Improvements of AI-driven emission estimation for point sources applied to high resolution 2-D methane-plume imagery

Thomas Plewa[a,b], André Butz[a,c,d], Christian Frankenberg[e,f], Andrew K. Thorpe[f], Julia Marshall[b]

[a]*Institute of Environmental Physics (IUP), Heidelberg University, Heidelberg, Germany*
[b]*Deutsches Zentrum für Luft- und Raumfahrt, Institut für Physik der Atmosphäre, Oberpfaffenhofen, Germany*
[c]*Heidelberg Center for the Environment (HCE), Heidelberg University, Heidelberg, Germany*
[d]*Interdisciplinary Center for Scientific Computing (IWR), Heidelberg University, Heidelberg, Germany*
[e]*Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena, CA 91125, USA*
[f]*NASA Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91125, USA*

Abstract

Anthropogenic methane (CH_4) sources have had a considerable impact on the Earth's changing radiation budget since pre-industrial times. Localized sources such as those resulting from the fossil fuel industry and waste treatment have been shown to make up a substantial fraction of the emission total, and CH_4 plumes from such sources are detectable through airborne and space-based hyperspectral imaging techniques. Here, we further develop a machine learning technique to estimate CH_4 emission rates from such plume images without the need for auxiliary data such as local wind speed information. We directly build upon the idea of previous research which used a convolutional neural network (CNN) called MethaNet and a library of large-eddy-simulations (LES) of turbulent CH_4 plumes as our synthetic data environment. Here we suggest appropriate error metrics and changes to the training procedure that reduce systematic biases present in previous studies. Our improved setup has a mean absolute percentage error (MAPE) of 10% for sources with flux rates above 40 $kgCH_4/h$, a Pearson correlation coefficient of 98% and is capable of providing meaningful error estimates for its predictions. This is a significant improvement to MethaNet and other studies and can be used as an efficient method for point source quantification in the future.

Keywords: Methane gas, Methane quantification, Deep learning, Point-source estimation, AVIRIS-NG, Greenhouse gas, CNN, LES

1. Introduction

Methane (CH_4) is the second-most important anthropogenically-influenced greenhouse gas next to carbon dioxide (CO_2) (**?**), and increased atmospheric concentrations are responsible for around 30% of the global warming since pre-industrial times (**??**). Because of its short atmospheric lifetime compared to CO_2, CH_4 is the focus of near-term global emission reduction measures (**?**). A large fraction of the anthropogenic CH_4 is emitted by industrial point sources (**??**).

While global mapping satellites such as the TROPOspheric Monitoring Instrument (TROPOMI) (**?**) are capable of identifying regions with high emissions and provide a top-down constraint on CH_4, they are not capable of resolving smaller underlying point sources that drive the increased enhancement. Inventory-based approaches on the other hand, usually underestimate those emissions (**??**), which suggests that a large part of those emissions are of unknown origin and could be caused by e.g. equipment malfunctions (**??**). Therefore, the identification and quantification of point sources is crucial to mitigate emissions and provides a cost effective, fast-activating way of reducing climate impacts (**??**).

This requires high spatial resolutions, which can be provided by spaceborne or airborne imaging absorption spectroscopy. Currently, multispectral area mappers with high spatial resolution such as Sentinel-2 are only capable of measuring super-emitters emitting a few tons of CH_4 per hour (**?**) and hyperspectral satellites such as the PRecursore IperSpettrale della Missione Applicativa (PRISMA) (**?**) or the Environmental Mapping and Analysis Program (EnMAP) (**?**) are able of detecting large emissions with high spatial resolution but less spatial coverage.

Thus, the spatial and temporal coverage of the current generation of spaceborne instruments alone is not sufficient to provide the necessary observational coverage. Upcoming missions such as CO2Image **?** or Carbon Mapper (**?**) will improve the coverage. In addition, airborne instruments such as the next-generation Airborne Visible/Infrared Imaging Spectrometer (AVIRIS-NG) (**??**) are able to supplement these data by inspecting crucial areas and providing images of CH_4 column enhancements with a spatial resolution of around 1 m to 5 m.

For the estimation of flux rates from image data, multiple methods have been proposed, such as the source pixel method (**?**), the Gaussian plume inversion (**??**), the integrated mass enhancement (IME) (**??**) or the cross-section flux method (**????**). These methods require the external input of effective local wind speed information. While 3D wind fields are readily available from various meteorological models, they are not always accurate, and the effective wind speed the plume has experienced is not uniquely defined. This makes uncertainties in the wind

speed a leading source of errors for these methods, and systematic under- or overestimation of the real wind conditions would translate into biased results.

The study of **?** provided evidence that the 2-D plume morphology contains useful information about the local wind speed conditions. This led to the idea of approaching the regression task of flux rate estimation using pattern recognition methods. While deep learning methods such as convolutional neural networks (CNNs) were used before in remote sensing application (**?????**), **?** introduced the method to the application of emission rate estimation of methane.

It showed the potential of CNNs for flux rate estimation without the need for external wind speed information, using synthetic remote sensing image data for the AVIRIS-NG instrument. The use of realistic synthetic data is crucial for the training processes of the CNNs, as it requires knowledge of the true emission rate.

The study sparked applications to other instrument including satellites such as PRISMA (**?**), Sentinel-2 (**?**) or the GHGSat-C1 (**?**) to estimate methane emissions or to the upcoming CO2M mission (**?**) to estimate the CO_2 emission rates of power plants. In **?** the short-wavelength infrared (SWIR) spectral radiances from the PRISMA satellite and the Red-Green-Blue (RGB) bands were used to predict the plume mask, the concentration map, detect a plume and then perform the regression task of emission rate estimation. **?** uses column-averaged mole fractions of CO_2 (XCO_2) and wind speed fields as a base input to estimate the emission rates of power plants and tests how the addition of estimated plume masks or NO_2 concentration fields to the base input affect the results. The study of **?** is closest to **?**, as it also only uses column-integrated CH_4 images as an input. However, it compares the performance of six established CNN architectures to find the best-performing architecture for this regression task. Their findings support the use of an EfficientNet-V2L (**?**). In **?** column-integrated CH_4 images are used to predict the plume mask and use the concentration map, the binary mask and the 10−m wind speed as an input to calculate the flux rate either by using a CNN or the IME method.

In this study we present changes to the training process that lead to an increased performance and resolve bias patterns present in previous studies. We illustrate these improvements in direct comparison to **?** as we use the same data for training and testing the model. In addition, we suggest an optimization metric that leads to meaningful error estimates for the predicted flux rates. Finally, we suggest an analysis that reveals limitations of the current applicability of the methodology and provides insights into this method that help further improve the emission rate estimation.

The improvements we suggest are very general and should be applicable to the studies mentioned above, as well as future studies to enhance model performance.

Section 2 provides technical details of the data used and showcases the process used to generate the realistic plume images. The neural network architecture and the details of our training process are described in Section 3. In Section 4 we present our results and an analysis of the model performance. In the final section we summarize and discuss the finding of our study.

2. Data

This section describes the data that were used for training, validating and testing the deep learning model. We use the simulation data previously described in **?** with the same split into disjoint training (80%), validation (15%) and testing data (5%). Therefore, we will only provide a brief motivation and summary, and highlight relevant aspects. The data aim to provide realistic images of column-integrated CH_4 plumes over urban, desert or agricultural areas, as they would be measured by the AVIRIS-NG instrument.

Large eddy simulations (LES) were used to generate time-resolved three-dimensional CH_4 distributions resulting from point source emissions under different geostrophic wind speed conditions. The complete LES model setup description can be found in **?** and the parameterization and initial parameters in **?**. Based on the findings of studies like **?** and **?**, the wind speed is assumed to be the most influential parameter regarding the spatial patterns of total-column CH_4. Together with the total column enhancement across the scene, these two quantities provide a strong foundation for the flux inversion of the scene. Assuming that the self-buoyancy of methane is negligible after it is mixed, we can scale the simulated plumes during the post-processing to obtain plumes simulating different flux rates. This makes it possible to mainly vary the geostrophic wind speed conditions in the LES and still obtain a dataset that offers a good representation of observable emission scenarios. The parameters selected for the range of the desired emission rates (0 kg h^{-1} to 2000 kg h^{-1}), the geostrophic wind speed conditions (1 m s^{-1} to 10 m s^{-1}), and the surface sensible and latent heat fluxes (400 W m^{-2} and 40 W m^{-2}) are based on typical field conditions that were found during the Four-Corners campaign (**?**).

The final plume library consist of 7000 3-D fields, which are equally distributed among the different geostrophic wind speed conditions. Every used scene has a size of 300×300 pixels with a total size of 1.5×1.5 km^2, which results in a spatial resolution of 5 × 5 m^2. For the application to airborne point source estimation, these fields are integrated vertically, weighted with the column averaging kernel of the AVIRIS-NG retrieval. To create a realistic plume image from these synthetic noise-free 2-D snapshots of column-integrated CH_4 plumes, a realistic noise estimate for such measurements is required. For this we use the same 3000 retrieved scenes from AVIRIS-NG flight lines over desert, urban or agricultural areas with the same split into training, validation and test data as in **?**. These scenes do not contain any plumes, but do contain random and correlated noise features that correspond to surface structures typical of the corresponding area. This adds the challenge of distinguishing between surface and plume features and thus provides realistic backgrounds representing typical observable scenes.

The synthetic plume image is then scaled by a random factor to generate different emission scenarios, is rotated by a random angle between −170° to 170° to generate different wind

directions, and is translated randomly by up to 30 pixels to simulate different emission locations. The augmented plume now gets added to a randomly selected and rotated realistic background noise scene and, as a final step, a masking threshold of 500 ppm-m is applied. The masking threshold applied in the last step of the augmentation is crucial, as it removes additional information that is contained in the simulation, but cannot realistically be obtained in a real measurement. Therefore, the threshold for the masking is ideally on par with the real-world performance of the instrument, or even slightly higher, to reduce the quality of a real observation to the quality that the machine learning model was trained on. The final result of this augmentation scheme for three different geostrophic wind speed conditions and three different realistic noise scenes can be seen in Fig. 1.

The distribution of the random flux rates is uniform, with all fluxes below $3\,\mathrm{kg\,h^{-1}}$ set to 0 to generate more cases without a plume present. From every turbulent realization in our LES plume bank we generate 50 scenes for the test data and 20 scenes for the validation data using the described augmentation scheme. This increases the amount of available data, especially for the test data, which provides us with better statistics when it comes to evaluating the model performance. For the training data we did new augmentations for every training epoch and used each turbulent realization five times per epoch.

3. Method

This section provides an overview of the methods that we employ and the changes we made to prior approaches. Section 3.1 describes our network architecture and the loss function we used for optimization. In Section 3.2 we describe our training process where we introduce changes that lead to an improved performance.

3.1. Deep learning setup

Over the years, numerous state-of-the-art machine learning, especially deep learning, algorithms for pattern recognition tasks have been developed (**?**, **?**, **?**, **?**, **?**, **?**, **?**). Most of theses algorithms make use of different convolutional neural network (CNN) architectures and were used for classification tasks on ImageNet data (**?**). For the application to a regression task, such as the emission rate estimation of point sources, these established architectures either have to be modified (**?**) or one has to create a different architecture (**?**, **?**). In the instances where different architectures were created, this has resulted in more simplistic networks.

Since we follow up on the work of **?**, we compared the custom architecture used there with an established one. We found that the established architecture, in our case a ResNet-50 (**?**), outperformed the simpler architecture of MethaNet. A ResNet architecture uses shortcut connections to pass the identity from one convolutional layer to the next. This allows very deep and easy-to-optimize neural networks to be built. An illustration of the network architecture can be seen in Fig. 2. We modified the last layer of the ResNet architecture by adding two fully connected layers and used a linear activation function at the end of the last layer.

During our training, the model performed better without a dropout layer, which was used in **?**. However, this could be due to changes in the training process, which we will discuss in Section 3.2. The last layer of our model requires two output parameters: one for the estimated flux and one for a variance estimate, which is required by our choice of loss function.

As a loss function we propose the use of the Gaussian negative log likelihood (GNLL)

$$\ell(\mathbf{x},\theta) = \frac{1}{2}\left(\log(\sigma^2(\mathbf{x},\theta)) + \frac{(\hat{j}(\mathbf{x},\theta) - j_{\mathrm{truth}})^2}{\sigma^2(\mathbf{x},\theta)}\right) + const, \quad (1)$$

with j denoting the flux, $\mathbf{x}$ the input image and θ the neural network parameters. This choice of loss function is reasonable due to the observed heteroscedasticity of the flux estimates that is visible in this and previous works. In addition, this allows us to use the estimated variance of the model as an uncertainty for the predicted emission rates. The estimate of the variance is the second output of the last layer of our neural network. The variance weights the deviation of the prediction from the ground truth, and in addition it adds an offset in the form of the normalization factor of the normal distribution. The neural network learns to balance these two aspects while minimizing the loss.

3.2. Training process

The training process is a crucial part that is relevant for the performance of a deep learning model. Here we want to address the changes we made and point out differences compared to previously published approaches. As mentioned in the last section we use a GNLL. However, all the changes we discuss here were initially implemented using a mean square error (MSE) loss function and would also improve results for this loss function.

The major improvement that we would like to suggest is related to the squared nature of the MSE or GNLL. In previous publications (e.g. **?**, **?**, **?**, **?**) a similar bias pattern can be observed. The model tended to underestimate large fluxes and overestimate smaller fluxes.

We found the cause of this problem to be a combination of upper and/or lower limits to the flux rates and the properties of the loss function. Due to the quadratic nature of the loss function, outliers are weighted heavily. This provides a strong incentive to reduce the spread as much as possible and, more importantly, it favours reduction of the spread of the predictions over the correct estimation of the mean value. The presence of an upper or lower bound allows for a solution with less outliers, which is a favourable outcome in terms of the computed loss but induces a bias.

For the case of an upper limit, this causes the model to reduce the spread of the data towards the upper bound, which leads to the observed underestimation of the mean value. Similar effects would also be observed when using non-quadratic

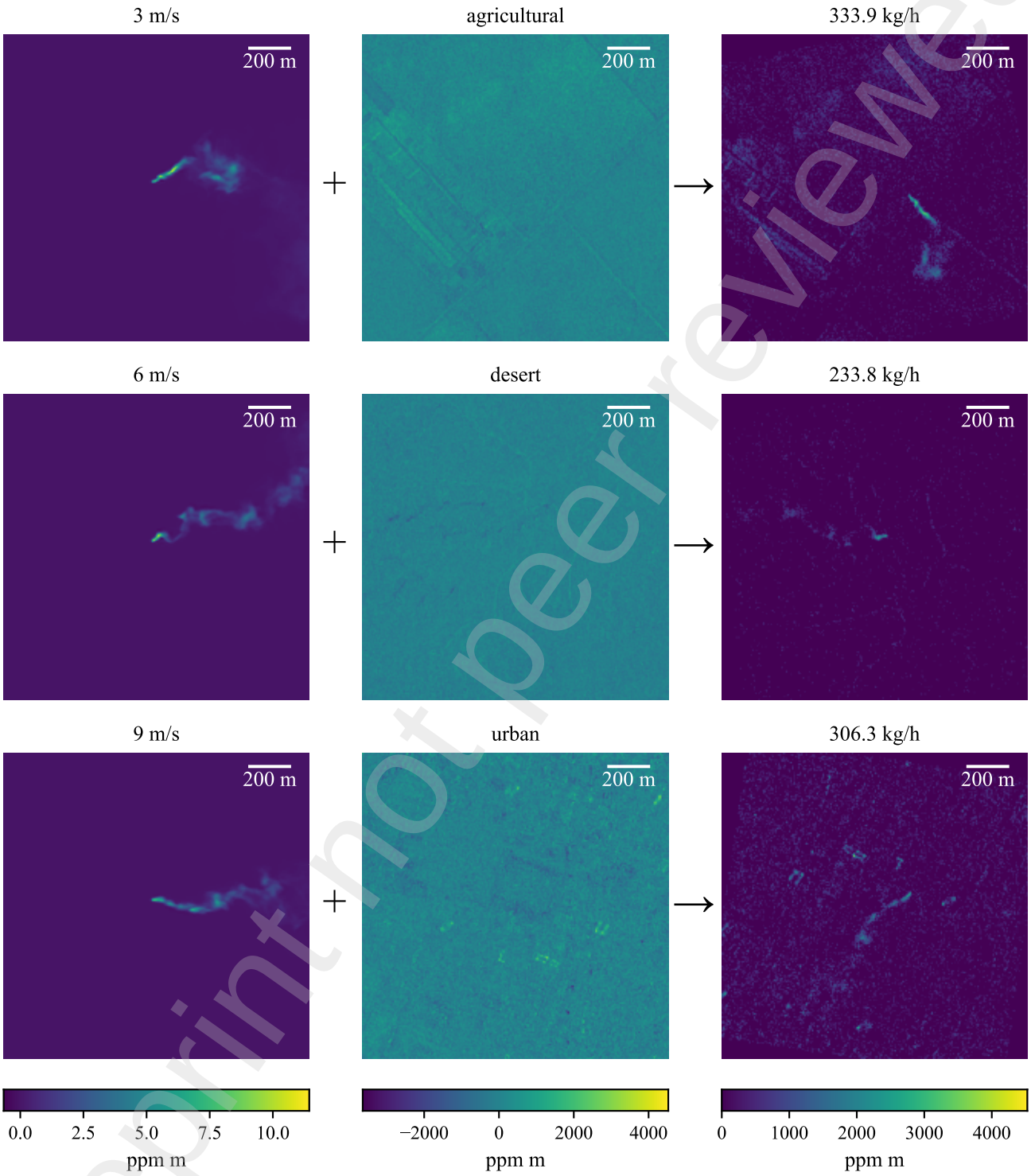


Figure 1: Examples for augmented realistic plume scenes consisting of a simulated plume at different geostrophic wind speeds (left) and a realistic background noise scene from agricultural, desert, or urban areas (center). The plumes are randomly rotated, shifted and scaled, and added to the randomly rotated background noise. The sum is masked with a threshold of 500 ppm m to get the scenes for analysis (right).

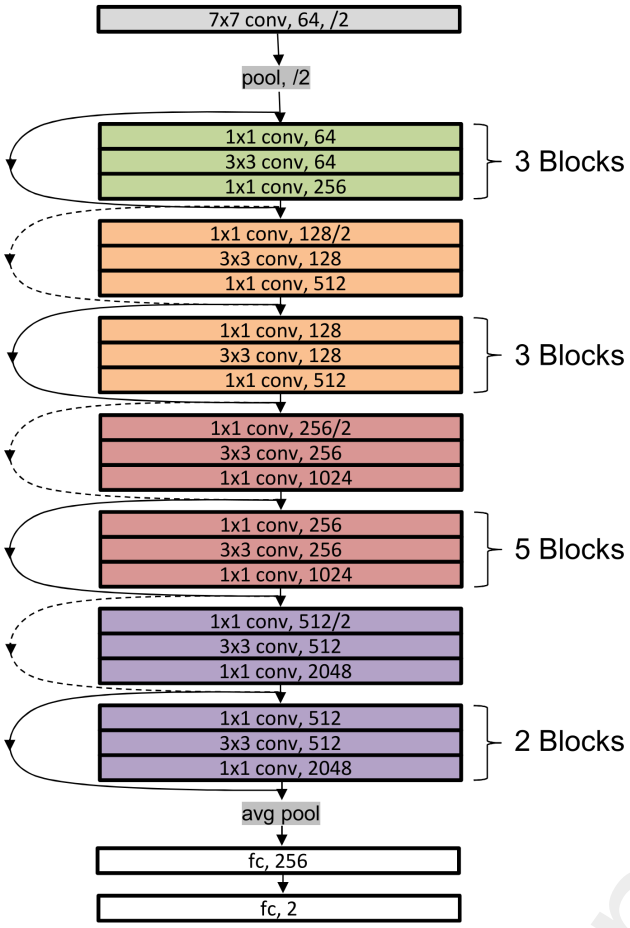


Figure 2: Illustration of the slightly modified ResNet architecture that we used during this study.

optimizer such as the mean absolute deviation (MAD), however these solutions would be less rewarding in terms of the computed loss.

This is undesirable as, in addition to the bias, it also makes the predictions for fluxes near the boundaries meaningless, as they are heavily influenced by systematic restrictions of the training and are thus not comparable to the rest of the predictions.

To solve this problem, we suggest to simply extend the flux domain of the desired range in the training data. If the upper and lower bounds are sufficiently far away from the domain of interest, the effect induced by its presence gets negligibly small. In our case, a maximum flux of 3500 kg h^{-1} for a validation and test range up to 2000 kg h^{-1} was sufficient.

For the lower flux rates the problem is more complex, as there is the emerging difficulty of noise dominating the measurements, which makes it impossible to predict the flux rates beyond a certain point. However, the impact of low flux plumes on the loss is also smaller, which makes the model less sensitive towards such a bias. In our case we selected a lower bound of 0 kg h^{-1}, since we wanted to include scenes without a plume for the training to make it a possible scenario. During our analysis of the model performance in Section 4.2 we can see that the model stabilizes quite quickly and the lower bound is less of a problem than the upper bound.

In addition to extending the training domain we used the mean percentage error (MPE) to prevent our model from overfitting, and in order to select a model that shows little bias. We used the MPE in addition to the GNLL in the form of a threshold to select our best model from the training process. By using the MPE as a threshold we mitigate the effects of underestimation caused by relative measures. For our specific application, we selected the model with the best GNLL loss that showed a MPE lower than 1% for fluxes larger than 100 kg h^{-1}.

We consider only estimates for sources with over 100 kg h^{-1} to exclude the impact of scenes where the model might not be able to perform proper estimates due to the large noise contribution. This allows us to prevent the model from overfitting and, in addition, it filters out poorly performing models, as the bias of the model is not represented well by the optimization metric. This is of particular importance for this application, as the predictions should be reliable and unbiased when averaged over many measurements.

4. Results

In this section, we present the results of our model on realistic noise scenes using the test dataset described in Section 2, following the training process described in Section 3.2. The model performance is shown in Section 4.1 and in Section 4.2 we analyse the performance and stability of the model.

4.1. Application to realistic noise scenes

The application of the deep learning model to the test data leads to the results shown in Fig. 3. The scattered data as well as the data clustered into flux rate ensembles show a nice linear behaviour and the means of the respective groups seem to be almost unbiased.

Fig. 4 shows the relative deviations of the flux groups to get a closer look at the deviations from the 1:1 line. This reveals some slight biases of up to 1.8% for scenes around 500 kg h^{-1} and a large deviation for the smallest flux group. This instability for low fluxes is expected for a relative deviation since at a certain point the plumes will no longer be recognizable and in addition the lower bound on the flux range causes a bias, which results in huge relative errors and thus is no reason for concern. Therefore, the model overall stays very stable and shows little to no bias over the whole desired range of fluxes.

For the different ensembles, around 85% of the estimates are within a distance of one standard deviation of their respective flux group and around 80% for the whole dataset. This indicates that neither the distribution of the estimated fluxes within an ensemble nor over the whole dataset is normally distributed and the ensemble standard deviations are not an accessible metric to quantify the uncertainty of the model.

However, given our selected GNLL loss metric, the model provides an uncertainty estimate in the form of a variance, which we can use to compute the standard deviation for every predicted emission rate. The predicted standard deviations

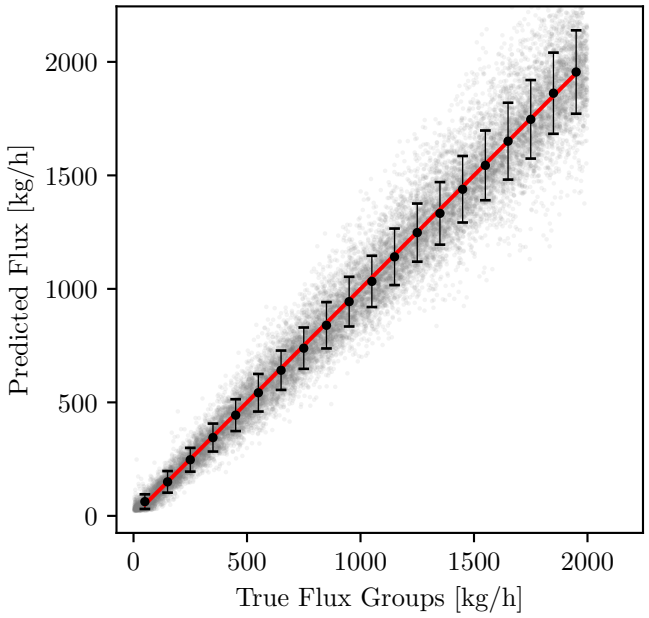


Figure 3: Scatter plot of the predicted fluxes against their ground truth. The data have been separated into groups containing scenes in a range of 100 kg h^{-1} to provide their means and standard deviations for a more quantitative estimate of their distribution.

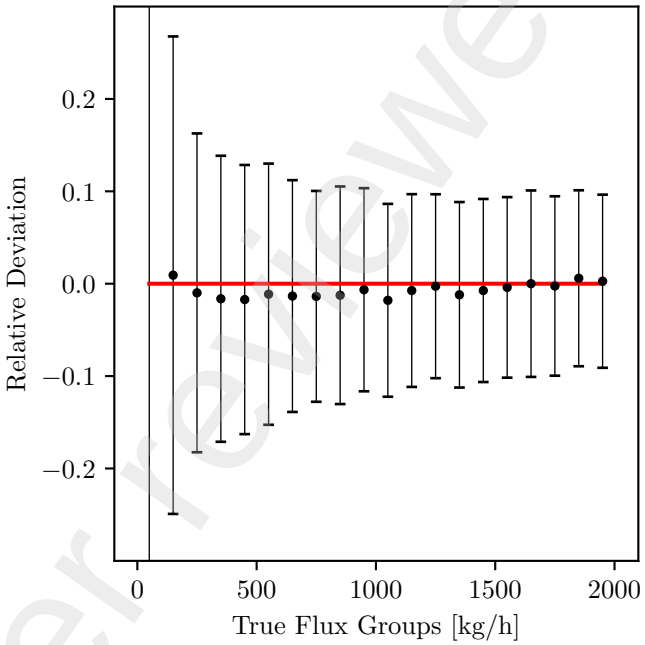


Figure 4: Plot of the relative deviations of the predicted fluxes, separated into groups spanning a range of 100 kg h^{-1}, against their ground truth.

colored according to their respective wind speed conditions are depicted in Fig. 5. The data show a tendency towards clustering in wind speed groups, especially for low wind speed situations. Fig. 6 shows the distribution of the difference of the ground truth from the estimated flux divided by the error estimate of our model for the respective emission prediction. The depicted bell curve is slightly skewed, which is to be expected due to the slight tendency towards underestimation of the model for which we did not apply any corrections.

However, when looking at the absolute deviations over the whole dataset of the predicted values, we can see the properties one would expect from a Gaussian distribution. This indicates that the variances estimated by the model of the whole dataset are meaningful, especially in direct comparison to the ensemble estimates.

4.2. Model analysis

In this section we take a closer look at the performance of our model and try to assess its stability. From the results presented in the previous section we can see that the predictions in an ensemble are not normally distributed, and thus the variance does not provide an accessible summary statistic.

Therefore, we decide to use the mean absolute percentage error (MAPE) to characterise the spread of the predictions and, in addition, the mean percentage error (MPE) to characterize the bias that our model has. To better compare to other work and to measure the linear correlation between the ground truths and our predictions, we also use the Pearson correlation coefficient (r). All these metrics for our model for fluxes larger than

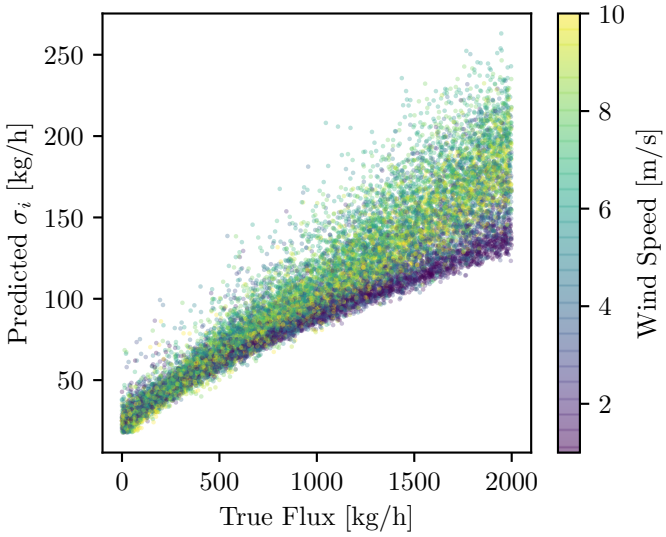


Figure 5: Plot showing the estimated standard deviations for all scenes against their respective true flux rate with a color grading representing their respective wind speed situations.

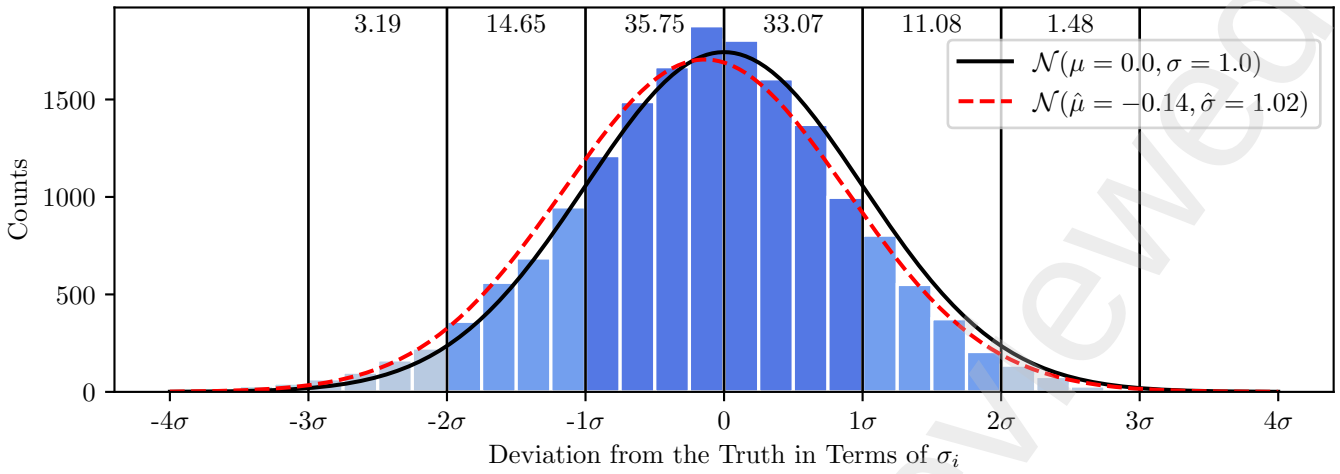


Figure 6: Plot showing the deviations of the predicted from the true flux rates with respect to the estimated standard deviation for the respective scene. The ideal distribution is depicted in black and the best fit to the data in red.

0, 40 or 100 kg h^{-1} can be found in Table 1, and a graphical presentation of the MPE and MAPE for different lower emission thresholds is given in Fig. 7.

It can be seen that the model performance starts to become stable for fluxes larger than around 40 kg h^{-1}. In addition to the grouping of data into flux groups, we also introduce a split into different wind speeds. The relative deviations in the different wind speed ensembles are depicted in Fig. 8 and Fig. 9. This analysis is motivated by a wind-speed-dependant bias that was visible in Fig. 6 in **?**, and also during our development process, that showed under-/overestimated flux rates for high/low wind speed situations.

The model seems to consistently underestimate the fluxes at very high wind speeds, mainly at 9 m s^{-1} and 10 m s^{-1}, over the whole flux range domain. The estimates at 6 m s^{-1} also show a bias. However, compared to the biases at high wind speeds, it varies with the data selection. That is to say, it is not present in other data splits or in the validation data (see Fig. A.12), while the biases at high wind speeds seem to be systematic.

In addition to the consistent wind-speed-dependent bias at the highest wind speeds, there is also a wind-speed-dependent bias that only affects low flux rates, as shown in Fig. 8 and Fig. 9. This instability is also reflected in the increase in the spread of the model.

A plausible explanation for the behavior at low flux rates is related to the masking threshold that is applied, and the increasing influence of the noise, which together make the regression task and the extraction of wind speed information increasingly difficult.

While the high/low bias for low/high wind speeds is visible in the test data, it is more pronounced in the validation data, as shown in Fig. A.12 and Fig. A.13. For the validation data, the point at which the clustering of wind speed situations in Fig. 5 seems to stop and the point at which the bias pattern starts to become apparent are very close to each other. This pattern fits a behaviour that one would expect if the wind speed can no longer be determined and is guessed instead, which would lead to the estimation of some sort of an average wind speed estimate.

Because the plume mass is transported faster by high wind speeds, the overall column-integrated CH_4 enhancement above the threshold in the scene is reduced. Therefore, the point at which scenes are dominated by noise effects is reached earlier than for lower wind speeds. This makes them more difficult to deal with, resulting in a mean estimate that is shifted towards high-wind-speed situations and a more severe impact at higher wind speeds, which matches the observed pattern.

For very low flux scenarios, the presence of a lower bound also leads to an overestimation, as discussed in Section 3.2. This wind-speed-dependent bias also has an impact on the slight skewness of the bell curve in Fig. 6, as it is caused by the emission rate estimates at the highest wind speeds. Fig. 10a shows the distribution of the deviations of the estimates with respect to their corresponding uncertainties, excluding the data from wind speeds above 8 m s^{-1}, which shows a clear improvement. In Fig. 10c and Fig. 10b the distributions for wind speeds of 9 m s^{-1} and 7 m s^{-1} as an example for a biased and a unbiased wind speed group are displayed. The overall distribution for the different wind speed ensembles are thus, with the exception of the 9 m s^{-1} and 10 m s^{-1} ensembles, approximately normally distributed. Even for very high wind speeds, the absolute deviation from the truth still is close to the properties a normal distribution.

Fig. 11 shows some example scenes for which not only the emission estimate was poor, but also the estimated error was underestimated, resulting in a prediction that is more than four standard deviations away from the true emission rate. Therefore, these scenes represent scenarios where the model not only predicts the emissions poorly, but also is overly confident in its performance. The scenes which fall into this category are usually at higher wind speeds and show signs of instability with respect to the direction of the wind (see Fig. 11c, Fig. 11a

threshold $kg\,h^{-1}$	MPE %	MAPE %	r %
0	3.24	13.86	97.98
40	−0.47	10.29	97.88
100	−0.72	9.48	97.67

Table 1: Summary statistics to describe the model performance over all scenes of the test set with a flux larger or equal to the given threshold.

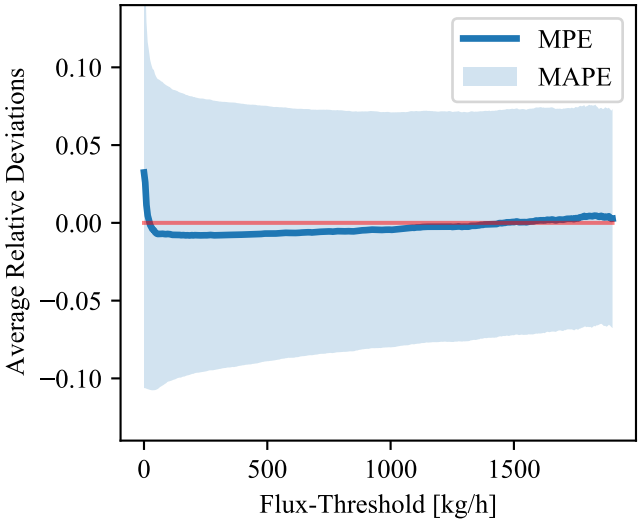
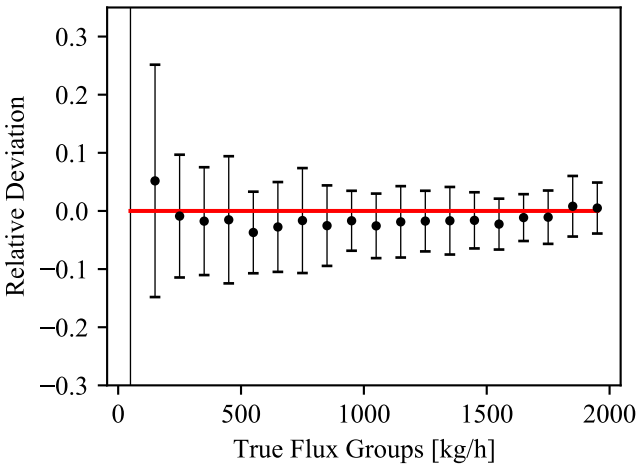
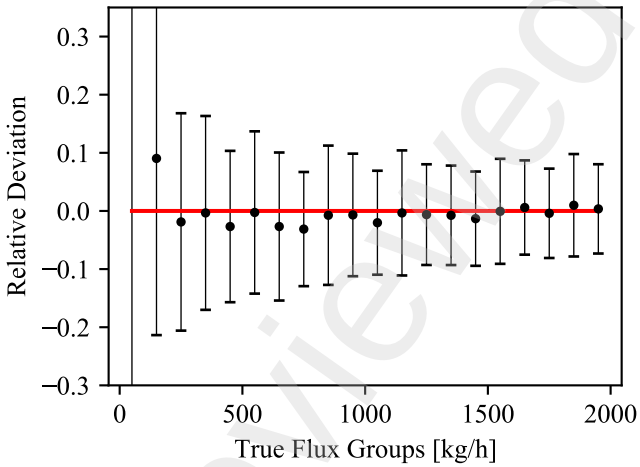


Figure 7: Plot of the MPE and the MAPE over the dataset, excluding all scenes with a flux rate smaller than the given threshold.

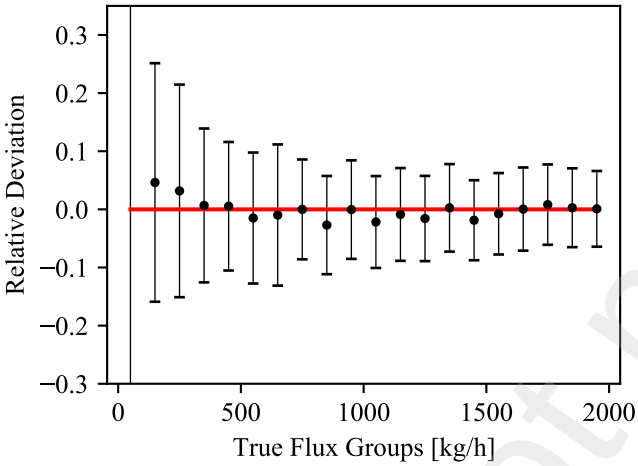
and Fig. 11b), or there is barely a plume visible in the scene (Fig. 11d). The same turbulent realisations with different background noise and at different flux rates tend to be poorly estimated, which indicates that the model is simply not capable of providing reasonable estimates for certain turbulent patterns. This is most likely linked to their scarcity in the training data.



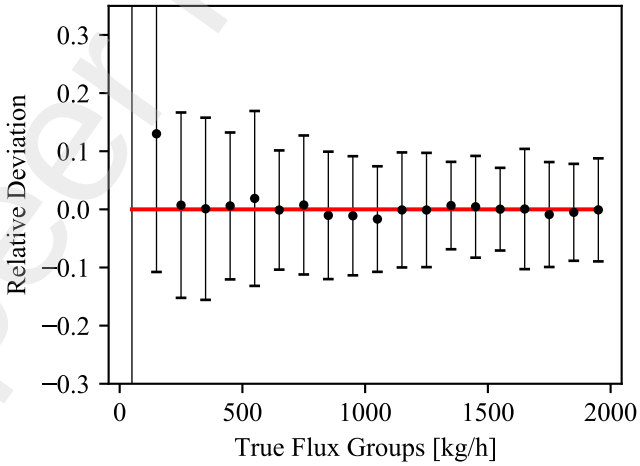
(a) Wind speed 1 m s^{-1}



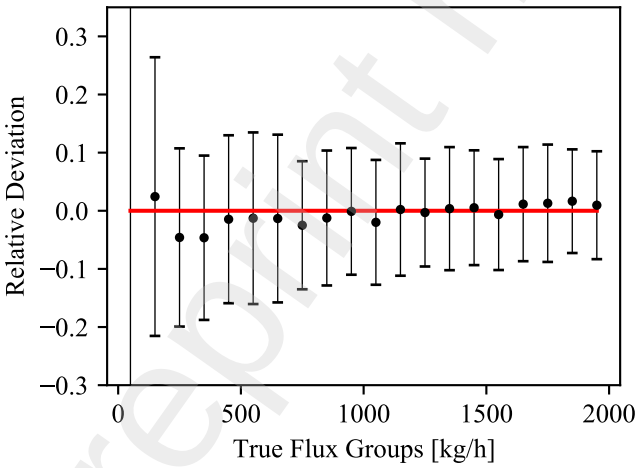
(b) Wind speed 2 m s^{-1}



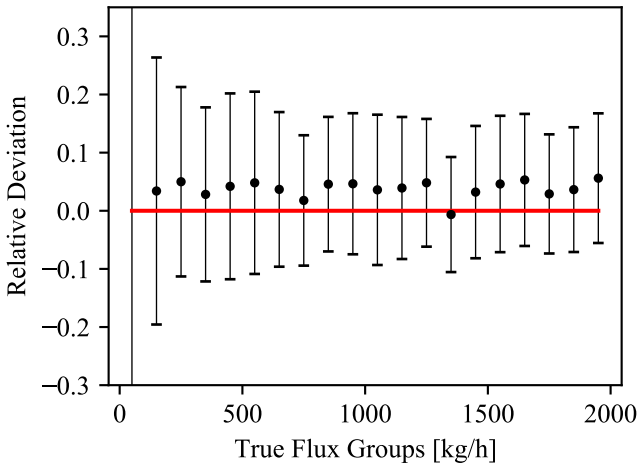
(c) Wind speed 3 m s^{-1}



(d) Wind speed 4 m s^{-1}

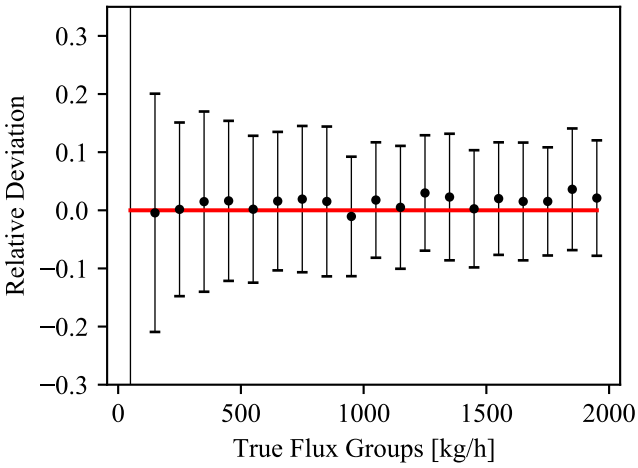


(e) Wind speed 5 m s^{-1}

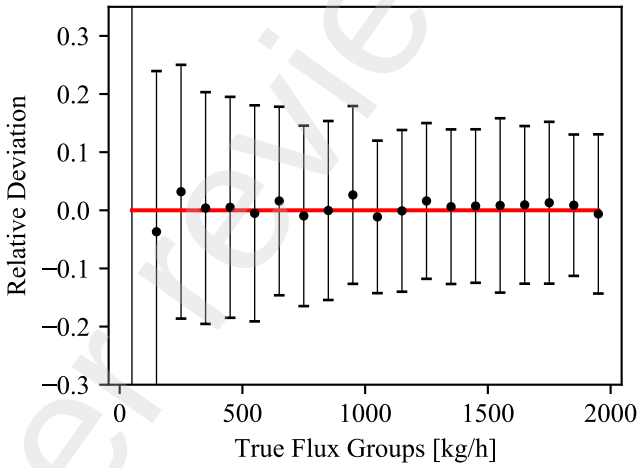


(f) Wind speed 6 m s^{-1}

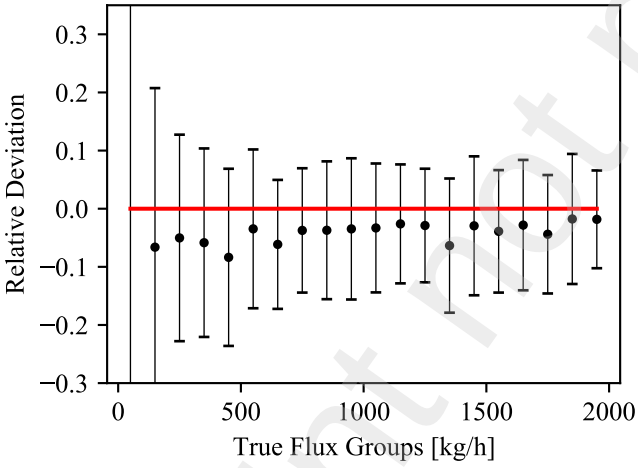
Figure 8: The figures show the relative deviations of flux rate ensembles (spanning a range of 100 kg h^{-1}) of the predicted flux rates from the true flux rates for different wind speed conditions.



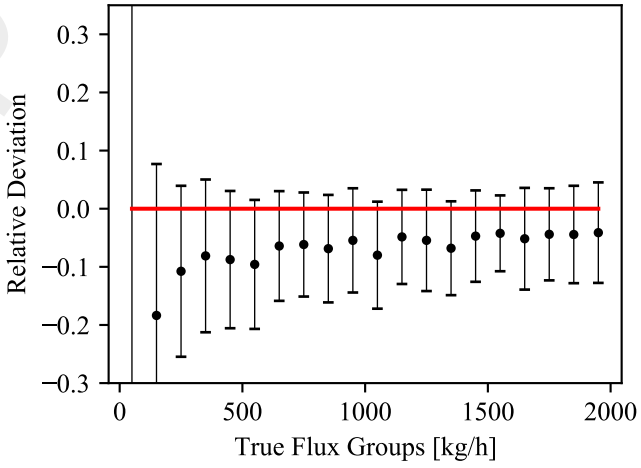
(a) Wind speed 7 m s^{-1}



(b) Wind speed 8 m s^{-1}

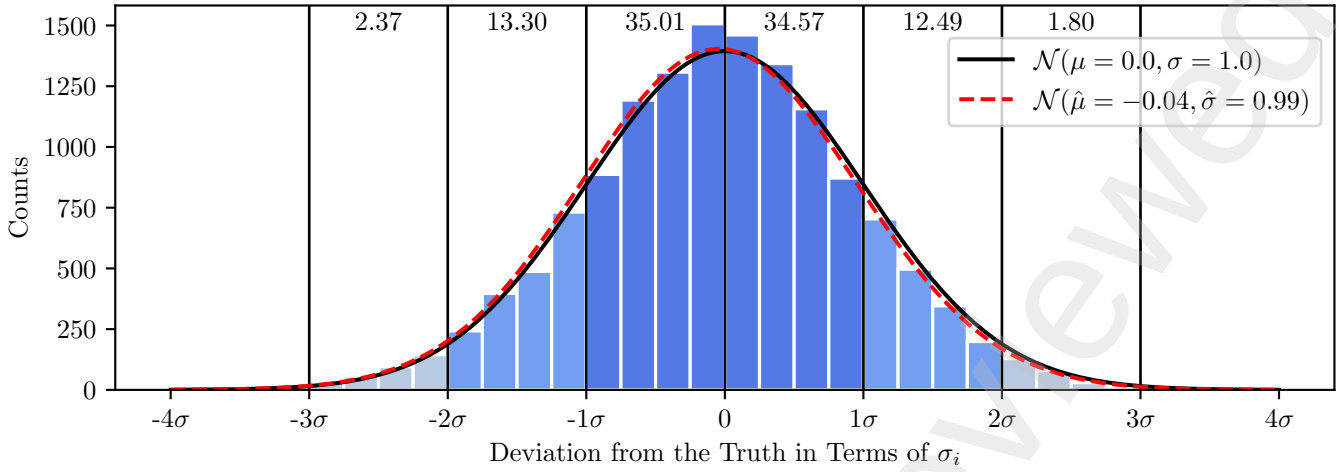


(c) Wind speed 9 m s^{-1}

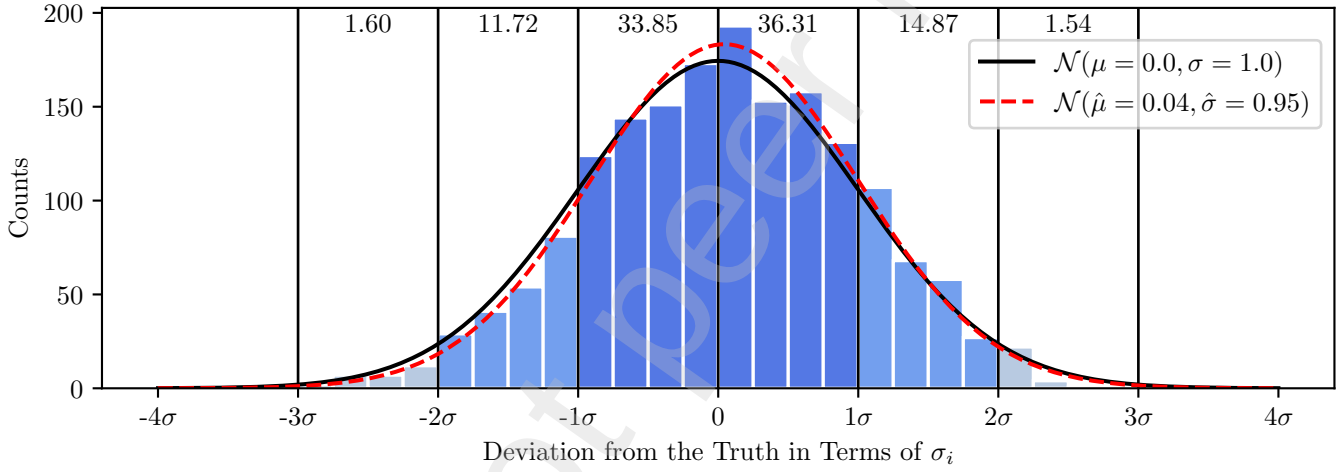


(d) Wind speed 10 m s^{-1}

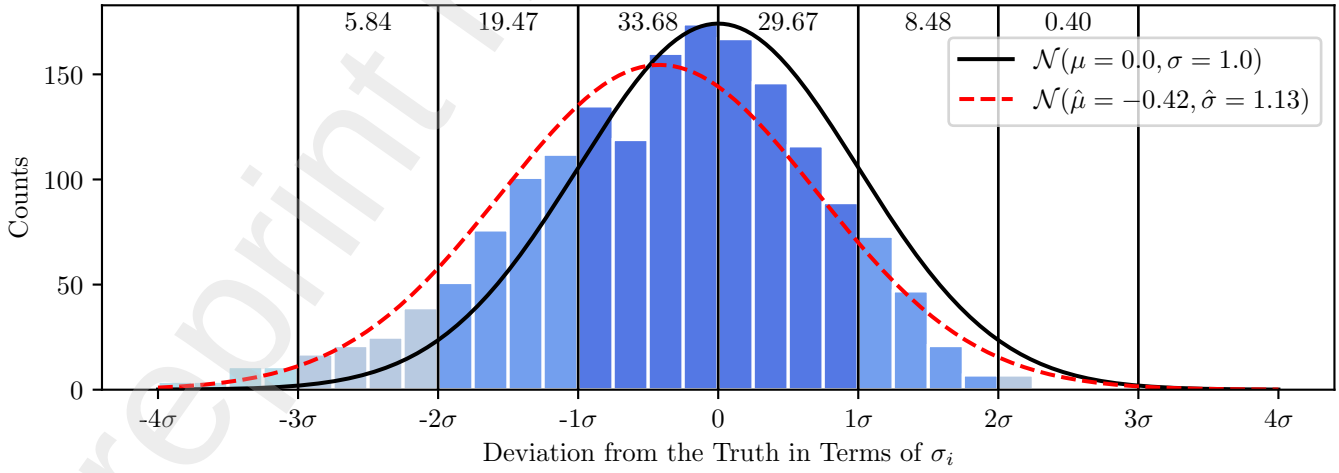
Figure 9: The figures show the relative deviations of flux rate ensembles (spanning a range of 100 kg h^{-1}) of the predicted flux rates from the true flux rates for different wind speed conditions.



(a) Wind speed ensembles from 1 m s^{-1} to 8 m s^{-1}

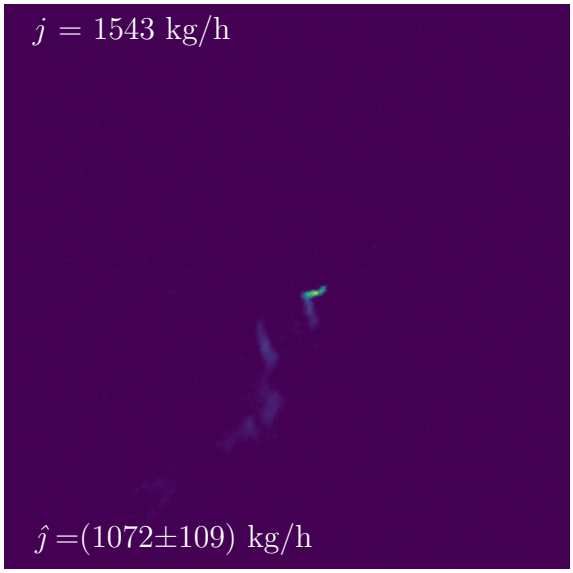


(b) Wind speed 7 m s^{-1}



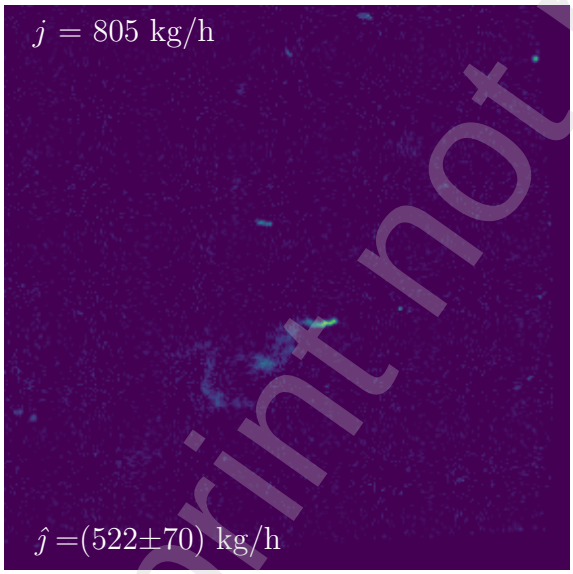
(c) Wind speed 9 m s^{-1}

Figure 10: Plots showing the deviations of the predicted from the true flux rates with respect to the estimated standard deviation for the respective scene for different wind speed ensembles. The ideal distributions are depicted in black and the best fits to the data in red.

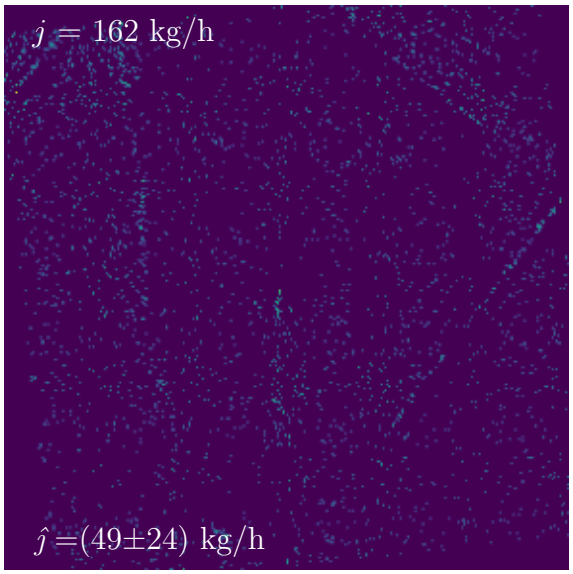


(a) Wind speed of 9 m s^{-1}

(b) Wind speed of 9 m s^{-1}



(c) Wind speed of 8 m s^{-1}



(d) Wind speed of 10 m s^{-1}

Figure 11: Examples of scenes with deviations from the true flux rate that are larger than 4 σ, measured by their respective error estimate.

5. Summary and conclusion

In this study, we applied a modified ResNet-50 to simulated high-resolution airborne imagery. We used images that correspond to 2-D column-integrated methane enhancements of different emission scenarios under different geostrophic wind speed conditions. The synthetic data were created by adding augmented LES plumes and realistic background noise scenes that were obtained from AVIRIS-NG flights. The emissions range from $0\,\mathrm{kg\,h^{-1}}$ to $2000\,\mathrm{kg\,h^{-1}}$ and the geostrophic wind speeds from $1\,\mathrm{m\,s^{-1}}$ to $10\,\mathrm{m\,s^{-1}}$.

We trained the neural network to solve the regression task of flux estimation from 2-D plume images without further input in the form of wind speed information. During this study, we developed improvements to the training process that result in better-performing and more reliable predictions compared to prior works. We argue for an extended training flux realm and the use of additional evaluation metrics for hyperparameter tuning on the validation data. In addition, we propose the use of the negative Gaussian loss likelihood as an optimization metric as it better fits the problem and allows for the prediction of estimation uncertainties for each prediction.

For our evaluation using the test dataset we use the mean percentage error (MPE) as a measure for the bias and the mean absolute percentage error as a measure for the spread of our predictions. Our results show a MPE of 3.24% and a MAPE of 13.86% over the whole dataset and a MPE of −0.47% or −0.72% and a MAPE of 10.29% or 9.48% for scenes with a flux higher than $40\,\mathrm{kg\,h^{-1}}$ or $100\,\mathrm{kg\,h^{-1}}$. Further, the model achieves a Pearson correlation coefficient (r) of 98%.

Therefore, the model provides accurate estimates and a spread that is a significant improvement in direct comparison to **?**. When comparing to other studies, where the dataset is different and the spatial resolution, noise properties and the flux ranges show different properties, a direct comparison of the summary statistics provides only limited information, but some improvements can be seen. By visual comparison we can see that our model shows less bias over the different flux ranges than comparable studies such as e.g. **?** Fig. 6, **?** Fig. 6 or **?** Fig. 9. This is, however, most likely a consequence of the different training process, and not due to the use of a ResNet-50, which performed worse than the EfficientNet-V2L in the study of **?**. In addition, we provide error estimations for every predicted flux rate, and these errors follow a normal distribution over the whole test dataset. Thus, unlike error estimates that are calculated from ensemble statistics, these seem to provide a meaningful measure of uncertainty for individual flux predictions. Given the high stability of the performance of our model over different flux ranges, the quality of the predictions, as well as the error estimates, we expect that it should generalize well on different, arbitrary observed flux distributions.

Our analysis of the model performance includes a separation of the data into their different geostrophic wind speed conditions. This revealed a wind-speed-dependent bias for scenes at 9 and $10\,\mathrm{m\,s^{-1}}$ that seems to be systematic and applies over all flux rates.

In addition, we can observe a systematic bias for scenes with low emission rates, which shows underestimations for high and overestimation for low wind speed conditions. Therefore, this bias follows the pattern that one would expect to observe when the wind speed can no longer be properly estimated but instead is guessed, to provide results that fit on average. This bias agrees with and adds to the results of **?**, which, in a plume detection task, found that plumes become undetectable at around $50\,\mathrm{kg\,h^{-1}}$. Our analysis adds to this by pointing out that plumes might be detectable at lower flux rates, but that there is not sufficient information left to extract wind speed information reliably, which causes instabilities in the emission estimation.

This along with the apparent clustering visible in the estimated variances (Fig. 5) indicates that there is actual wind speed information that is being extracted from the plume image.

The wind-speed-dependent biases result in a negative impact on the error estimation of the network, especially when it comes to high wind speeds. The average model performance remains unaffected by these biases, with the deterioration for low fluxes only causing an increased spread. However, this makes the model not entirely independent of the distribution of the data with respect to the wind speed conditions.

Therefore, the model performance would degrade for specific conditions: on average very high wind speeds, or for low flux rates and on average either very low or very high wind speed conditions. This could be addressed through the training of specialized models for extreme wind speed conditions, especially when targeting low flux rates.

It should be noted that, as in **?**, the LES plumes were simulated with flat topography and with the emissions released at ground level. Inspection of scenes where the model performs poorly and shows a large deviation from the truth with respect to the estimated uncertainty, as seen in Fig. 11, reveals that the model has problems with unstable turbulent realizations at high wind speeds. While many CH_4 emission sources are in flat areas, this suggests that for more complicated topographies, where transport is more complicated and such scenarios may be more frequent, carefully selected LES training data would be required.

We have shown in direct comparison that the changes we introduced to the training process and the hyperparameter tuning increase the model performance of deep learning methods when it comes to flux estimation. The methods described here should be transferable to other applications of this methodology for other measurement instruments. Furthermore, we present a thorough analysis pipeline for deep-learning-based models when it comes to flux estimation, which allows for the model performance to be characterized and highlights its limitations. The limitations observed during this analysis reveal weaknesses that can be addressed in future studies to further increase the performance of the methodology. Along with the addition of uncertainty estimates on the predictions, the characterization of the limitations of the method are crucial when it comes to the use of this method for real data. Our model is able to provide reliable estimates for a large range of wind speed conditions and emission rates, and should be applicable to past and future AVIRIS-NG flight campaigns. This however is out of the scope

of this study and will be part of future studies.

Funding

This work was supported by the BMWK-funded project CO2KI (FZK50EE2212).

CRediT authorship contribution statement

Thomas Plewa: Conceptualization, Methodology, Investigation, Visualization, Software, Formal analysis, Data Curation, Writing - Original Draft. **André Butz**: Conceptualization, Resources, Supervision, Funding acquisition, Writing - Review & Editing. **Christian Frankenberg**: Conceptualization, Resources, Writing - Review & Editing. **Andrew K. Thorpe**: Data Curation, Writing - Review & Editing. **Julia Marshall**: Conceptualization, Resources, Supervision, Funding acquisition, Writing - Review & Editing

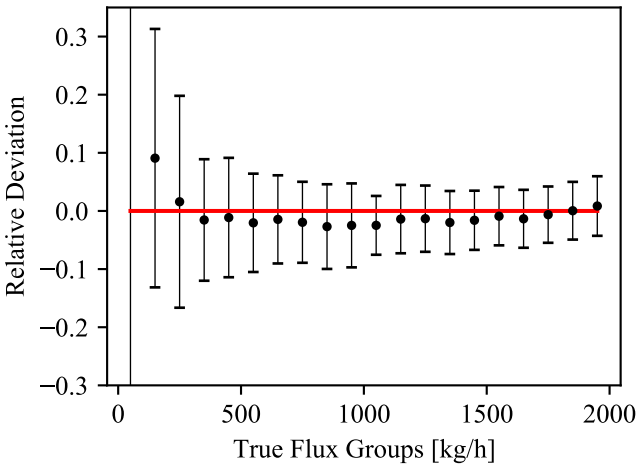
Declaration of Competing Interest

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

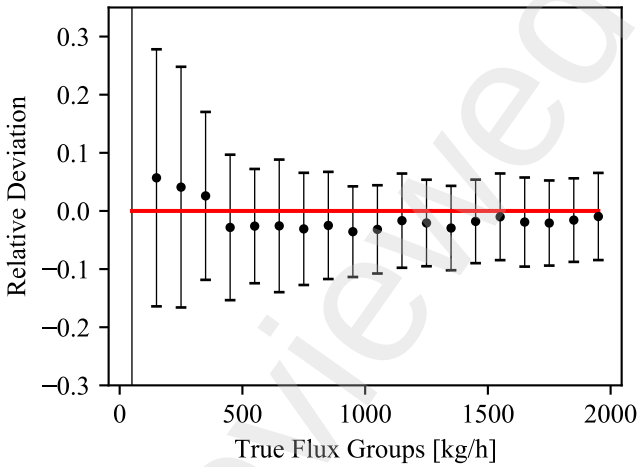
Acknowledgements

This work used resources of the Deutsches Klimarechenzentrum (DKRZ) granted by its Scientific Steering Committee (WLA) under project ID bd1231 and bb1170.

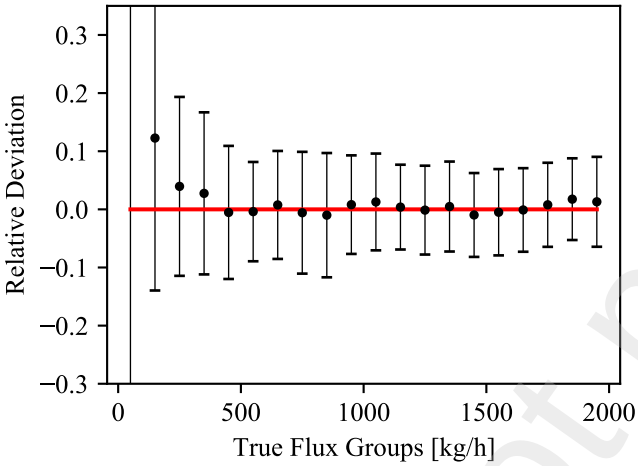
Appendix A. Appendix



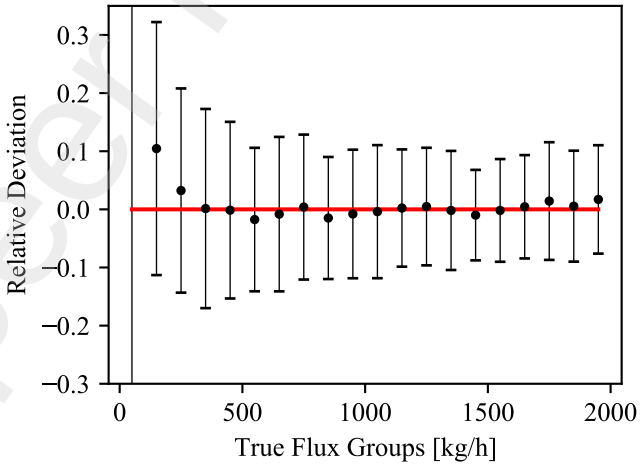
(a) Wind speed 1 m s^{-1}



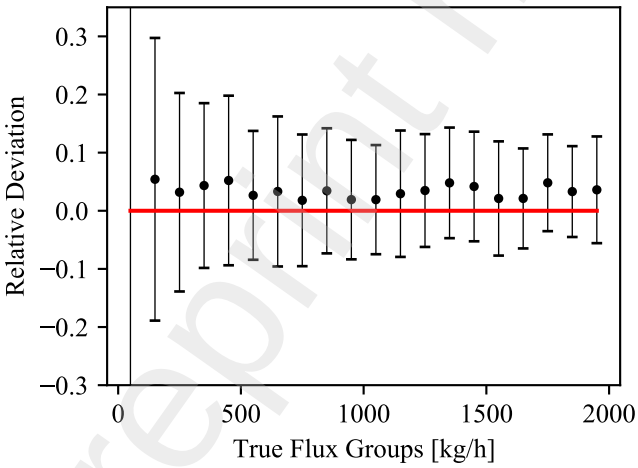
(b) Wind speed 2 m s^{-1}



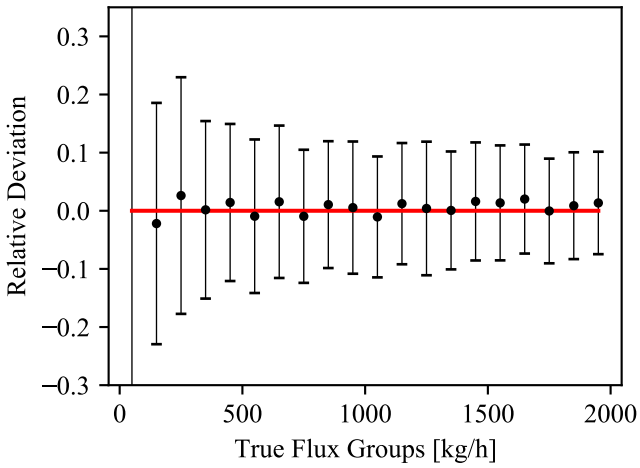
(c) Wind speed 3 m s^{-1}



(d) Wind speed 4 m s^{-1}

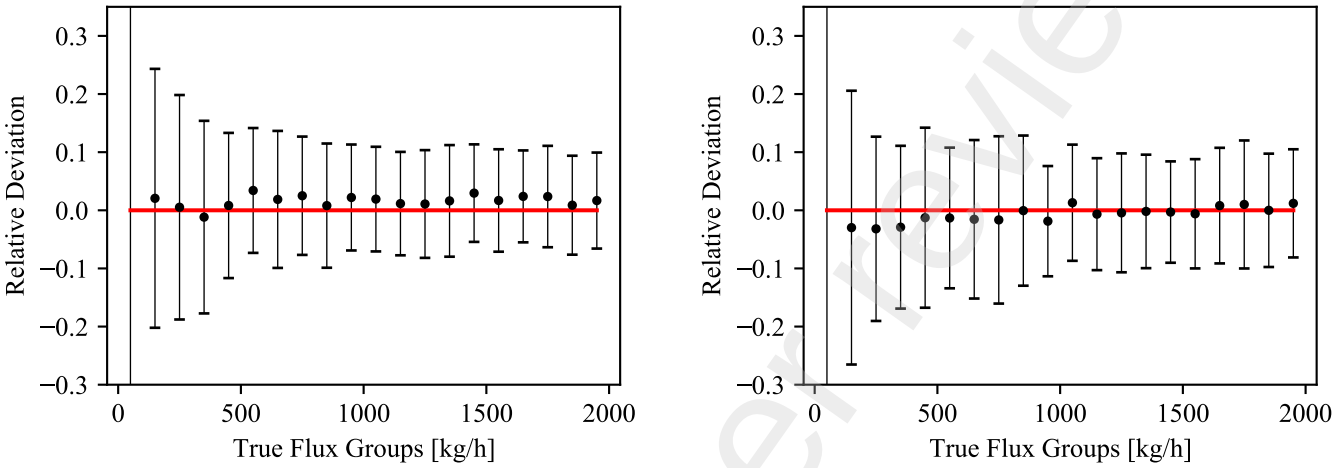


(e) Wind speed 5 m s^{-1}



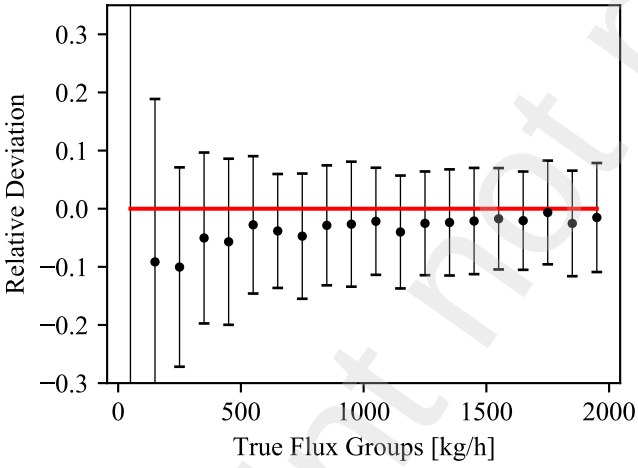
(f) Wind speed 6 m s^{-1}

Figure A.12: The figures show the relative deviations of flux rate ensembles (spanning a range of 100 kg h^{-1}) of the predicted flux rates from the true flux rates for different wind speed conditions for the validation data.

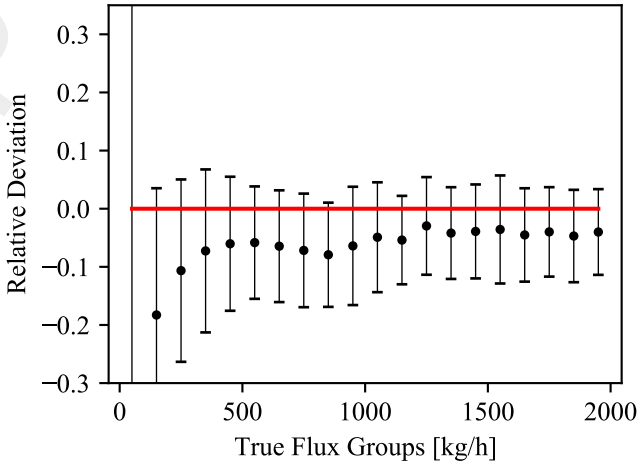


(a) Wind speed 7 m s^{-1}

(b) Wind speed 8 m s^{-1}



(c) Wind speed 9 m s^{-1}



(d) Wind speed 10 m s^{-1}

Figure A.13: The figures show the relative deviations of flux rate ensembles (spanning a range of 100 kg h^{-1}) of the predicted flux rates from the true flux rates for different wind speed conditions for the validation data.