



A case study application of machine-learning for the detection of greenhouse gas emission sources

Jacob T. Shaw^a, Grant Allen^{a,*}, David Topping^a, Stuart K. Grange^{b,c}, Patrick Barker^a, Joseph Pitt^{a,1}, Robert S. Ward^d

^a Centre for Atmospheric Science, Department of Earth and Environmental Science, University of Manchester, Manchester, UK

^b Empa, Swiss Federal Laboratories for Materials Science and Technology, Switzerland

^c Wolfson Atmospheric Chemistry Laboratories, University of York, York, UK

^d British Geological Survey, Environmental Science Centre, Nottingham, UK

ARTICLE INFO

Keywords:

Machine-learning
Methane
Greenhouse gas
Emission
Hydraulic fracturing
Climatology

ABSTRACT

Conclusively linking local, episodic enhancements in greenhouse gas concentrations to a specific emission source can be challenging, particularly when faced with multiple proximal sources of emissions and variable meteorology, and in the absence of co-emitted tracer gases. This study demonstrates and evaluates the efficacy of using machine-learning tools to detect episodic emissions of methane (CH₄) from a shale gas extraction facility in Lancashire (United Kingdom). Two machine-learning tools (*rmweather* and *Prophet*) were trained using a two-year climatological baseline dataset collected prior to gas extraction operations at the facility. The baseline dataset consisted of high-precision trace gas concentrations and meteorological data, sampled at 1 Hz continuously between 2016 and 2019. The models showed good overall predictive capacity for baseline CH₄ concentrations, with R² values of 0.85 and 0.76 under optimised training conditions for *rmweather* and *Prophet*, respectively. CH₄ concentrations were then forecast for an 18-month period from the onset of operations at the shale gas facility (in 2018). Forecast values were compared with true measurements to detect anomalous deviations that may indicate the presence of new emission events associated with the operational facility. Both models successfully detected two periods in which CH₄ emissions were known to have occurred (December 2018 and January 2019) via anomalous deviations between modelled and measured concentrations. This work demonstrates the application of machine-learning models for the detection of CH₄ emission events from newly built industrial sources, when used in combination with real-time atmospheric monitoring and a baseline dataset collected prior to installation.

1. Introduction

The introduction of a new atmospheric emission source (such as a new industrial facility) can adversely impact the local atmospheric environment, which may be of concern for local air quality, human health, and greenhouse gas emissions accounting (Manisalidis et al., 2020). However, detecting, interpreting and quantifying the impact of a novel emission source can be difficult, especially in the absence of a comparative historical or long-term dataset prior to its installation. Furthermore, the relationship between local meteorology and pollutant concentration is dynamic. It can therefore be difficult to determine if

short-term changes in pollutant concentrations are due to the direct impact of a new emission source, or due to a shift in local or regional weather patterns and extant sources upwind (Rao and Zurbenko, 1994; Libiseller et al., 2005; Wise and Comrie, 2005; Grange et al., 2018).

An atmospheric baseline climatology provides a set of measurement data that are statistically representative of the typical local environmental background conditions when collected over sufficiently long time periods (typically >12 months, Shaw et al., 2019). Such baselines can facilitate the identification of new (and future) atmospheric emission sources introduced subsequent to the baseline, which may perturb local atmospheric composition on a short-term or long-term basis. An

Peer review under responsibility of Turkish National Committee for Air Pollution Research and Control.

* Corresponding author.

E-mail address: Grant.Allen@manchester.ac.uk (G. Allen).

¹ Now at: School of Chemistry, University of Bristol, Bristol, BS8 1 TS, United Kingdom.

<https://doi.org/10.1016/j.apr.2022.101563>

Received 28 April 2022; Received in revised form 6 September 2022; Accepted 8 September 2022

Available online 10 September 2022

1309-1042/© 2022 Turkish National Committee for Air Pollution Research and Control. Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

atmospheric baseline allows for the incremental impacts of a novel polluting activity to be quantitatively assessed, helping to inform policy decisions concerning environmental protection and compliance with industry regulation, such as environmental permit requirements.

A number of recent studies have employed machine-learning tools to diagnose and interpret patterns and changes in local atmospheric composition. For example, Grange et al. (2018) used a random forest (RF) model to examine patterns in meteorology and particulate matter over 20 years of atmospheric measurements in Switzerland. By normalising the particulate matter data to meteorology (effectively removing the influence of varying meteorological conditions), Grange et al. (2018) were able to diagnose typical conditions that led to elevated and harmful concentrations of local particulate matter. The same weather-normalisation approach was used to detect the influence of specific air quality policy interventions on atmospheric concentrations of SO₂ and NO_x in the UK (Grange and Carslaw, 2019). Their machine-learning procedure highlighted reductions in NO₂ and NO_x that were not otherwise readily observable in the raw concentration data. Similarly, Vu et al. (2019) used de-trended air quality data to assess the impact of clean air policy interventions on Beijing air quality, finding significant reductions in particulate matter, NO₂, SO₂ and CO despite large variability in year-on-year meteorology.

In addition to meteorological normalisation, predictive RF models have been used to forecast a counterfactual dataset (or business-as-usual scenario) based on patterns in historical data, following an intervention or event that caused a change to the prevailing environmental conditions. Such an approach can aid in quantifying the impact(s) of an event, through comparisons against real-world measurements. For example, the air quality impacts of severe Australian wildfires were quantified by using a RF machine-learning algorithm to predict pollutant concentrations in the absence of wildfires (Ryan et al., 2021). Further, the forecasting of counterfactual datasets have found considerable application in diagnosing the air quality impacts caused by the COVID-19 pandemic as a result of lockdown measures introduced throughout 2020 (e.g. Cole et al., 2020; Grange et al., 2021; Petetin et al., 2020; Grange et al., 2021; Lovrić et al., 2021; Shi et al., 2021; Wyche et al., 2021). Some studies have used both meteorological normalisation and counterfactual prediction jointly: Brancher (2021) used meteorological normalisation to interpret reductions in O₃ caused by a lockdown in Vienna, as well as business-as-usual predictions to quantify those changes. Alternative machine-learning approaches also exist: for example, Topping et al. (2020) used an open-source time-series forecasting model (*Prophet*) to interpret a reduction in NO₂ concentrations during the COVID-19 lockdown in the city of Manchester, UK. Together, these studies show the predictive power of machine learning in deconvolving meteorological influences to understand drivers of emission source change. However, previous studies have focussed on detecting and analysing reductions in air quality indicators, and have not assessed the feasibility of using machine learning for identifying pollutant increases as a result of newly introduced emission sources. Identifying new (or changing) emission sources is crucial for establishing impacts on air quality and greenhouse gas emissions. However, increases in air pollution can often be difficult to attribute directly to a specific emission source, especially in the presence of multiple polluting sources. Variable meteorology can further complicate analysis. Machine learning offers a practical tool for deconvolving complex atmospheric datasets, and aiding in the detection and attribution of atmospheric pollution to new emission sources.

In this study, we explore the potential for using machine-learning tools to detect the influence of a new or episodic emission source, using known emission events from hydraulic fracturing in Lancashire (United Kingdom) as a case study to demonstrate capability (Shaw et al., 2020; Shah et al., 2020). Machine learning, when suitably configured and evaluated, offers a versatile and accessible tool for atmospheric data analysis and interpretation, which may not necessarily require specialist expertise in atmospheric science. This could make such tools useful to industry and regulators for leak detection and monitoring purposes.

2. Methods

2.1. Observational data

The observational data used in this work were measured at a fixed-site monitoring station located near to Preston New Road (PNR) in Little Plumpton, Lancashire (United Kingdom). Fig. 1 shows the atmospheric monitoring station location on private land adjacent to an exploratory shale gas facility. The monitoring station has been in operation since February 2016, recording meteorological parameters (wind speed, wind direction, air temperature, and atmospheric pressure) alongside concentrations of atmospheric methane (CH₄), carbon dioxide (CO₂), nitrogen oxides (NO_x; NO and NO₂), ozone (O₃) and particulate matter. The monitoring station is 430 m to the east of the shale gas facility (constructed between 2016 and 2018), which began operational (exploratory) hydraulic fracturing in October 2018. The position of the monitoring station was selected to be downwind in the prevailing wind direction (westerly in the UK) of the facility. Meteorological data were measured using a Lufft WS-500UMB compact weather station with a 2D sonic anemometer for wind speed and wind direction. Greenhouse gas (CH₄ and CO₂) mole fractions (henceforth concentrations) were measured using an Ultra-portable Greenhouse Gas Analyzer (UGGA; Los Gatos Research Inc., USA), calibrated in the field using three gas standards traceable to the World Meteorological Organisation (WMO) greenhouse gas scales. The UGGA air inlet was positioned 3 m above ground. For more information regarding the monitoring station operation, instrumentation, calibration and data quality control procedures, see Shaw et al. (2019) and Purvis et al. (2019). The atmospheric data collected were a component of a wider environmental monitoring programme that also included ground and surface water monitoring, and seismic monitoring (see Ward et al., 2017, 2018, 2019, 2020; Smedley et al., 2022).

A 1-min average atmospheric baseline dataset derived from the measurements at the monitoring facility has previously been reported and interpreted for diurnal, weekly, seasonal, and inter-annual trends and statistics for CH₄ and CO₂ concentrations (see Shaw et al., 2019). Statistical analyses of the baseline data were used to define specific threshold concentration criteria (described in Shaw et al., 2019), exceedances of which represented statistically anomalous events. The thresholds were set as the 99th percentile values for three variables calculated during the baseline period. The three variables were: the hourly-average CH₄ concentration, the CH₄:CO₂ ratio, and the product of wind speed and CH₄ concentration. Thresholds were calculated for each calendar month to account for seasonal variability in the measurements. Hourly-periods in which all three of these 99th percentile thresholds were exceeded were flagged for further examination. The thresholds were exceeded in December 2018 and January 2019, coincident with operator-reported venting of CH₄ from the flare stack at the nearby shale gas facility. Although the Shaw et al. (2019) threshold criteria were successfully used to detect and quantify CH₄ emissions in this instance, the statistical criteria defined in Shaw et al. (2019) may have limited applicability to other locations of interest (e.g. other shale gas sites) and may potentially miss small or short-term anomalous emission events. With suitable training data and optimization, machine-learning tools can provide a dynamic, universal, and automated method to overcome the limitations of more fixed statistical criteria.

Fig. 2 shows 10-min average CH₄ and CO₂ concentrations measured from 1 February 2016 through to 31 December 2019 at the PNR monitoring site (Fig. 1). The CH₄ data exhibited an underlying baseline similar to the northern-hemispheric background (~1900 ppb over this period) and expected seasonality, with intermittent but strong short-term enhancements (occasionally >20 ppm over background). These transient features were the result of local pollution episodes and other emission sources upwind, with some influence from long-range transport from remote urban sources. The largest CH₄ concentrations were

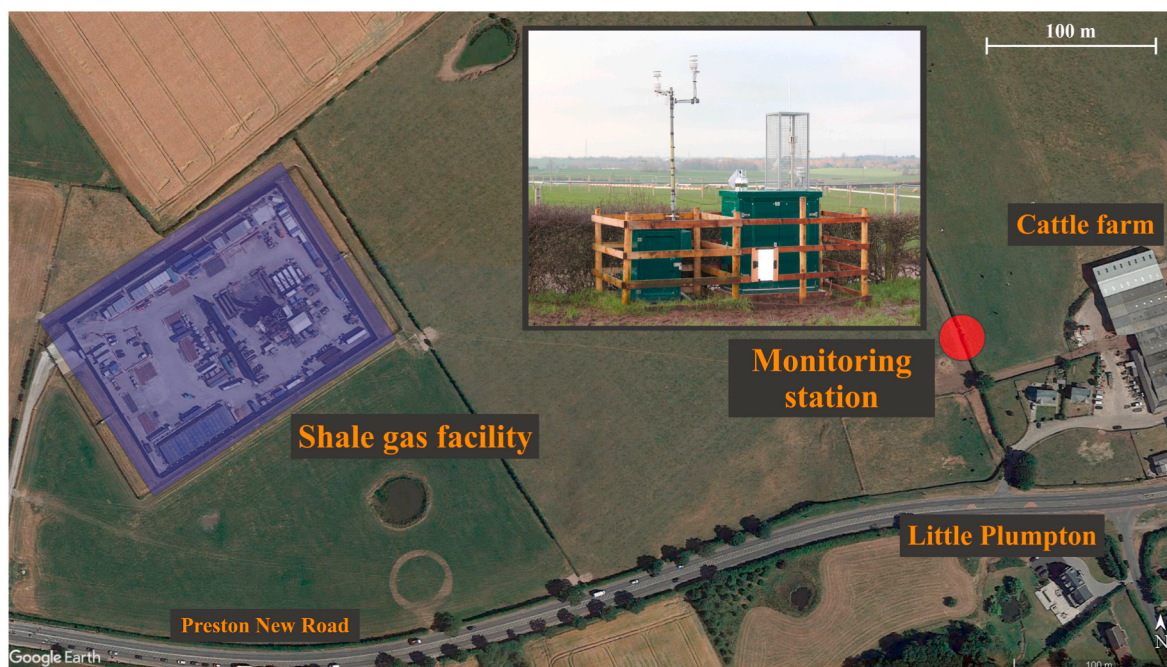


Fig. 1. Location of the fixed-site monitoring station (red circle), cattle farm, and shale gas extraction facility (blue rectangle) in Little Plumpton, Lancashire, UK (Google Maps © 26/11/2018). The monitoring station is approximately 430 m to the east of the shale gas facility boundary, and 100 m north of Preston New Road. The buildings 100 m to the east of the measurement site are part of a dairy and cattle farm. Other potential sources of CH₄ pollution include: leaks from natural gas infrastructure on Preston New Road, a motorway (M55; 1.3 km to the north), and a landfill site (2.6 km to the southwest) (see Lowry et al. (2020) for more details).

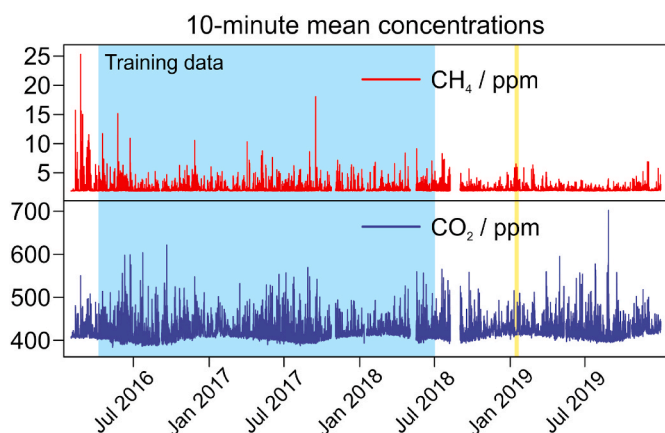


Fig. 2. 10-minute mean CH₄ and CO₂ concentrations between February 2016 and December 2019 measured at the PNR monitoring station. The blue highlighted region indicates the data used to train the machine-learning tools. Data used to test the models for CH₄ prediction were randomly selected (and withheld) from the training data. Data after the training period could have been impacted by emissions from the shale gas facility, and were compared against model predictions. The known emission event from the shale gas site is shown highlighted in yellow (January 2019).

typically associated with low wind speeds ($<2 \text{ m s}^{-1}$) and easterly or northerly winds, during which measurements were influenced by emissions from the nearby cattle farm, a major source of CH₄ (see Shaw et al., 2019 and Lowry et al., 2020). The CO₂ data showed expected seasonal variation in CO₂ background as a result of biospheric respiration, but was again dominated by transient spikes in concentration expected to be associated with local sources.

It is challenging to readily identify the known CH₄ emission event in January 2019 from the CH₄-concentration time series alone (see Fig. 2). In other words, other local CH₄ sources (such as a nearby dairy farm and local waste facilities) often manifested similar (or greater) concentration

enhancements in time series data. These extraneous sources resulted in CH₄ concentrations that obscured those resulting from the emission event, making detection difficult. However, by simultaneously examining complementary data such as wind direction, meteorology, and other trace gases, it is possible to differentiate sources. This concept was successfully realised by Shaw et al. (2019) using the measurement statistics. This work evaluates a more generalised machine-learning approach for the same event detection task. The case study here is limited to CH₄ venting at the PNR shale gas site in 2018 and 2019 as a moratorium on the use of hydraulic fracturing was announced by the UK Government in November 2019, subject to future safety evaluation.

2.2. Machine learning tools

Many different machine-learning models exist, and have been used widely in environmental studies (Zheng et al., 2021). Two machine-learning tools (*rmweather* and *Prophet*) were chosen for this work, and were used to derive statistical relationships between atmospheric CH₄ concentrations and complementary input parameters as a function of time. These two tools were selected as they were both freely and publicly available, with comprehensive and easy-to-follow instructions, and therefore potentially easily accessible to industry and regulatory stakeholders. The *rmweather* model was built specifically to handle atmospheric measurements. Data measured prior to operational activity at the shale gas facility (between 1 April 2016 and 30 June 2018; shown highlighted in blue in Fig. 2) were used to train the machine-learning tools to develop predictive models for forecasting CH₄ based on the prevailing meteorological conditions in the absence of operational shale gas activity. The training dataset consisted of 27 months of data.

The trained models were then used to forecast a counterfactual (or business-as-usual) time-series of CH₄ concentrations, involving periods of known operational site activity, using the auxiliary observational data (meteorological and CO₂ concentration) as input parameters. The counterfactual dataset represents an alternate reality in which the shale gas facility was never operational. From the perspective of the models,

the counterfactual dataset is in the future. The counterfactual dataset consisted of 18 months of data (from 1 July 2018 and up to the end of 2019). This counterfactual dataset was then compared with the measured dataset to test the capability of the machine-learning tools for forecasting the measured CH₄, and to identify perturbations from the baseline due to emissions from known operational activities at the shale gas facility. This allows us to test how a known emission source manifests in the measured data, compared with its absence in the predicted data. The following section describes the two machine learning tools used and their configuration for the tests in this study.

2.2.1. *rmweather*

The *rmweather* package utilises a random forest (or decision forest) machine learning approach to deconvolve the relationships between meteorological variables and air pollutant concentration (see Grange et al. (2018) for more detail on the *rmweather* model). The relationship between a pollutant's concentration and additional, non-meteorological variables (such as another pollutant) can also be analysed.

The *rmweather* tool randomly assigned 20% of observations within the training data as testing data. The other 80% of the training data was used to train the model. The test data was withheld from training procedures, and was used to assess model performance. The *rmweather* tool was applied to the training data (Fig. 2) with the following variables used as predictors for CH₄ concentration: date, day-of-year, weekday, air temperature, wind direction, wind speed, CO₂ concentration, and surface atmospheric pressure. Different time resolutions, or time-averaging periods, were investigated to determine the optimal data frequency for maximising model fit and minimising errors (see Fig. 3). Additional model hyperparameters were also varied iteratively to determine the optimal parameters for maximising model fit and minimising error (see Supplementary Information Section 1).

2.2.2. *Prophet*

Prophet is a time-series forecasting model used to decompose temporal trends in data (see Taylor and Letham (2018) for a description of the model algorithm). Time-series forecasting approaches fit data against yearly, weekly, and daily periodicity. *Prophet* is an open-source time-series forecasting model developed for this purpose for Facebook™. Time-series forecasting models are generally used to analyse non-linear trends to predict future growth whilst allowing for seasonal

effects. The model is capable of deconvolving the impacts of additional parameters, known as regressors, such as meteorological variables. The forecast trend prediction was set to assume zero growth as forecasting was only necessary for ~18 months post training. Such year-on-year changes are negligible compared to the magnitude of transient emission events, which typically result in many ppm enhancements (three orders of magnitude greater than mean annualised global growth). However, a zero annual growth assumption would not be appropriate if applied over longer (e.g. decadal) timescales.

The *Prophet* tool was applied to the training data (Fig. 2) with the following variables used as regressors (predictors) for CH₄ concentration: air temperature, wind direction, wind speed, CO₂ concentration, and atmospheric pressure, in addition to fits against daily, weekly, and yearly periodicity. The effect of model hyperparameters was tested to optimise model performance (see Supplementary Information Section 2).

3. Results and discussion

3.1. Model development and model performance

The two machine-learning models were trained using 27 months of observational data measured at the PNR site. Here we evaluate model performance in terms of predictive accuracy for the test set of data – a set of randomly selected observations withheld from the training data. Testing here was performed on a standard modern desktop computer as a simple qualitative example. The test data provide an independent dataset that is representative of the broader training data, and can be used for model validation (Bennett et al., 2013).

Fig. 3 shows time-series comparisons of measured and model-predicted CH₄ concentrations for the test set of data. It should be noted that the test data is a set of randomly selected observations from within the training period (as described in Section 2.2.1). This conventional approach to evaluating model performance using randomised data subsets (Bennett et al., 2013) may be expected to negatively affect the accuracy of *Prophet* predictions as *Prophet* is fundamentally based on forecasting time-series of continuous, evenly distributed data. Results of an alternative testing strategy, in which the models were tested against a period of continuous data, is presented in Supplementary Information Section 5. For the purposes of presenting a direct comparison between

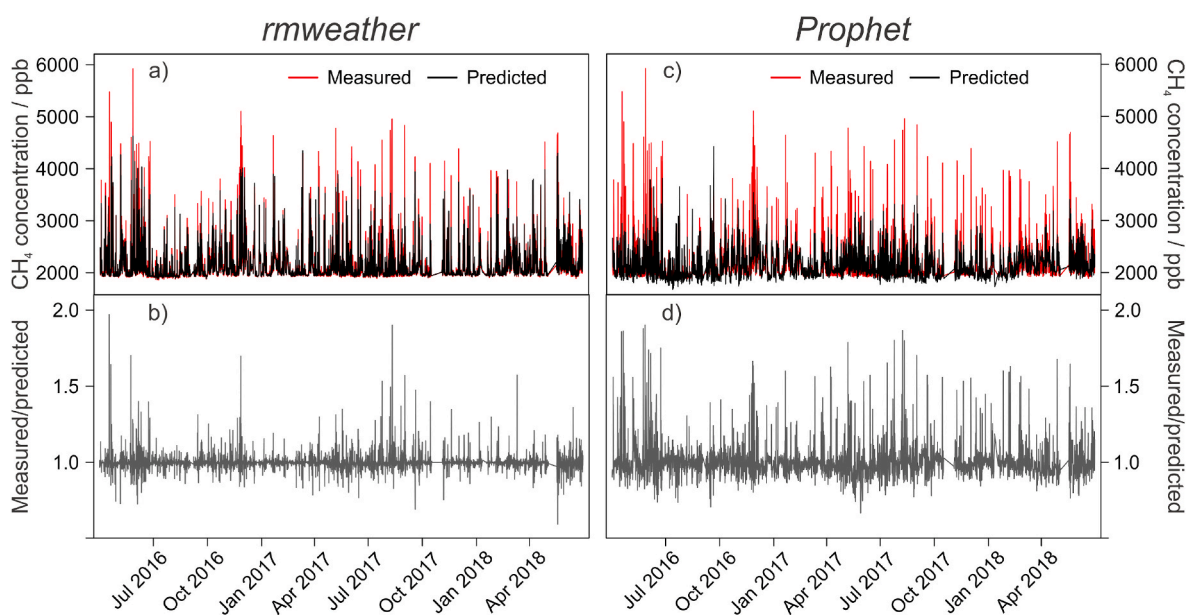


Fig. 3. Time-series of 30-min average a) measured and *rmweather* predicted CH₄ in the test set; b) ratio of measured-to-*rmweather*-predicted CH₄, representing model bias; c) measured and *Prophet* predicted CH₄ in the test set, and; d) ratio of measured-to-*Prophet*-predicted CH₄, representing model bias.

the models, the *Prophet* model output shown here was trained using a time-averaging period of 30 min, despite this not being the optimal period for *Prophet* model performance (see [Supplementary Information Section 3](#)).

The *rmweather* model performed well when predicting the test set observations, and generally captured patterns in CH₄ throughout the training period. Mean absolute error was 55.4 ppb (median = 18.9 ppb), and mean absolute percentage error was 2.2% ([Table S1](#)). The overall mean bias was -0.24 ppb, indicating a marginal under-prediction relative to measurements. The model generally performed better at lower CH₄ concentrations (i.e. those typical of background conditions) with absolute error increasing with increasing CH₄ concentration ([Fig. S4](#)). This is likely due to the scarcity of high CH₄ concentration events, and the relatively rare conditions associated with them (easterly winds, low wind speeds, low atmospheric boundary layer etc.). The measured-to-predicted ratio shown in [Fig. 3b](#) provides a continuous indication of model prediction relative to the measurements. Ratio values greater than 1.0 indicate model under-prediction relative to the measurements, and values less than 1.0 show model over-prediction relative to measurement. However, there were data for which the model underestimated CH₄ by factors of up to 2.0. These excursions occurred for the highest measured CH₄ concentrations (>4000 ppb), suggesting that the model was unable to fully capture transient events based on meteorological conditions alone. This was likely due to unusual emission events from local sources impacting local CH₄ concentrations abnormally.

Prophet also did performed reasonably well predicting CH₄ in the test set ([Fig. 3](#)), although metrics were slightly worse than those for *rmweather* ([Table S1](#)). Mean absolute error was 136 ppb (median = 80.5 ppb), and mean absolute percentage error was 5.7%. As for *rmweather*, *Prophet* showed greater capacity for predicting lower CH₄ concentrations, typical of background conditions ([Fig. S4](#)). As was the case for *rmweather*, there was no clear temporal bias in *Prophet* predictions, with the model doing a reasonable job throughout the training data and across different seasons. *Prophet* also underestimated many of the highest CH₄ concentrations, and to a greater extent than *rmweather*.

[Fig. 4](#) shows smoothed histogram density plots of the distribution of CH₄ concentrations within the test set of data, both as measurements, and as values predicted by *rmweather* and by *Prophet*. There was a clear overlap between the density of CH₄ concentrations predicted by *rmweather* and the measurements, with the distributions almost identical across background CH₄ values (those around 2000 ppb), and the long tail distribution of CH₄ enhancements (those >2000 ppb). *Prophet* did not capture the same distribution. Although the mode of the *Prophet* CH₄

distribution was accurate, the model appeared to smear the density distribution about the mode relative to that observed for the measurements. *Prophet* also modelled lower CH₄ than those measured, even predicting concentrations that would be highly unusual in the troposphere given the global CH₄ background value.

[Fig. 5](#) shows the relationships between measured and predicted CH₄ concentration, and the measured-to-predicted ratios plotted as a function of wind direction. The scatter plot confirms the underestimation by the models of the largest measured CH₄ concentrations, as observed in [Fig. 3](#). The slopes of the linear regression were 0.81 and 0.53 for *rmweather* and *Prophet* respectively. These values indicate underestimations of CH₄ concentration enhancements, with much greater underestimation by the *Prophet* model. Model prediction for background CH₄ concentrations (~2000 ppb) was much better. R² values of 0.85 for *rmweather* and 0.53 for *Prophet* indicate reasonable model fit ([Table S1](#)).

A wind rose plot for *rmweather* demonstrates very little model bias with respect to wind direction - most of the measured-to-predicted ratio values were between 0.95 and 1.05 with the majority of values outside of this range occurring under easterly wind conditions. This is expected to be due to the nearby dairy farm, a major source of CH₄ that may not always operate in a consistent and predictable manner. On the other hand, *rmweather* performed exceptionally well for westerly wind directions, with a large proportion of measured-to-predicted ratios between 0.99 and 1.01, as well as very few ratios greater than 1.05 or less than 0.95. This is expected to be due to the relatively clean (in terms of CH₄ pollution) and well-mixed air sampled from the west ([Shaw et al., 2019](#)) and is encouraging for accurate model forecasting beyond the training period. For *Prophet*, many of the underestimates of CH₄ occurred during easterly winds, with an almost equal amount of underestimation and overestimation of CH₄ occurring during westerly winds.

[Fig. 6](#) shows a hierarchy of the importance, or correlation, of variables towards CH₄ prediction, as output by the *rmweather* package. The importance values output by *rmweather* are defined as the permutation importance differences of prediction error, and are unit less. These statistics represent the increase in prediction ability after the inclusion of each explanatory variable, and therefore give an indication of the importance of each variable for representing the dependent variable (CH₄, in this case) in a random forest model. The lower plots show the partial dependencies of the individual variables towards CH₄ over their respective ranges, where each of the other variables were fixed to their mean values. CO₂ and wind direction data showed the greatest correlation with CH₄, with the influence of the nearby cattle and dairy farm clearly visible in the high CH₄ concentrations diagnosed when wind directions were roughly 90° (easterly), as expected. Atmospheric pressure, wind speed and weekday had the least influence on CH₄, with only a small impact across most of their ranges of values. This is expected as sources of CH₄ are not typically limited to weekday human activity (such as traffic volume, which may be expected to be more associated with CO₂).

The overall good predictive capacity demonstrated here for both *rmweather* and *Prophet* indicates that machine-learning models are able to use measured data to estimate CH₄ concentrations. CH₄ concentration prediction was typically better for background values, which may be expected due to the high density of those values in the dataset. CH₄ concentration enhancements were typically underestimated (by 20% for *rmweather*, and 50% for *Prophet*) but were still reasonable considering the transient and relative rarity of such events. Importantly, these results are remarkable considering the models have no knowledge of local emission sources, in terms of their location, their relative strength, or their temporal patterns in emissions. Model improvements could potentially be made upon including ancillary data such as emission source information (which would require flux analysis), or even by including other atmospheric data such as NO_x or O₃.

The following section will analyse the performance of the machine-learning models for forecasting CH₄ beyond the training data (into the

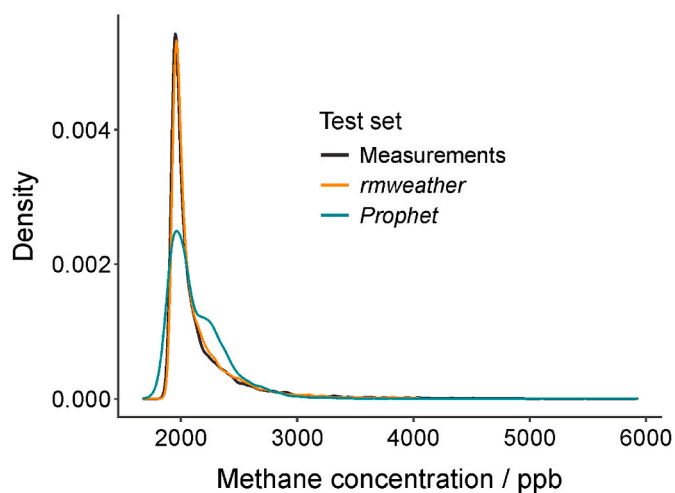


Fig. 4. Density plot of the distribution of CH₄ concentrations within the test set data, as measurements, and as *rmweather* and *Prophet* predictions.

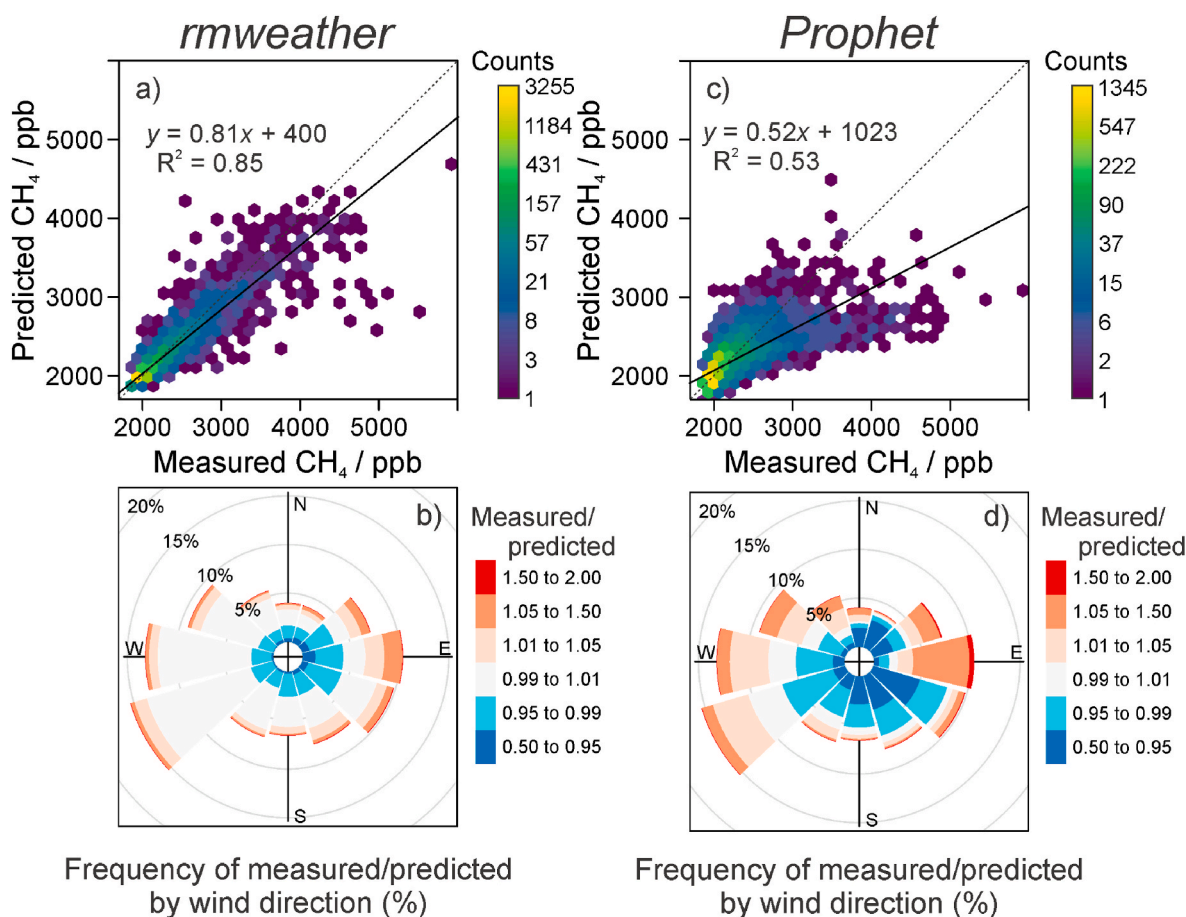


Fig. 5. a) Scatter plot of measured and *rmweather*-predicted CH₄ for the test set. The solid line shows the linear regression through the data, and the dotted line indicates a 1:1 fit; b) Wind rose plot of measured-to-predicted ratios for *rmweather*. The radii of the wedges indicate the percentage of values occurring in each wind direction; c) scatter plot of measured and *Prophet*-predicted CH₄ for the test set, and; d) wind rose plot of measured-to-predicted ratios for *Prophet*.

future, from the model's perspective). In this case, the forecasts represent a "business-as-usual" scenario, in which the conditions and CH₄ emission sources present in the training data persist, and therefore the forecast is counterfactual to the measured data.

3.2. Forecasting CH₄ and emission detection

The trained and optimised machine-learning models were used to forecast CH₄ beyond the training period (after 1 July 2018), corresponding to a period of time in which the shale gas facility became operational and exploratory hydraulic fracturing began. As the models were trained in the absence of these activities, any deviation between model-forecast CH₄ and actual measured CH₄ may be indicative of emissions from this new CH₄ source.

Fig. 7 shows a comparison of measured CH₄ and counterfactual CH₄ concentrations forecast by both *rmweather* and *Prophet*. For the most part, both models continued to correctly predict CH₄ data beyond the training period, including many of the short-term CH₄ enhancements, despite some difficulties in simulating the precise magnitudes. The models' performance decreased when forecasting CH₄ concentrations beyond the training period, with R^2 values of 0.60 and 0.46 for *rmweather* and *Prophet* respectively (Fig. S7). However, it should be noted that decreased performance would be expected as this period includes CH₄ emissions for which the models were not trained to predict. The central panels of Fig. 8 show the measured-to-forecast ratio as a function of time, to give an indication of model deviation from the measurements. *rmweather* appeared to overestimate background CH₄, forecasting background CH₄ concentrations approximately 20 ppb

greater than those measured. Overestimation was also particularly clear in *Prophet* forecasts for summer 2019. *rmweather* overestimations could be a result of the irregular annual increase in global CH₄ background concentration (Dlugokencky, 2022) but it is unknown as to what caused overestimations in *Prophet* forecasts. Therefore, it is recommended that machine-learning models be trained on at least 12-months of data in order to capture any seasonal cycle, with an additional caveat that even training on several years of data may not improve long-term forecasts if inter-annual growth is highly variable.

Filtering the measured-to-forecast ratio values for westerly wind directions ($270^\circ \pm 45^\circ$) allows for analysis of only those periods that may have been influenced by CH₄ emissions from the shale gas facility (the potential new source of CH₄ emission) introduced in mid-2018. The mean measured-to-forecast ratio under westerly wind conditions was $0.97 (\pm 0.06 \text{ at } 1\sigma)$ and $0.99 (\pm 0.06 \text{ at } 1\sigma)$ for *rmweather* and for *Prophet* respectively, indicating improved predictive capacity for CH₄ during periods of westerly winds (relative to all data). This is likely a result of the dominance of well-mixed but clean Atlantic airflow (with little variability in CH₄) from the west (Shaw et al., 2019), and correlates with the performance observed during model testing above.

Whilst the counterfactual CH₄ forecast by both machine-learning tools showed reasonable agreement with measurements, a number of instances where the models substantially underestimated the measured CH₄ occurred within the 18-month period, indicated by clear excursions in the measured-to-forecast ratio to values greater than 1.5 (Fig. 7 westerly winds plots). Two distinct periods of excursions were immediately obvious, occurring in December 2018 and January 2019, consistent with the two periods of known CH₄ emission. This clearly

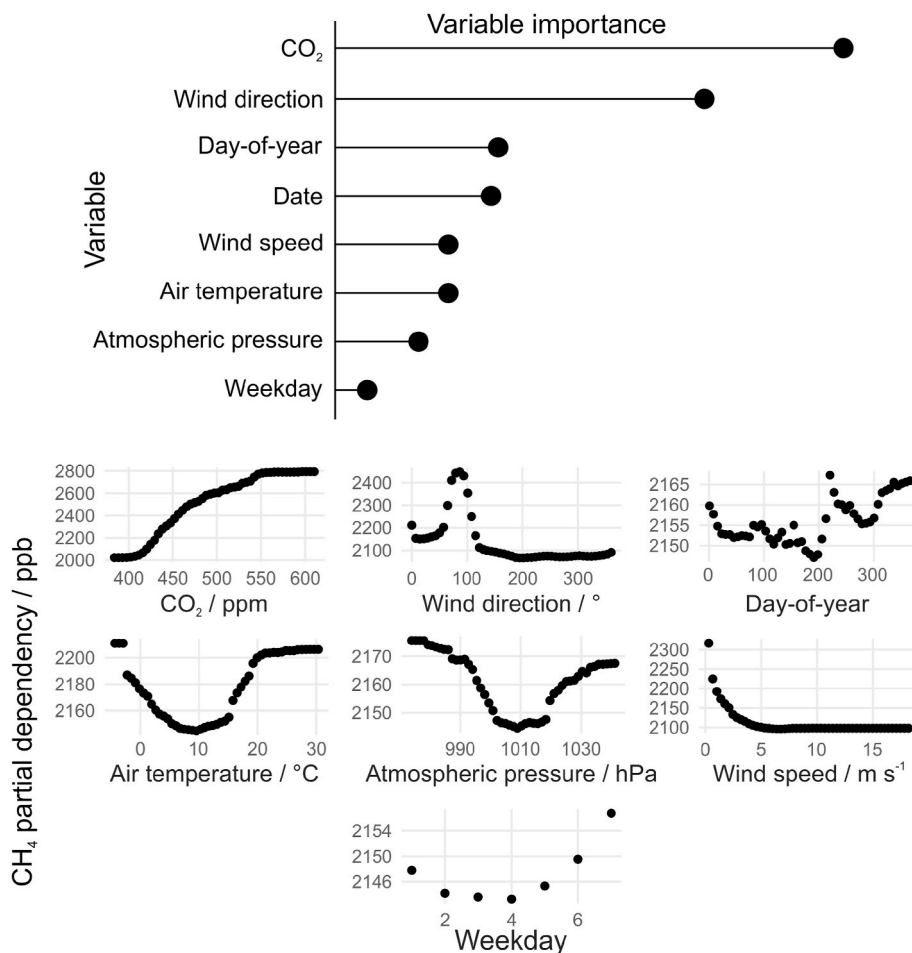


Fig. 6. Top panel; hierarchy of variable importance towards predicting, or correlation with, CH₄ at PNR, determined by the *rmweather* package. Bottom panels; partial dependencies of each of the variables, showing the determined variability in CH₄ concentration across each variable's range of values (assuming all other variables are fixed at their mean value).

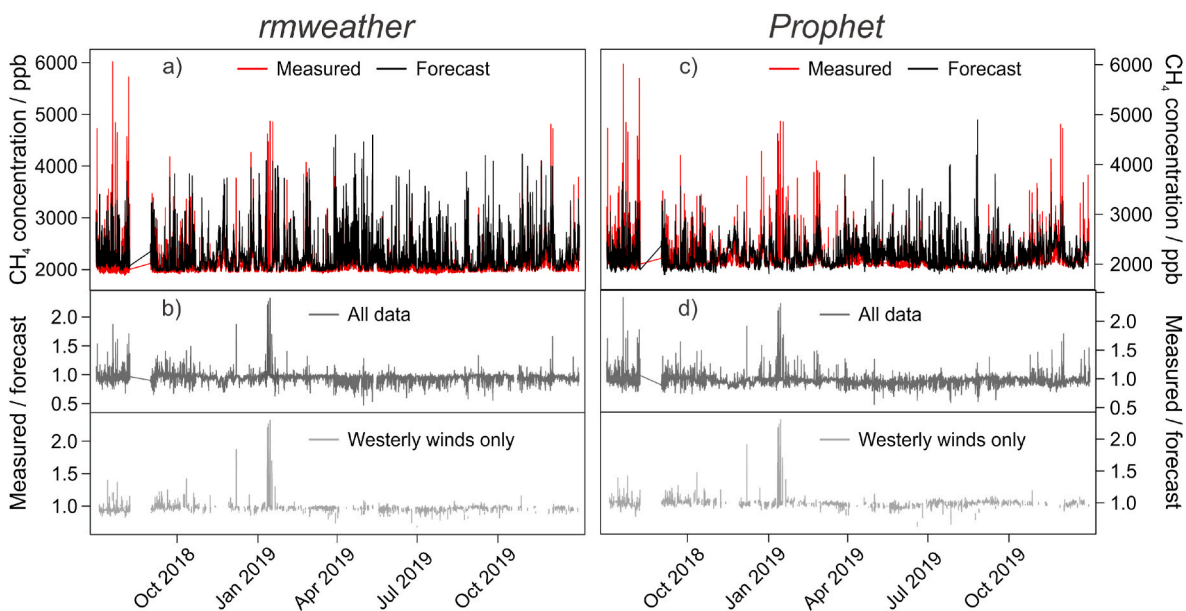


Fig. 7. Time-series of 30-min average a) measured and *rmweather* forecast CH₄ after development of the shale gas facility; b) ratio of measured-to-*rmweather*-forecast CH₄; c) measured and *Prophet* forecast CH₄ after development of the shale gas facility, and; d) ratio of measured-to-*Prophet*-forecast CH₄.

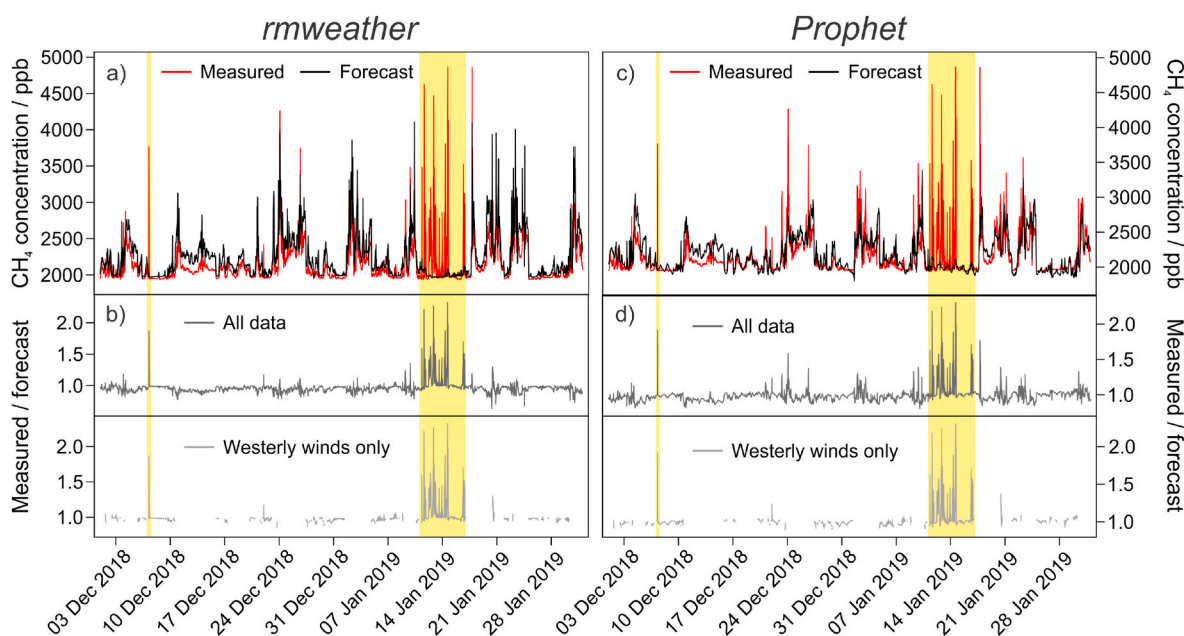


Fig. 8. Time-series of 30-min average a) measured and *rmweather* forecast CH₄ in December 2018 and January 2019; b) ratio of measured-to-*rmweather*-forecast CH₄; c) measured and *Prophet* forecast CH₄ in December 2018 and January 2019, and; d) ratio of measured-to-*Prophet*-forecast CH₄. Highlighted areas indicate periods in which CH₄ emissions were known to have occurred (as reported by the operators).

demonstrates that the machine learning models are capable of identifying anomalous emission events, the primary goal of this study. These two periods emissions are shown in detail in Fig. 8.

Fig. 8 shows a comparison of measured and forecast (*rmweather* and *Prophet*) CH₄ in December 2018 and January 2019. Local residents reported emissions of ‘vapours’ from the shale gas facility on the morning of December 7th 2018 via a local social media group. The operator reported performing flowback operations on a shale gas well between 10th and 16th January 2019 (see Shaw et al., 2020). Enhancements in CH₄ were measured during both of these periods, coincident with westerly winds at the time. Neither of the models were able to accurately forecast the CH₄ during these events, predicting CH₄ concentrations of roughly 2000 ppb (consistent with the global CH₄ background) rather than enhancements of ~5000 ppb (averaged over 30-min). This discrepancy strongly indicates the prevalence of a new emission source that was not present during the training period. The models underestimated the CH₄ by up to a factor of 2.5 during the flowback period (Jan 10th – 16th), as indicated by the measured-to-forecast ratios.

The same two periods of CH₄ enhancements were also identified using the manually derived threshold criteria described by Shaw et al. (2019) and Shaw et al. (2020). However, such threshold criteria require specialist proficiency and expertise in the interpretation of meteorology and atmospheric chemistry. This study demonstrates the use of machine-learning tools for the same purpose, and for reaching the same conclusion, but with clear advantages in usability, repeatability, and versatility for other measurement locations. Machine-learning tools and the prediction of counterfactual observations could be used to improve automated “alarm” systems, to alert operators to the presence of unexpected and dangerous pollutant emissions. Such systems could allow for rapid responses to reduce leaks and equipment failure.

4. Conclusions

This work demonstrates the use of two machine-learning approaches for the identification of transient CH₄ emissions associated with horizontal hydraulic fracturing for shale gas extraction in Lancashire (United Kingdom). In principle, the methods demonstrated here should be applicable for the detection of any atmospheric pollutant, in conjunction

with real-time monitoring and a representative baseline dataset collected prior to the introduction of any new emission source. The approach is beneficial for interpreting the incremental change of new emission sources and their local environmental impact, which can address concerns surrounding human health, air quality, and greenhouse gas emissions. Machine-learning approaches can offer a versatile and user-friendly tool for identifying and interpreting new emission sources, via the rapid processing of large datasets to deconvolve parameter relationships.

In this study, two independent machine-learning approaches were tested: *rmweather*, a random forest model, and *Prophet*, a time-forecasting model. These machine-learning models were trained on 27 months of baseline data measured prior to the development of the shale gas facility. The tools were then used to forecast counterfactual CH₄ for an 18-month period after shale gas extraction operations began. Forecast and prediction procedures were rapid, with both models taking less than 30 min to process the four years of measurement data on a standard desktop computer.

Both models showed capability for predicting the measured CH₄, with optimised R² values of 0.85 and 0.76 for *rmweather* and for *Prophet* respectively. Both models were proficient at identifying transient enhancements in CH₄ as a result of pre-existing emission sources. Both models demonstrated efficacy for detecting new sources of CH₄ emission through the significant deviation of the forecast counterfactual CH₄ from the measured CH₄ during periods of known emission from the shale gas facility. We recommend the use of *rmweather* for emission detection applications as it performed better for atmospheric data analysis, and random forest tools do not rely entirely on decomposing temporal data trends that may exhibit high variability over short time scales. However, we acknowledge that other machine-learning algorithms also exist which have not been tested here for their applicability for detecting episodic atmospheric emissions. Future work should examine the efficacy of a wider variety of machine-learning models for this application. Future work could also examine the potential of using extraneous data, measured beyond a single-site monitoring station, to further constrain parameter relationships. For example, measurements of global background CH₄ could potentially be used to improve characterisation of the local CH₄ background. To conclude, machine-learning tools offer a

rapid, reliable, and low-effort approach for detecting and identifying the impacts of new emission sources, and may find potential use as an automated alert system for unexpected and dangerous pollutant emissions.

Author contributions

JTS – conceptualization, investigation, data curation, formal analysis, methodology, writing. GA – principal investigator (PI), conceptualization, funding acquisition, investigation, project administration, supervision, writing. DT – methodology. SKG – methodology, software, writing. PB – investigation. JRP – data curation, investigation. RSW – principal investigator (PI), funding acquisition, project administration.

Declaration of Competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by two complementary projects both led by the British Geological Survey (BGS): an environmental monitoring project (www.bgs.ac.uk/lancashire) jointly funded by the BGS National Capability Programme and a grant awarded by the Department for Business, Energy and Industrial Strategy (BEIS; Grant code: GA/18F/017/NEE6617R); and the EQUIPT4RISK project (NE/R01809X/1), which is part of the NERC/ESRC Unconventional Hydrocarbon Research Programme (<http://www.ukuh.org/>).

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.apr.2022.101563>.

References

- Bennett, N.D., Croke, B.F.W., Guariso, G., Guillaume, J.H.A., Hamilton, S.H., Jakeman, A.J., Marsili-Libelli, S., Newham, L.T.H., Norton, J.P., Perrin, C., Pierce, S. A., Robson, B., Seppelt, R., Voinov, A.A., Fath, B.D., Andreassian, V., 2013. Characterising performance of environmental models. *Environ. Model. Software* 40, 1–20. <https://doi.org/10.1016/j.envsoft.2012.09.011>.
- Brancher, M., 2021. Increased ozone pollution alongside reduced nitrogen dioxide concentrations during Vienna's first COVID-19 lockdown: significance for air quality management. *Environ. Pollut.* 284, 117153 <https://doi.org/10.1016/j.envpol.2021.117153>.
- Cole, M.A., Elliott, R.J.R., Li, B., 2020. The impact of the WUHAN COVID-19 lockdown on air pollution and health: a machine learning and augmented synthetic control approach. *Environ. Resour. Econ.* 76, 553–580. <https://doi.org/10.1007/s10640-020-00483-4>.
- Dlugokencky, E., 2022. NOAA/ESRL. Retrieved from. http://www.esrl.noaa.gov/gmd/ccgg/trends_ch4. Last access: June.
- Grange, S.K., Carslaw, D.C., Lewis, A.C., Boleti, E., Hueglin, C., 2018. Random forest meteorological normalisation models for Swiss PM₁₀ trend analysis. *Atmos. Chem. Phys.* 18, 6223–6239. <https://doi.org/10.5194/acp-18-6223-2018>.
- Grange, S.K., Carslaw, D.C., 2019. Using meteorological normalisation to detect interventions in air quality time series. *Sci. Total Environ.* 653, 578–588. <https://doi.org/10.1016/j.scitotenv.2018.10.344>.
- Grange, S.K., Lee, J.D., Drysdale, W.S., Lewis, A.C., Hueglin, C., Emmenegger, L., Carslaw, D.C., 2021. COVID-19 lockdowns highlight a risk of increasing ozone pollution in European urban areas. *Atmos. Chem. Phys.* 21, 4169–4185. <https://doi.org/10.5194/acp-21-4169-2021>.
- Libiseller, C., Grimvall, A., Waldén, J., Saari, H., 2005. Meteorological normalisation and non-parametric smoothing for quality assessment and trend analysis of tropospheric ozone data. *Environ. Monit. Assess.* 100, 33–52. <https://doi.org/10.1007/s10661-005-7059-2>.
- Lovrić, M., Pavlović, K., Vuković, M., Grange, S.K., Haberl, M., Kern, R., 2021. Understanding the true effects of the COVID-19 lockdown on air pollution by means of machine learning. *Environ. Pollut.* 274 <https://doi.org/10.1016/j.envpol.2020.115900>, 115900.
- Lowry, D., Fisher, R.E., France, J.L., Coleman, M., Lanoisellé, M., Zazzeri, G., Nisbet, E. G., Shaw, J.T., Allen, G., Pitt, J., Ward, R.S., 2020. Environmental baseline monitoring for shale gas development in the UK: identification and geochemical characterisation of local source emissions of methane to atmosphere. *Sci. Total Environ.*, 134600 <https://doi.org/10.1016/j.scitotenv.2019.134600>, 708.
- Manisalidis, I., Stavropoulou, E., Stavropoulos, A., Bezirtzoglou, E., 2020. Environmental and health impacts of air pollution: a review. *Front. Public Health* 8, 14. <https://doi.org/10.3389/fpubh.2020.00014>.
- Petetin, H., Bowdalo, D., Soret, A., Guevara, M., Jorba, O., Serradell, K., García-Pando, C. P., 2020. Meteorology-normalized impact of the COVID-19 lockdown upon NO₂ pollution in Spain. *Atmos. Chem. Phys.* 20, 11119–11141. <https://doi.org/10.5194/acp-20-11119-2020>.
- Purvis, R.M., Lewis, A.C., Hopkins, J.R., Wilde, S.E., Dunmore, R.E., Allen, G., Pitt, J., Ward, R.S., 2019. Effects of 'pre-fracking' operations on ambient air quality at a shale gas exploration site in rural North Yorkshire, England. *Sci. Total Environ.* 673, 445–454. <https://doi.org/10.1016/j.scitotenv.2019.04.077>.
- Rao, S.T., Zurbenko, I.G., 1994. Detecting and tracking changes in ozone air quality. *J. Air Waste Manage.* 44, 1089–1092. <https://doi.org/10.1080/10473289.1994.10467303>.
- Ryan, R.G., Silver, J.D., Schofield, R., 2021. Air quality and health impact of 2019-20 Black Summer megafires and COVID-19 lockdown in Melbourne and Sydney, Australia. *Environ. Pollut.* 274, 116498 <https://doi.org/10.1016/j.envpol.2021.116498>.
- Shah, A., Ricketts, H., Pitt, J.R., Shaw, J.T., Kabbabe, K., Leen, J.B., Allen, G., 2020. Unmanned aerial vehicle observations of cold venting from exploratory hydraulic fracturing in the United Kingdom. *Env. Res. Comm.* 2, 021003 <https://doi.org/10.1088/2515-7620/ab716d>.
- Shaw, J.T., Allen, G., Pitt, J., Mead, M.I., Purvis, R.M., Dunmore, R., Wilde, S., Shah, A., Barker, P., Bateson, P., Bacak, A., Lewis, A.C., Lowry, D., Fisher, R., Lanoisellé, M., Ward, R.S., 2019. A baseline of atmospheric greenhouse gases for prospective UK shale gas sites. *Sci. Total Environ.* 684, 1–13. <https://doi.org/10.1016/j.scitotenv.2019.05.266>.
- Shaw, J.T., Allen, G., Pitt, J., Shah, A., Wilde, S., Stamford, L., Fan, Z., Ricketts, H., Williams, P.I., Bateson, P., Barker, P., Purvis, R., Lowry, D., Fisher, R., France, J., Coleman, M., Lewis, A.C., Risk, D.A., Ward, R.S., 2020. Methane flux from flowback operations at a shale gas site. *JAPCA J. Air Waste Ma* 70. <https://doi.org/10.1080/10962247.2020.1811800>.
- Shi, Z., Song, C., Liu, B., Lu, G., Xu, J., Vu, T.V., Elliott, R.J.R., Li, W., Bloss, W.J., Harrison, R.M., 2021. Abrupt but smaller than expected changes in surface air quality attributable to COVID-19 lockdowns. *Sci. Adv.* 7, 3. <https://doi.org/10.1126/sciadv.abd6696>.
- Smedley, P.L., Shaw, J.T., Allen, G., Crewdson, E., 2022. *Environmental Monitoring in the Fyde, Lancashire Phase 6 Final Report*. British Geological Survey, Nottingham, UK.
- Taylor, S.J., Letham, B., 2018. Forecasting at scale. *Am. Statistician* 72, 37–45. <https://doi.org/10.1080/00031305.2017.1380080>.
- Topping, D., Watts, D., Coe, H., Evans, J., Bannan, T.J., Lowe, D., Jay, C., Taylor, J.W., 2020. Evaluating the use of Facebook's Prophet model v0.6 in forecasting concentration of NO₂ at single sites across the UK and in response to the COVID-19 lockdown in Manchester, England. *Geosci. Model Dev.* <https://doi.org/10.5194/gmd-2020-270> submitted for publication.
- Vu, T.V., Shi, Z., Cheng, J., Zhang, Q., He, K., Wang, S., Harrison, R.M., 2019. Assessing the impact of clean air action on air quality trends in Beijing using a machine learning technique. *Atmos. Chem. Phys.* 19, 11303–11314. <https://doi.org/10.5194/acp-19-11303-2019>.
- Ward, R.S., Smedley, P.L., Allen, G., Baptie, B.J., Daraktchieva, Z., Horleston, A., Jones, D.G., Jordan, C.J., Lewis, A., Lowry, D., Purvis, R.M., Rivett, M.O., 2017. *Environmental Baseline Monitoring Project: Phase II Final Report*. British Geological Survey, Nottingham, UK.
- Ward, R.S., Smedley, P.L., Allen, G., Baptie, B.J., Cave, M.R., Daraktchieva, Z., Fisher, R., Hawthorn, D., Jones, D.G., Lewis, A., Lowry, D., Luckett, R., Marchant, B.P., Purvis, R.M., Wilde, S., 2018. *Environmental Baseline Monitoring: Phase III Final Report*. British Geological Survey, Nottingham, UK.
- Ward, R.S., Smedley, P.L., Allen, G., Baptie, B.J., Barkwith, A.K.A.P., Bateson, L., Bell, R. A., Bowes, M., Coleman, M., Cremen, G., Daraktchieva, Z., Gong, M., Howarth, C.H., Fisher, R., Hawthorn, D., Jones, D.G., Jordan, C., Lanoisellé, M., Lewis, A.C., Lister, T.R., Lowry, D., Luckett, R., Mallin-Martin, D., Marchant, B.P., Miller, C.A., Milne, C.J., Novellino, A., Pitt, J., Purvis, R.M., Rivett, M.O., Shaw, J., Taylor-Curran, H., Wasikiewicz, J.M., Werner, M., Wilde, S., 2019. *Environmental Monitoring: Phase 4 Final Report*. British Geological Survey, Nottingham, UK.
- Ward, R.S., Smedley, P.L., Allen, G., Baptie, B.J., Barker, P., Barwith, A.K.A.P., Bates, P., Bateson, L., Bell, R.A., Coleman, M., Cremen, G., Crewdson, E., Daraktchieva, Z., Gong, M., Howarth, C.H., France, J., Lewis, A.C., Lister, T.R., Lowry, D., Luckett, R., Mallin Martin, D., Marchant, B.P., Miller, C.A., Milne, C.J., Novellino, A., Pitt, J., Purvis, R.M., Rivett, M.O., Shaw, J., Taylor-Curran, H., Wasikiewicz, J.M., Werner, M., Wilde, S., 2020. *Environmental Monitoring: Phase 5 Final Report*. British Geological Survey, Nottingham, UK.
- Wise, E.K., Comrie, A.C., 2005. Extending the Kolmogorov-Zurbenko filter: application to ozone, particular matter, and meteorological trends. *J. Air Waste Manage.* 55, 1208–1216. <https://doi.org/10.1080/10473289.2005.10464718>.
- Wyche, K.P., Nichols, M., Parfitt, H., Beckett, P., Gregg, D.J., Smallbone, K.L., Monks, P. S., 2021. Changes in ambient air quality and atmospheric composition and reactivity in the South East of the UK as a result of the COVID-19 lockdown. *Sci. Total Environ.* 755, 142526 <https://doi.org/10.1016/j.scitotenv.2020.142526>.
- Zheng, L., Lin, R., Wang, X., Chen, W., 2021. The development and application of machine learning in atmospheric environment studies. *Rem. Sens.* 13, 4839. <https://doi.org/10.3390/rs13234839>.