 imagery

Anna Vaughan[1], Gonzalo Mateo-García[2,3], Luis Gómez-Chova[3], Vít Růžička[4], Luis Guanter[5,6], and Itziar Irakulis-Loitxate[5,7]

[1] Computer Laboratory, University of Cambridge, UK
[2] Trillium Technologies Ltd., London, UK
[3] Image Processing Laboratory, University of Valencia, Valencia, Spain
[4] University of Oxford, Oxford, UK
[5] Universitat Politècnica de València, Valencia, Spain
[6] Environmental Defense Fund, Reguliersgracht 79, 1017 LN Amsterdam, the Netherlands
[7] International Methane Emission Observatory, United Nations Environment Program, Paris, France.

**Correspondence:** Anna Vaughan (av555@cam.ac.uk)

**Abstract.** We present a deep learning model, CH4Net, for automated monitoring of methane super-emitters from Sentinel-2 data. When trained on images of 23 methane super-emitter locations from 2017-2020 and evaluated on images from 2021 this model detects 84% of methane plumes compared with 24% of plumes for a state-of-the-art baseline while maintaining a similar false positive rate. We present an in depth analysis of CH4Net over the complete dataset and at each individual super-emitter site. In addition to the CH4Net model we compile and open source a hand annotated training dataset consisting of 925 methane plume masks as a machine learning baseline to drive further research in this field.

## 1 Introduction

As a potent greenhouse gas responsible for approximately 25% of warming since the industrial revolution (Stocker, 2014; Varon et al., 2021) with rapidly increasing atmospheric concentrations (Tollefson, 2022), curbing methane emissions is an important step in combating the climate crisis. Anthropogenic emissions emanate from diverse sources, principally associated with livestock, agriculture, landfills, and the fossil fuel industry (oil and gas extraction and coal mining) (Saunois et al., 2020; Maasakkers et al., 2022). Of particular interest for rapid suppression of emissions are super-emitters, defined to be sources in the top 1% of global anthropogenic methane emitters, corresponding to an approximate flow rate of 25 kg/h (Zavala-Araiza et al., 2017). These sources contribute a substantial fraction of all methane emissions in the oil and gas sector (Alvarez et al., 2018), providing an opportunity to rapidly limit emissions with mitigation at a reasonable cost (Lauvaux et al., 2022).

Over the past five years, remote sensing instruments have been extensively utilised for detecting and monitoring super-emitters (Irakulis-Loitxate et al., 2022; Lauvaux et al., 2022; Varon et al., 2021; Maasakkers et al., 2022; Irakulis-Loitxate et al., 2021). To monitor these point sources, it is necessary to use point source imagers, instruments with a spatial resolution of less than 60 m (Jacob et al., 2022). In addition to this, the ideal instrument would also have global coverage, rapid revisit time, and high spectral resolution in the 1700 and 2300 nm short wave infrared spectral windows where methane absorption is the strongest. Unfortunately, no currently available instrument has all of these desired characteristics.

Hyperspectral instruments, for example PRISMA and EnMAP, produce more accurate methane retrievals because they are more sensitive to small concentrations (Jacob et al., 2022; Guanter et al., 2021). However, they have limited swaths (30km) and image acquisitions need to be tasked –request to the ground segment to acquire a particular area of interest– therefore they have limited data availability.

An alternative approach is to utilise multispectral imagery such as Sentinel-2 (Drusch et al., 2012) and Landsat-8 and 9 (Roy et al., 2014). These instruments have relatively rapid revisit time (approximately five days for Sentinel-2 and 16 days for Landsat at the equator) and high (20-30m) spatial resolution. They, however, have significantly degraded spectral resolution compared to hyperspectral instruments, resulting in a lower sensitivity to methane (Sherwin et al., 2023). Recent works have demonstrated successful detection and quantification of large plumes from Sentinel-2 imagery (Varon et al., 2021; Ehret et al., 2022; Irakulis-Loitxate et al., 2022). These approaches are based on temporal differences and ratios between Sentinel-2 bands 11 (1560–1660 nm) and 12 (2090–2290 nm). Band 12 strongly overlaps with the methane absorption feature, while band 11 provides an estimate of the background at a relatively similar wavelength. Varon et al. (2021) present a series of approaches differencing between S2 bands 11 and 12 to quantify methane emissions. Their most successful approach quantifies emissions down to a rate of 3 t/h (tons of CH4 emitted per hour) by taking the difference of bands 11 and 12 comparing two consecutive passes, however, remains sensitive to surface artefacts. Ehret et al. (2022) take a similar approach projecting onto a time series of 30 previous images with two-stage linear regression and a manual verification step to identify the presence of false positives caused by surface artefacts. There are two significant limitations with these methods. The first and most important is that they remain sensitive to surface artifacts, often requiring manual verification. The second is that a time series of images is required.

In this study, we ask the question: “for a known set of methane super-emitters, is it possible to accurately identify plumes in Sentinel-2 imagery to monitor future emissions?”. This has the important application of assessing whether mitigation work on existing emissions has been successful. We train a machine learning model, CH4Net, to segment methane plumes from a single image. In contrast to previous methods, CH4Net learns background characteristics of the sites by processing multiple passes over each location during training without the need for a time series of previous images, reference image, or manual verification step. Machine learning has been successfully applied to segmenting plumes in hyperspectral data (Groshenry et al., 2022; Jongaramrungruang et al., 2022; Schuit et al., 2023), however, this methodology has not yet been applied to Sentinel-2 imagery as a sufficiently large dataset of verified plumes was unavailable. We first collect and annotate a dataset of methane plumes from known super-emitters in Turkmenistan (Irakulis-Loitxate et al., 2022), a semi-arid region with strong emissions providing the best-case scenario for multispectral methane imaging. This is used to train a deep learning model to segment methane plumes from the background. We evaluate this model for a future time period for the training locations. In addition, we show that the model can successfully be applied to monitor a super-emitter at a new location in the same region unseen at training time. The aims of this paper are as follows:

1. Collect and label a machine learning dataset of methane plumes in Sentinel-2 imagery.

2. Develop an automated plume segmentation system. In contrast to existing works, this is a fully automated system that does not require a time series of Sentinel-2 images or identification of a reference image at test time.

3. Apply this system to track emissions from a selection of known methane super-emitters during a future time period.

Section 2 presents an overview of dataset collection, the CH4Net architecture and training procedure. Results are presented in Section 3 and 4, with conclusions and a discussion in Section 5.

## 2 Methods

### 2.1 Dataset collection and processing

We first collect and manually annotate a dataset of methane plumes from Sentinel-2 images from 2017-2021 consisting of 10,046 0.01×0.01 degree images ( 200×200 pixels) from Sentinel-2 L1C scenes centred on 23 known super-emitter locations in Turkmenistan (Irakulis-Loitxate et al., 2022). Several locations identified are in close proximity to each other, and are combined into a single scene. For a map and complete list of included sites, see Figure 1 and Table 1. For each site all available images were downloaded using the Sentinel Hub API, each consisting of the 13 scaled and harmonized Sentinel-2 channels (Sinergise Ltd., 2023). Images containing clouds are deliberately not discarded to allow the model to learn a mapping robust to these features without the need for costly pre-processing steps. We note that the model output is therefore predicting whether a plume is visible in the scene or not; it is possible that an emission may be present but is covered by clouds. Cloudy scenes could easily be discarded if necessary for a particular application by applying a cloud detection model (Jeppesen et al., 2019; López-Puigdollers et al., 2021; Aybar et al., 2022).

We frame methane detection as a binary segmentation problem, where a pixel is classified as either 0, if not part of a plume, or 1, if part of a plume. To manually label the plumes, enhanced images were created for each time-step using the multi-band multi-pass (MBMP) method developed by Varon et al. (2021). A clear-sky reference image was chosen for each location, with the multi-band multi-pass image given by

$$MBMP = \frac{cR_{12} - R_{11}}{R_{11}} - \frac{c'R'_{12} - R'_{11}}{R'_{11}}$$

where $R_{11}$ and $R_{12}$ are the raw Sentinel-2 band 11 and 12 observations for the current image, $R'_{11}$ and $R'_{12}$ are the raw Sentinel-2 band 11 and 12 observations for the reference image, and $c$ ($c'$) is calculated by least-squares regression of $R_{11}$ against $R_{12}$ ($R'_{11}$ against $R'_{12}$) for all pixels. These images were used to manually identify and label the extent of the methane plumes for each time-step. For examples of the MBMP images and corresponding hand-labelled plumes, see Figure 2. It is emphasized that these MBMP images are used as an auxiliary tool to guide annotation only and are not included as input predictors to the final model.

Each data point consists of the 13 Sentinel-2 bands interpolated to a common resolution of 10m together with the hand-labelled plume mask for a total of 925 scenes containing a plume and 9121 without. The resolution of 10m is chosen as adding the highest resolution RGB channels improves the model performance, so all data is interpolated to this resolution to avoid loss

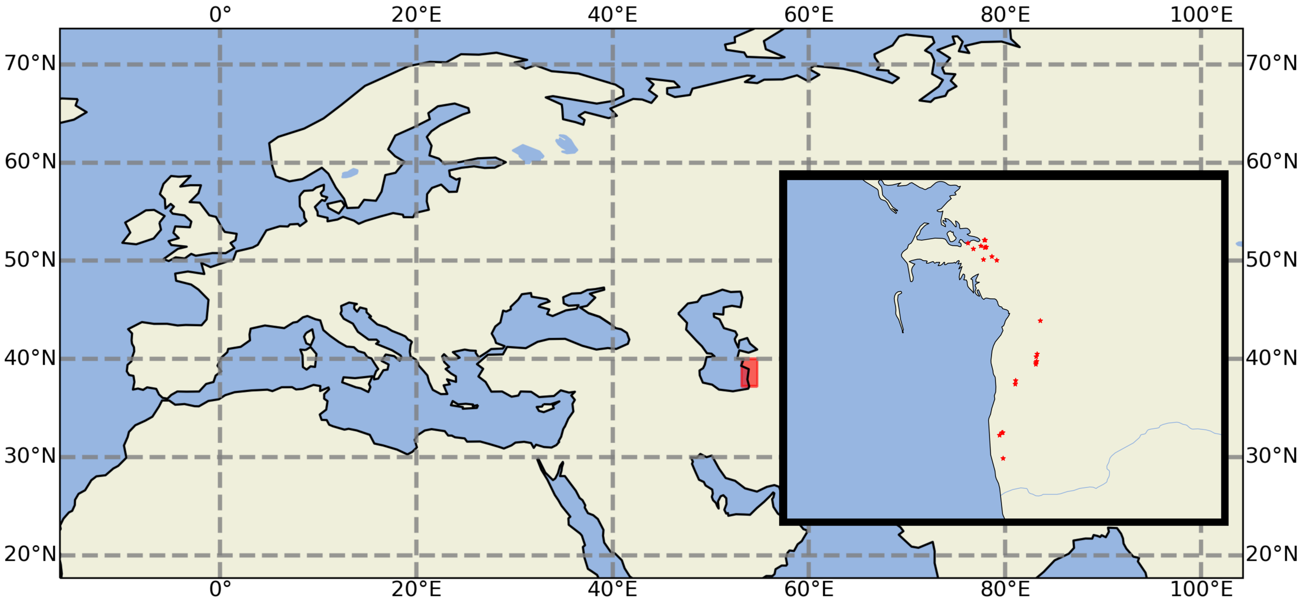

**Figure 1.** Locations of the 23 super emitters included in the dataset showing the study region shaded in red and precise locations (inset).

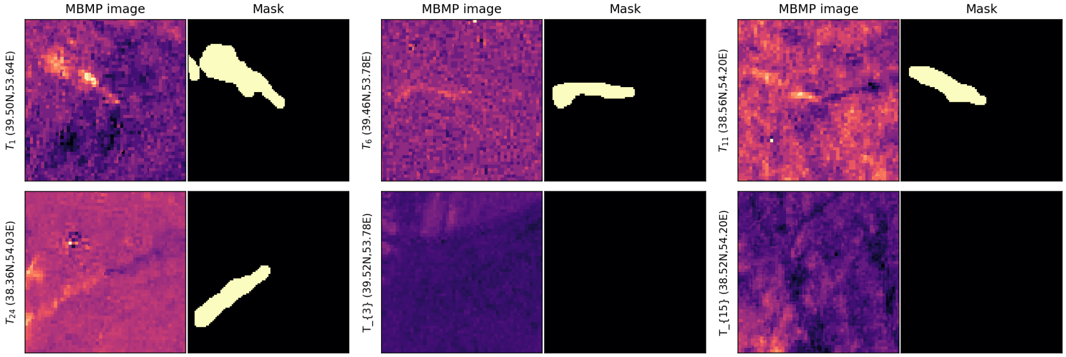

**Figure 2.** Examples of the MBMP images and corresponding hand annotated masks.

of information. We emphasize that only a single timestep is required at test time, unlike in previously proposed methods where multiple timesteps are required. This removes the requirement to identify a clear sky reference image or series of images, which typically requires manual selection, and is simpler to deploy and maintain.

This dataset is split into train, test and validation sets:

- train: all images from 2017-2020 excluding the validation set

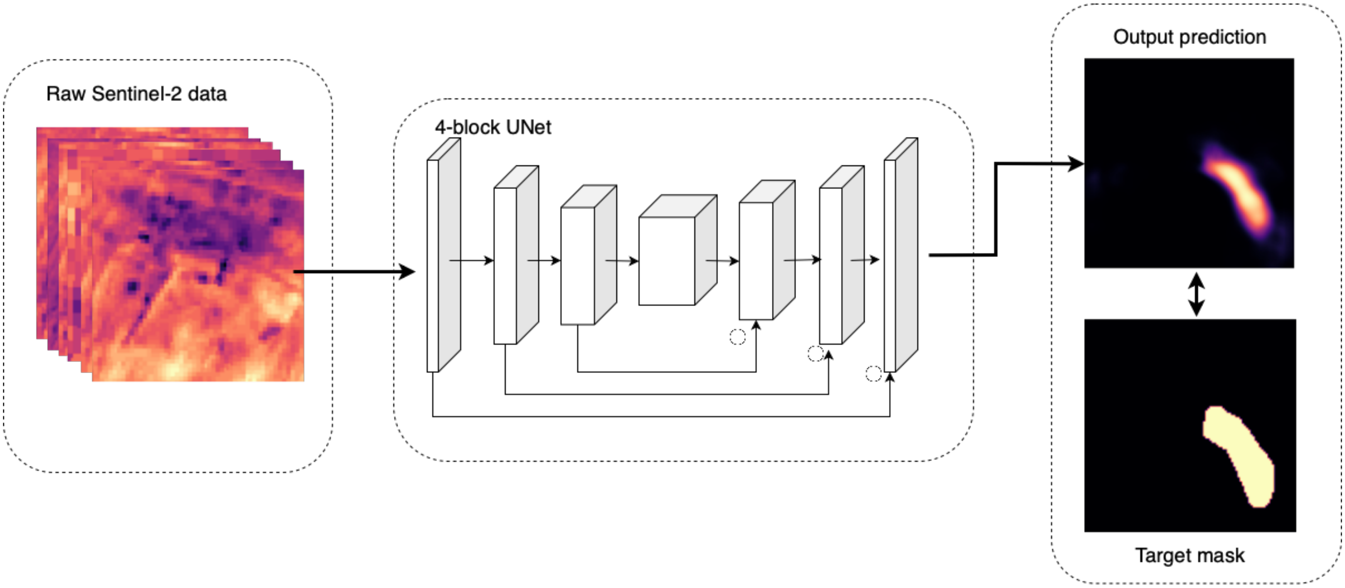

**Figure 3.** Schematic of the CH4Net model architecture showing the Sentinel-2 bands input to the UNet and probabilistic output compared to the hand-annotated mask.

- validation: a held out randomly subsampled selection of 256 train images stratified by plume presence.
- test: all images from 2021

The validation split is used for model selection and we use the test set to report results. As a baseline, we consider a MBMP approach based on that outlined by (Irakulis-Loitxate et al., 2022). To calculate the baseline prediction the multiband-multipass image is constructed for each image. This is denoised using a Gaussian filter, then thresholded to identify clusters of pixels with values more than two standard deviations below than the mean. Resulting clusters are kept as a predicted plume if they contain more than 115 pixels.

## 2.2 Model architecture and training

The detection model uses a simple and flexible UNet architecture Ronneberger et al. (2015) consisting of 4 encoder blocks (2D convolution layer, batch norm, ReLU activation, 2D convolution layer, batch norm ReLU activation, maxpool) followed by four decoder blocks (transposed 2D convolution layer, 2D convolution layer, batch norm, ReLU activation, 2D convolution layer, batch norm ReLU activation) with skip connections between blocks of corresponding scale. Channel output dimensions for each of these blocks are $\{128, 256, 512, 512, 256, 128, 64, 128, 1\}$ with kernel sizes of 3 for all convolution layers and 2 for the max pooling layers. For a complete schematic of the model see Figure 3. This model takes the Sentinel-2 bands as input and outputs a pixelwise prediction of the probability (between 0 and 1) of the pixel being part of a methane plume.

The UNet is trained on the training dataset described above with Binary Cross-Entropy loss, Adam optimisation (Kingma and Ba, 2014) and a learning rate of $1e-4$ for 250 epochs. As the dataset is unbalanced with significantly more negative than positive images, at each epoch $n$ negative images are randomly sampled, where $n$ is the total size of the positive image set. To prevent over-fitting, augmentation is applied by cropping a random 100x100 pixel scene from the larger image tiles. In order to investigate the optimal predictor set, the UNet is trained with both bands 11 and 12 only as predictors (11+12), and all bands (ALL).

## 3 Results: All images

We first evaluate the skill of CH4Net at correctly identifying whether a given image contains a methane plume. This is referred to as scene-level prediction, as opposed to pixel-level prediction. For scene-level prediction the probabilistic predictions are transformed to a binary prediction by defining a methane plume as a contiguous region of greater than 115 pixels with probability greater than or equal to 0.25. The 115 pixel threshold is chosen as this is the size of the smallest plume contained in the training set, while the 0.25 threshold is selected to maximize the balanced accuracy score. A scene is classified as 1 (containing a plume) if such a feature is present and 0 otherwise.

Accuracy, balanced accuracy, precision, recall, false positive rate, and false negative rate for both the ALL and 11+12 experiments over the 2021 images are shown in the upper portion of Table 1. The model with all bands included as predictors outperforms that with only bands 11 and 12, indicating that other bands add value for methane detection, or for the reduction of false positives. Results over the test set for the model with all bands included (bands 11+12 only, the MBMP baseline) are accuracy 0.80 (0.69, 0.50), balanced accuracy 0.76 (0.75, 0.71), precision 0.30 (0.24,0.11), recall 0.84 (0.61,0.24), false positive rate 0.24 (0.23,0.23) and false negative rate 0.16 (0.39,0.76). The model with all bands included outperforms that with only bands 11 and 12 on all metrics except for false positive rate which is slightly higher. CH4Net outperforms the baseline substantially on all metrics except for the false positive rate which is very slightly higher for ALL and the same for 11+12. The new model detects 83% of all plumes in the validation set compared to 24% for the baseline whilst producing a similar number of false positives, a large improvement in performance.

A more challenging task is to assess prediction skill at a pixel level, quantified by balanced accuracy and IoU over all pixels. Results on these metrics are shown in the lower section of Table 1. The model trained with all bands achieves a balanced accuracy (IoU) of 0.66 (0.57) compared to 0.66 (0.55) for the model with just bands 11 and 12, indicating that inclusion of other channels also improves performance at the pixel level. Both CH4Net models outperform the baseline, which achieves a balanced accuracy of 0.51 and IoU of 0.50.

**Table 1.** Scene and pixel level metrics over the test dataset (year 2021) for CH4Net trained with the complete 13 band predictor set (ALL), the bands 11 and 12 only predictor set (11+12) and MBMP baseline.

| Scene level metrics | | | |
|---|---|---|---|
| | ALL | 11+12 | MBMP Baseline |
| Accuracy | **0.80** | 0.69 | 0.50 |
| Balanced accuracy | **0.76** | 0.75 | 0.71 |
| False positive rate | 0.24 | **0.23** | **0.23** |
| False negative rate | **0.16** | 0.39 | 0.76 |
| Precision | **0.30** | 0.24 | 0.11 |
| Recall | **0.84** | 0.61 | 0.24 |
| Pixel level metrics | | | |
| Balanced accuracy | **0.66** | **0.66** | 0.51 |
| IoU | **0.57** | 0.55 | 0.50 |

## 4 Results by site

For a more nuanced assessment of skill at each individual location in the training set we produce predictions for all available images during the 2021 test period at each of the 23 sites. Results for each site are presented in Table 2. In all cases, these are generated using the optimal predictor set with all bands (ALL).

At a scene level, high accuracy is observed for a majority of sites, with accuracy greater than 75% for 19 out of 23 sites, and ranging from 0.57 to 0.71 for remaining sites. False positive rates range from 0.01 to 0.4, and false negative rates from 0.0 to 0.75, though are below 0.2 for a majority of sites.

At a pixel level, balanced accuracy ranges from 0.62 to 1.0, with 17 out of the 23 sites above 0.75. IoU (only defined for cases where at least one mask is available) ranges from 0.54 to 0.68.

To better understand the successes and limitations of this approach, we present several case studies, two of locations with excellent prediction quality (sites $T_7$ and $T_{17}$) and two with poor prediction quality (sites $T_1$ and $T_{11}$).

### 4.1 Case studies: sites $T_7$ and $T_{17}$ (high quality predictions)

For example, consider site $T_7$ where the prediction system has a balanced accuracy score of 0.83, with false positive rate of 0.20 and false negative rate of 0.12 for a site where 39% of scenes in the test set contain an emission. Figure 4 compares predictions to the observed values for scene-level classification. Overall predictions are in good agreement with observations, correctly identifying two emissions early in 2021 followed by a period of high emission activity which subsides towards the end of the year.

Predictions at site $T_{17}$ provide an example of correct prediction of multiple sporadic emission events over the course of the 2021 year. For this site the scene level accuracy is 0.90, false positive rate 0.11, false negative rate 0.0 and pixel level balanced

**Table 2.** CH4Net performance evaluated on all available images at the 23 super-emitter sites for 2021,showing (L-R) site ID, site longitude, site latitude, percentage of images containing a plume, scene level accuracy, scene level precision, scene level recall, false positive rate, false negative rate, pixel level balanced accuracy and pixel level balanced intersection over union (IoU)

| Site | longitude | latitude | % positive | Accuracy | Precision | Recall | FPR | FNR | Balanced accuracy (pixel) | IoU (pixel) |
|---|---|---|---|---|---|---|---|---|---|---|
| $T_1$ | 53.6367 | 39.49687 | 17.0% | 0.57 | 0.27 | 0.92 | 0.5 | 0.08 | 0.85 | 0.55 |
| $T_2$ | 53.77274 | 39.52148 | 0.0% | 0.94 | - | - | 0.06 | - | 1.0 | - |
| $T_3$ | 53.77903 | 39.52137 | 0.0% | 0.9 | - | - | 0.1 | - | 1.0 | - |
| $T_4$ | 53.74292 | 39.4739 | 1.0% | 0.9 | 0.06 | 1.0 | 0.1 | 0.0 | 0.93 | 0.55 |
| $T_5$ | 53.78836 | 39.46428 | 1.0% | 0.75 | 0.05 | 1.0 | 0.26 | 0.0 | 0.62 | 0.51 |
| $T_6$ | 53.77502 | 39.4616 | 38.0% | 0.9 | 0.8 | 0.96 | 0.14 | 0.04 | 0.81 | 0.68 |
| $T_7$ | 53.77921 | 39.45965 | 39.0% | 0.83 | 0.74 | 0.88 | 0.2 | 0.12 | 0.75 | 0.6 |
| $T_8$ | 53.68117 | 39.44955 | 0.0% | 0.93 | - | - | 0.07 | - | 1.0 | - |
| $T_9$ | 53.76506 | 39.36045 | 23.0% | 0.71 | 0.4 | 0.47 | 0.21 | 0.53 | 0.58 | 0.53 |
| $T_{10}$ | 53.83516 | 39.38584 | 0.0% | 0.93 | - | - | 0.07 | - | 1.0 | - |
| $T_{11}$ | 53.87509 | 39.35498 | 8.0% | 0.84 | 0.17 | 0.25 | 0.11 | 0.75 | 0.6 | 0.55 |
| $T_{12}$ | 54.23498 | 38.85515 | 15.0% | 0.85 | 0.5 | 0.27 | 0.05 | 0.73 | 0.59 | 0.56 |
| $T_{13}$ | 54.20931 | 38.57959 | 0.0% | 0.82 | - | - | 0.18 | - | 0.99 | - |
| $T_{14}$ | 54.20049 | 38.55747 | 37.0% | 0.75 | 0.62 | 0.85 | 0.3 | 0.15 | 0.77 | 0.63 |
| $T_{15}$ | 54.20393 | 38.51871 | 0.0% | 0.95 | - | - | 0.05 | - | 1.0 | - |
| $T_{16}$ | 54.19769 | 38.50798 | 0.0% | 0.95 | - | - | 0.05 | - | 1.0 | - |
| $T_{17}$ | 54.19764 | 38.49393 | 10.0% | 0.9 | 0.5 | 1.0 | 0.11 | 0.0 | 0.97 | 0.65 |
| $T_{18}$ | 54.02832 | 38.33078 | 16.0% | 0.75 | 0.39 | 0.92 | 0.28 | 0.08 | 0.76 | 0.55 |
| $T_{19}$ | 54.03149 | 38.36017 | 0.0% | 0.6 | - | - | 0.4 | - | 0.98 | - |
| $T_{20}$ | 53.89857 | 37.90825 | 16.0% | 0.77 | 0.41 | 0.92 | 0.26 | 0.08 | 0.75 | 0.59 |
| $T_{21}$ | 53.91623 | 37.9286 | 1.0% | 0.99 | 0.5 | 1.0 | 0.01 | 0.0 | 0.71 | 0.63 |
| $T_{22}$ | 53.92431 | 37.92913 | 23.0% | 0.75 | 0.48 | 0.71 | 0.23 | 0.29 | 0.63 | 0.54 |
| $T_{23}$ | 53.92702 | 37.71665 | 0.0% | 0.6 | - | - | 0.4 | - | 0.98 | - |

accuracy and IoU is 0.97 and 0.65, respectively. A more detailed view of predictions at a pixel scale is shown in Figure 5. This shows the observation mask compared to prediction overlaid on the RGB imagery for every available Sentinel-2 image in 2021. Both the occurrence and morphology of each plume is largely well captured, though two false positives are observed.

### 4.2 Case studies: sites $T_1$ and $T_{11}$ (low quality predictions)

We next examine two cases with comparatively poor prediction quality. Results for site $T_1$ are the worst out of all locations with at least one emission during 2021, with an accuracy of 0.57, false positive rate of 0.5 and false negative rate of 0.08. A time series of predictions compared to observations is shown in the upper panel of Figure 6. This demonstrates that the model

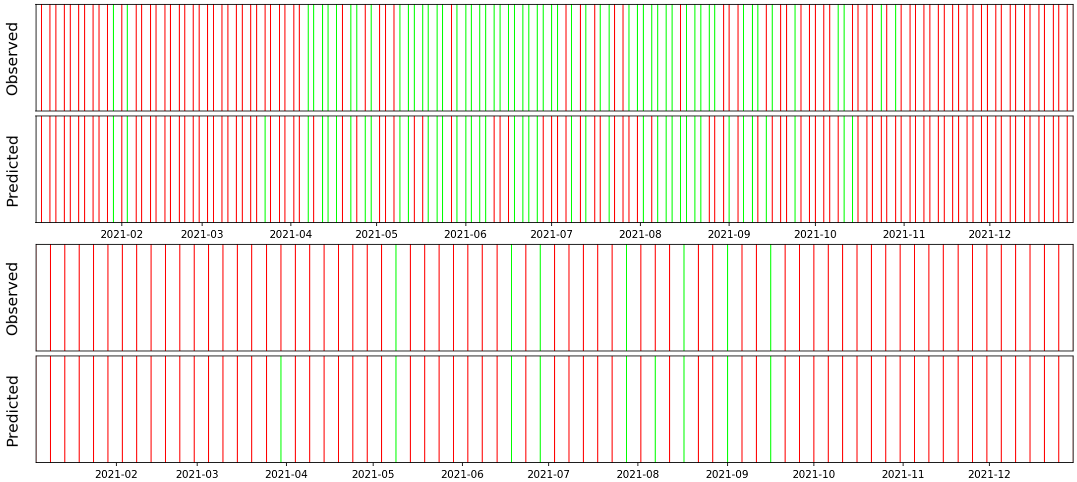

**Figure 4.** Time series of predictions for sites $T_7$ (top) and $T_{17}$ (bottom) over the test year (2021). Green (red) lines indicate that a plume was (not) observed or predicted. Observed ground truth values are shown in the upper time series and CH4Net predictions on the lower time series.

produces a high number of false positives, particularly through the second half of the year. Closer examination of individual predictions images indicates that there are three primary sources of false positives. Artifacts in the image (e.g., Fig. 7(a)) and thin clouds (e.g., Fig. 7(b)) produce occasional false positives throughout the time series. During the second half of 2021 multiple false positives are produced coinciding with a bright surface artifact visible in both the RGB and MBMP images (e.g., Fig. 7(c)). It is possible that this is a methane emission source, however, it is not labelled as such during the manual labelling as either the wind speed is too low to produce a clear plume or alternatively the emissions are weak with only the area immediately at the source detectable with the limited detection capability of Sentinel-2.

Site $T_{11}$ is an example of a site with multiple false negatives. For this location, the scene accuracy is 0.84, with a false positive rate of 0.11 however the false negative rate is the highest for all sites at 0.75. The prediction time series for this site is shown in the lower panel of Figure 6. Here the false negatives appear to arise in cases with heterogeneous background (which also often results in an increase in false positives). This is consistent with recent work indicating that the detection capability of Sentinel-2 is significantly lower in cases with a strongly heterogeneous background (Gorroño et al., 2023).

## 5 Conclusions

We have implemented CH4Net, the first fully automated system for monitoring known methane super-emitter sites, and produced the first large scale dataset of methane plumes in Sentinel-2 imagery. Model skill was assessed on multiple scene-level

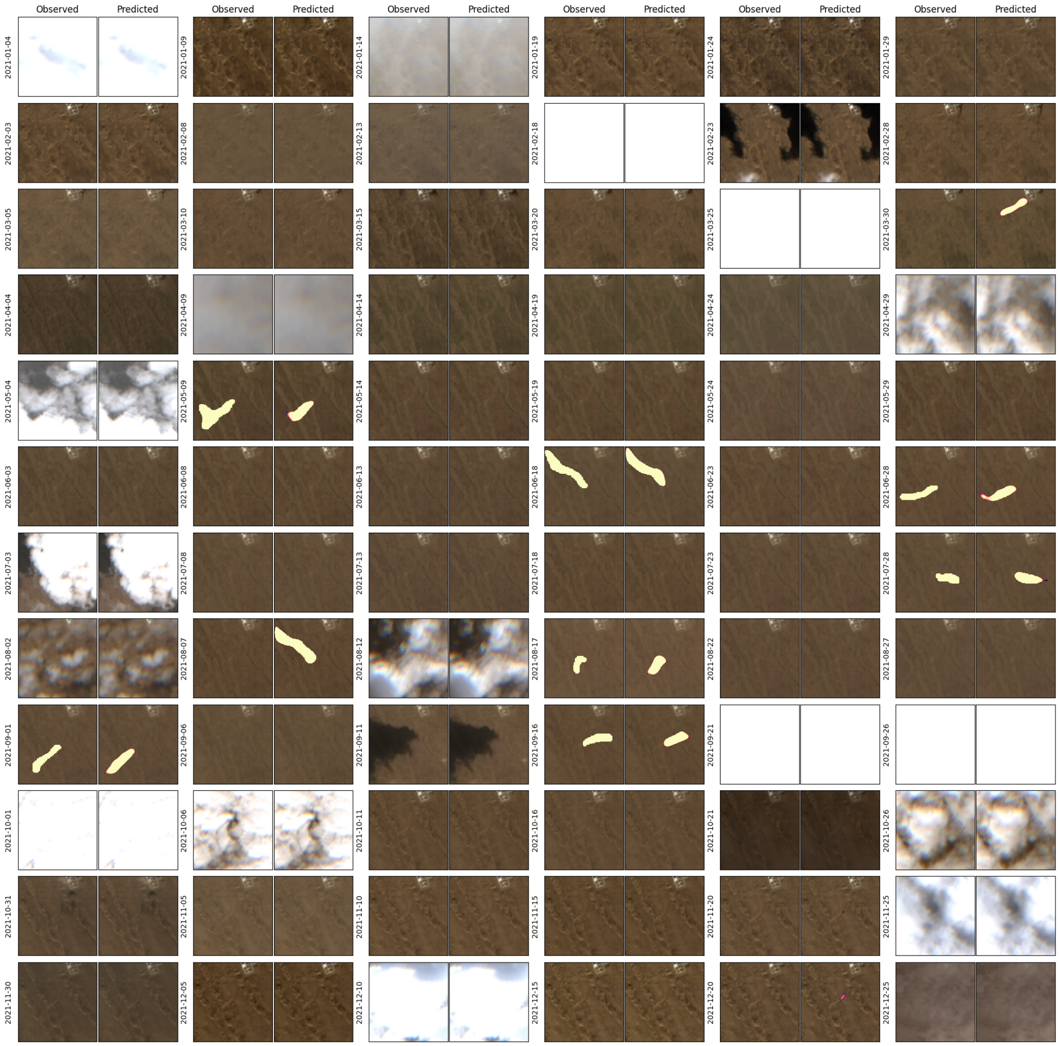

**Figure 5.** CH4Net pixel-level predictions for every image over site $T_{17}$ during 2021. For each time-step the observed mask (left) and probabilistic prediction (right) are shown overlaid on the RGB image.

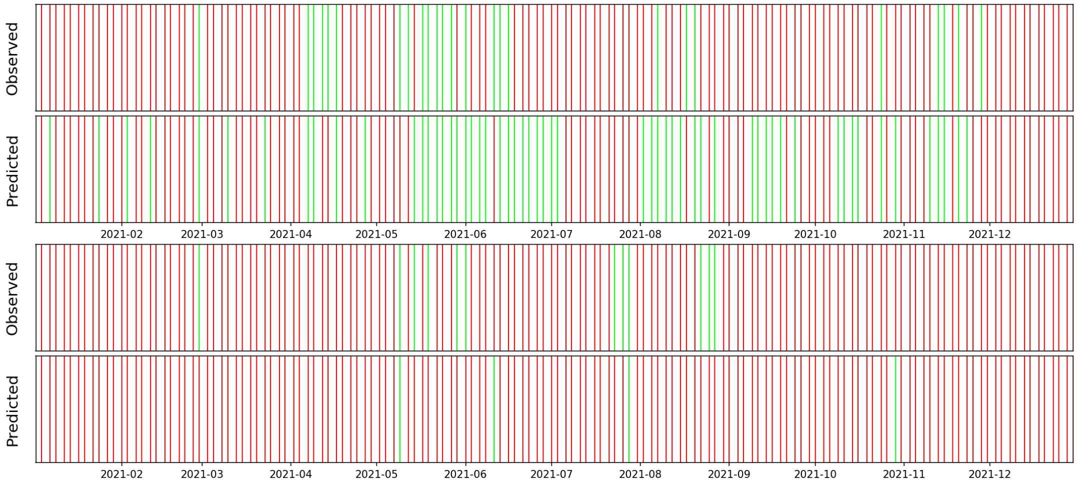

**Figure 6.** Time series of predictions for sites $T_1$ (top) and $T_{11}$ (bottom) over the test year (2021). Green (red) lines indicate that a plume was (not) observed or predicted. Observed ground truth values are shown in the upper time series and CH4Net predictions on the lower time series.

and pixel-level metrics, demonstrating that overall predictions are of high quality, though several sources of false positives and false negatives remain to be addressed. CH4Net comprehensively outperforms the multiband multipass baseline on all metrics except false positive rate where both methods perform similarly. These results offer promise for implementing ongoing tracking of known sources to mitigate emissions and provide early warnings when an event is observed.

In contrast to existing methods for methane plume detection in Sentinel-2 images (Varon et al., 2021; Ehret et al., 2022; Irakulis-Loitxate et al., 2022), this model requires only a single pass to generate predictions at test time and is fully automated. This creates a significant advantage in allowing large volumes of data to be processed without requiring costly manual verification. We believe that this is a significant breakthrough since, as it has been shown in other works (e.g. Irakulis-Loitxate et al. (2022)), emissions from a single site often recur over a long period of time. With this model we can envision a system that, when a new location is added, we can label past data, retrain the model and use it to produce notifications of new plumes on incoming Sentinel-2 acquisitions over that location. This is very useful to verify that leaks have been permanently fixed and to notify the emitters if this is not the case.

Further work is required in several areas to extend these results. One avenue for future work is improving the current monitoring methodology. For the dataset, priority for future work in this area is to collect further data over new areas and test whether CH4Net is suitable for application to other semi-arid locations. Furthermore, the accuracy of each mask could further be improved by having multiple annotators providing a mask for each image and taking the intersection over the proposed masks. A current shortcoming of this work is that the output of CH4Net provides only a binary mask as opposed to quantifying

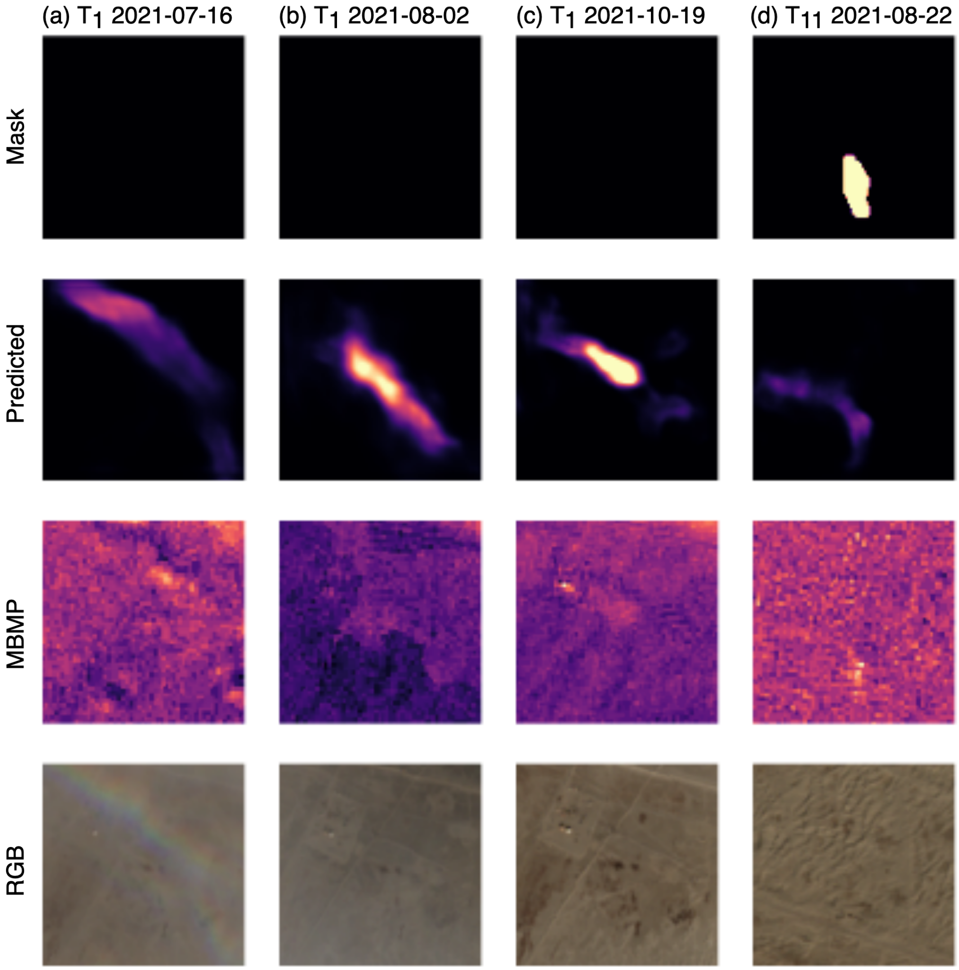

**Figure 7.** Examples of false positives and negatives for sites $T_1$ and $T_{11}$, showing: (a) false positive at site $T_1$ resulting from image artefact, (b) false positive at site $T_1$ resulting from thin cloud (not easily visible in the RGB window), (c) false positive at site $T_1$ resulting from potential low intensity methane source and (d) false negative at site $T_{11}$ resulting from strongly heterogeneous background.

the methane concentration at each pixel. Direct prediction of this quantity would allow for both emission occurrence and volume to be monitored. There are also a number of improvements that could be explored to improve the modelling methodology, including implementing scene level classification with a classification head, and implementing more sophisticated segmentation models such as vision transformers (Dosovitskiy et al., 2020). We hope that providing this dataset and baselines will lead to further work on machine learning models for this task.

A second avenue for future work is to explore training a similar model for scanning sentinel-2 images to discover new super-emitter sites. This would require collecting a much larger dataset of heterogeneous images (images from different locations and biomes), and training a model capable of limiting false positives in areas with highly heterogeneous backgrounds.

*Code and data availability.* Code and hand-annotated masks are available at https://github.com/annavaughan/CH4Net. Sentinel-2 data are available from Sentinel Hub https://www.sentinel-hub.com/.

*Author contributions.* A.V designed the study, implemented the code, labelled the dataset, conducted the experiments and wrote the first draft. All authors contributed to the analysis of results and final version of the paper.

*Competing interests.* The authors declare no competing interests.

*Acknowledgements.* Early stages of this project were funded as part of European Space Agency 3CS grant of Trillium Technologies reference `Starcop 1-2022-00380`. Authors gratefully acknowledge the support of the Trillium team and ESA technical officer. A.V. acknowledges the UKRI Centre for Doctoral Training in the Application of Artificial Intelligence to the study of Environmental Risks (AI4ER), led by the University of Cambridge and British Antarctic Survey, and studentship funding from Google DeepMind. G.M.-G. and L.G.-C. acknowledge support from the Spanish Ministry of Science and Innovation (project PID2019-109026RB-I00 funded by MCIN/AEI/10.13039/501100011033) and the European Social Fund. L.G.-C. and L.G. acknowledge support from the GVA PROMETEO programme (project CIPROM/2021/056).

## References

Alvarez, R. A., Zavala-Araiza, D., Lyon, D. R., Allen, D. T., Barkley, Z. R., Brandt, A. R., Davis, K. J., Herndon, S. C., Jacob, D. J., Karion, A., et al.: Assessment of methane emissions from the US oil and gas supply chain, Science, 361, 186–188, 2018.

Aybar, C., Ysuhuaylas, L., Loja, J., Gonzales, K., Herrera, F., Bautista, L., Yali, R., Flores, A., Diaz, L., Cuenca, N., Espinoza, W., Prudencio, F., Llactayo, V., Montero, D., Sudmanns, M., Tiede, D., Mateo-García, G., and Gómez-Chova, L.: CloudSEN12, a global dataset for semantic understanding of cloud and cloud shadow in Sentinel-2, Scientific Data, 9, 782, number: 1 Publisher: Nature Publishing Group, 2022.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929, 2020.

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., et al.: Sentinel-2: ESA's optical high-resolution mission for GMES operational services, Remote sensing of Environment, 120, 25–36, 2012.

Ehret, T., De Truchis, A., Mazzolini, M., Morel, J.-M., D'aspremont, A., Lauvaux, T., Duren, R., Cusworth, D., and Facciolo, G.: Global tracking and quantification of oil and gas methane emissions from recurrent sentinel-2 imagery, Environmental science & technology, 56, 10 517–10 529, 2022.

Gorroño, J., Varon, D. J., Irakulis-Loitxate, I., and Guanter, L.: Understanding the potential of Sentinel-2 for monitoring methane point emissions, Atmospheric Measurement Techniques, 16, 89–107, 2023.

Groshenry, A., Giron, C., Lauvaux, T., d'Aspremont, A., and Ehret, T.: Detecting Methane Plumes using PRISMA: Deep Learning Model and Data Augmentation, arXiv preprint arXiv:2211.15429, 2022.

Guanter, L., Irakulis-Loitxate, I., Gorroño, J., Sánchez-García, E., Cusworth, D. H., Varon, D. J., Cogliati, S., and Colombo, R.: Mapping methane point emissions with the PRISMA spaceborne imaging spectrometer, Remote Sensing of Environment, 265, 112 671, 2021.

Irakulis-Loitxate, I., Guanter, L., Liu, Y.-N., Varon, D. J., Maasakkers, J. D., Zhang, Y., Chulakadabba, A., Wofsy, S. C., Thorpe, A. K., Duren, R. M., et al.: Satellite-based survey of extreme methane emissions in the Permian basin, Science Advances, 7, eabf4507, 2021.

Irakulis-Loitxate, I., Guanter, L., Maasakkers, J. D., Zavala-Araiza, D., and Aben, I.: Satellites Detect Abatable Super-Emissions in One of the World's Largest Methane Hotspot Regions, Environmental Science & Technology, 56, 2143–2152, 2022.

Jacob, D. J., Varon, D. J., Cusworth, D. H., Dennison, P. E., Frankenberg, C., Gautam, R., Guanter, L., Kelley, J., McKeever, J., Ott, L. E., et al.: Quantifying methane emissions from the global scale down to point sources using satellite observations of atmospheric methane, Atmospheric Chemistry and Physics, 22, 9617–9646, 2022.

Jeppesen, J. H., Jacobsen, R. H., Inceoglu, F., and Toftegaard, T. S.: A cloud detection algorithm for satellite imagery based on deep learning, Remote sensing of environment, 229, 247–259, 2019.

Jongaramrungruang, S., Thorpe, A. K., Matheou, G., and Frankenberg, C.: MethaNet–An AI-driven approach to quantifying methane point-source emission from high-resolution 2-D plume imagery, Remote Sensing of Environment, 269, 112 809, 2022.

Kingma, D. P. and Ba, J.: Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980, 2014.

Lauvaux, T., Giron, C., Mazzolini, M., d'Aspremont, A., Duren, R., Cusworth, D., Shindell, D., and Ciais, P.: Global assessment of oil and gas methane ultra-emitters, Science, 375, 557–561, 2022.

López-Puigdollers, D., Mateo-García, G., and Gómez-Chova, L.: Benchmarking Deep Learning Models for Cloud Detection in Landsat-8 and Sentinel-2 Images, Remote Sensing, 13, 2021.

Maasakkers, J. D., Varon, D. J., Elfarsdóttir, A., McKeever, J., Jervis, D., Mahapatra, G., Pandey, S., Lorente, A., Borsdorff, T., Foorthuis, L. R., et al.: Using satellites to uncover large methane emissions from landfills, Science Advances, 8, eabn9683, 2022.

Ronneberger, O., Fischer, P., and Brox, T.: U-net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, pp. 234–241, Springer, 2015.

Roy, D. P., Wulder, M. A., Loveland, T. R., Woodcock, C. E., Allen, R. G., Anderson, M. C., Helder, D., Irons, J. R., Johnson, D. M., Kennedy, R., et al.: Landsat-8: Science and product vision for terrestrial global change research, Remote sensing of Environment, 145, 154–172, 2014.

Saunois, M., Stavert, A. R., Poulter, B., Bousquet, P., Canadell, J. G., Jackson, R. B., Raymond, P. A., Dlugokencky, E. J., Houweling, S., Patra, P. K., et al.: The global methane budget 2000–2017, Earth system science data, 12, 1561–1623, 2020.

Schuit, B. J., Maasakkers, J. D., Bijl, P., Mahapatra, G., Van den Berg, A.-W., Pandey, S., Lorente, A., Borsdorff, T., Houweling, S., Varon, D. J., et al.: Automated detection and monitoring of methane super-emitters using satellite data, Atmospheric Chemistry and Physics Discussions, pp. 1–47, 2023.

Sherwin, E. D., Rutherford, J. S., Chen, Y., Aminfard, S., Kort, E. A., Jackson, R. B., and Brandt, A. R.: Single-blind validation of space-based point-source detection and quantification of onshore methane emissions, Scientific Reports, 13, 3836, https://doi.org/10.1038/s41598-023-30761-2, number: 1 Publisher: Nature Publishing Group, 2023.

Sinergise Ltd., S. L.: Sentinel Hub, https://www.sentinel-hub.com, 2023.

Stocker, T.: Climate change 2013: the physical science basis: Working Group I contribution to the Fifth assessment report of the Intergovernmental Panel on Climate Change, Cambridge university press, 2014.

Tollefson, J.: Scientists raise alarm over'dangerously fast'growth in atmospheric methane., Nature, 2022.

Varon, D. J., Jervis, D., McKeever, J., Spence, I., Gains, D., and Jacob, D. J.: High-frequency monitoring of anomalous methane point sources with multispectral Sentinel-2 satellite observations, Atmospheric Measurement Techniques, 14, 2771–2785, 2021.

Zavala-Araiza, D., Alvarez, R. A., Lyon, D. R., Allen, D. T., Marchese, A. J., Zimmerle, D. J., and Hamburg, S. P.: Super-emitters in natural gas infrastructure are caused by abnormal process conditions, Nature communications, 8, 1–10, 2017.