Response of the authors to comments by reviewer 2 on the manuscript: "Dynamic savanna burning emission factors based on satellite data using a machine learning approach"

General comments:

This study collected a large dataset of savanna emission factors (EFs), including over 4500 EF bag measurements of CO_2, CO, CH_4 and N_20 during 129 individual fires from 2017 to 2022. Based on this in-situ observations, the authors identified the drivers of EF variability and implemented this variability into global models through dynamic EFs. The optimized machine learning reduced the error in EF estimates by 60-85% compared to static biome averages. They also found seasonal drying resulted in a decrease of the EFs with the fire season progressing, with a stronger trend in open savannas than woodlands. Overall, this is an important study to understand the variability and mechanisms for biome-specific carbon emissions, particularly at the spatial scales. The generated global EF products can be used to better estimate fire-induced greenhouse gas emissions. However, I do have some concerns on the methodology parts, which may need to be addressed before publication.

Roland Vernooij (corresponding author) on behalf of the authors:

We sincerely thank the reviewer for the time and effort in assessing our manuscript, and the constructive comments which helped to improve the quality of this paper. Please find below our point-to-point response to the review. The revised text and updated figures are included in the updated manuscript. A separate 'track-changes' document is included to highlight the changes to the manuscript. Tables and figures referred to in the answers are added at the bottom of this document.

Reviewer 2, detailed comments	**Author's response, reasoning and comments**
1). Biomass burning EFs are highly dynamic both at the spatial and temporal scales for a given fire. For example, EFs may differ a lot as the fire spreads across different vegetation covers and terrain/moisture gradients at the local scale. How well did the collected EF bag measurements represent the total or averaged EFs for each selected fire? Is there a consistent spatial-temporal framework to integrate the concrete EF measurements to reflect the total EFs for all involved fires? Such processing details need to be provided for better understanding the uncertainty of "in-situ" measurement itself.	You raise a valid point which we can only address empirically. In Vernooij et al. (2022), we have described comparisons on EFs measured using a measurement mast and the UAS method (Figs 2 and 3). In this comparison we found that a limited amount of bag samples (8-12) resembled the fire-averaged EFs of the mast relatively well considering the spatiotemporal heterogeneity. The strategy was therefore to take 8-12 samples at a location until visible smoke from both flaming and smouldering had passed the drone and continue to the next location. The fire-average EF of the individual fires is calculated by adding up the EMRs from all the individual bags and calculating the EFs based in that sum. This means samples with high EMRs have a stronger effect on the fire-average EF than low-EMR samples.

	To make sure the features are representative of the vegetation consumed in the fire, we tried to target larger homogenous areas. Rather than marking the fire with features from the entire fire scar, the features are assigned to the samples based on location and timestamp of the sample. The average of all the sample features is then assigned to the fire-WA EF, meaning we only assign data from when and where we sampled. It might be possible to better quantify the spatial and temporal variability within a fire when continuous measurements can be done while being airborne. Currently, however, the equipment is too heavy to be carried by the drone but this may change in the future and at that stage the reviewer's question may be addressed better. We added the following text (P3 L32): "Fires were lit with the aim of being representative of early dry season (EDS, often prescribed) fires and late dry season (LDS) non-prescribed fires. Although some backfires were sampled during the initial phase of the fires, the majority of samples were obtained from the faster 'head' fires, which consumed most of the biomass. Fire sizes generally ranged between 2 to 10 hectares based on UAS drone imagery described by Eames et al. (2021), with exceptions of some fires that would not light and conversely, some fires that burned several hundred hectares. In the EDS, fire size was primarily limited by environmental conditions and fires ceased burning as humidity increased overnight whereas in the LDS, fire size was confined by low-fuel areas like burn scars, roads and prepared fire breaks. Particularly in the LDS, this means a limited fire size does not necessarily indicate limited fire intensity. Emissions were sampled at altitudes between 5–50 m depending on flame height for a duration of 35 seconds, resulting in 0.7 litres per gas sample. On average, we took 35 samples per fire. The sampling methodology involved taking samples from

	a fire passing a certain point –while correcting for wind direction and severity– until no more visual smoke passed the drone anymore. From earlier work (Vernooij et al., 2022a), where we compared the average of these measurements to results using continuous measurements taken at a mast, we have some confidence in the fidelity of this approach."
2). How did the fire induced EFs match the possible drivers at the spatial scale? Are they overlayed by the actual size of each fire, or just at the grid size of 0.25 degree? The latter may introduce large uncertainties.	For the training data we assigned the features to the fire using the highest resolution available. For instance, the fractional tree cover (FTC) would be assigned based on the 500 x 500-meter pixel value (or the weighted average of several of these pixels). For features with a strong diurnal pattern (e.g. VPD, RH or temperature) we took the hourly data, and interpolated this to assign the feature value at the minute of sampling. However, the spatial resolution of these datasets is typically relatively coarse (0.1°) introducing uncertainty. In our global analysis we indeed averaged out feature data over the 0.25° grid cell (filtering out non-savanna vegetation), and subsequently computed the EFs. This was done to match the spatial resolution of GFED4 and analyze global patterns. When looking at smaller regions, using the native resolution of the features (e.g. Figure 2) may reduce these uncertainties but we expect our data to be used mostly within coarse-scale applications.
3). The authors tested a series of machine learning methods and concluded that random forest performed best. Such a part may need data support. Past experiences suggest that the gradient boost MLs such as lightGBM and Xgboost tend to be better than random forest.	We appreciate the advice; We did not use the tools you mention but will include this in future work to test whether the results improve further. When we started this research we have used a suite of approaches (using the scy-kit learn "GradientBoostingRegressor()" function and GridSearchCV hyperparameter tuning) and actually found that the RFs performed slightly better than GBMs, although the difference was very

	small. In comparison, we also tried multilinear regressions, single decision trees and a simple neural network which all performed worse than RFs and GBMs. Since we include the MCE (which is often used as the sole predictor of EFs) from our RF model into the EF models as a predictor, one could argue that the EF models have a sort of gradient boosting step on that RF MCE model. Table 1 (bottom of this document) lists the RMSE and R^2 for different model runs (in which "All data" includes field observations, "Sat data includes Landsat and lower resolution >500m satellite data and "LR satellite data" only includes >500m resolution data). The random forest models have improved after this initial assessment. However, in line with your comment, we also found that in various runs the GBM regressor outperformed the RF regressor.
4). To predict BB EFs, the authors included a series of factors for each group driver (seen in Table 2). However, it seems that some of them are highly correlated, e.g., NDVI VS LAI VS FPAR, VPD VS evaporative stress index VS Relative humidity. The rationale for including these redundant factors may need to be clarified. In addition, given the potential uncertainty in remote sensing and reanalysis data, it may not be wise to include all predictors without doing a feature selection. One way to include a specific driver or not is to compare its effect with a randomly generated variable. If its effect is equal or worse than the random variable, it may not be included in the final training.	We fully agree; many of the features from the full set of features (Table 2) are strongly seasonal and therefore correlated strongly to other features, or even calculated based on one another. The analyses using the full set was mainly to detect which performed better. We then did a feature selection going from a broader set of features (e.g. Fig. 4), to the five features that explained most of the observed variability (Fig. 5, listed in the bottom right box of the panels). The eventual models are trained based on only these features. In the discussion we state (P16 L8): "Cross-correlation between the features meant that feature importance scores (Fig. 4) varied over various model runs based on the test-train data split and bootstrap resampling. For example, a decision tree split based on VPD is most likely very similar to soil moisture or RH, and FTC in national parks is often closely correlated to the MAR, with our measurement sites in Brazil being the notable exception. Although we conducted model runs for various feature-subsets and selected the best, different features may

	also perform well in explaining much of the variability. For features with very high co-variation (e.g., FPAR and LAI or FWI and ISI), this meant only one feature was selected for the trimmed-down model even when both features scored high on the initial assessment." Rather than taking the features with the highest feature scores, we realize that many of the features are correlated and therefore explain the same variability. We took the correlation between features into account when selecting the features for the final models. We added a Pearson correlation matrix (below this table, and as table A2 in the appendix of the manuscript) that lists the correlation between satellite features as well as the direct correlations of satellite features with the target variables.
5). Satellite data over tropical regions usually suffers from the contamination of clouds. When deriving the global EFs, how the authors gap-filled relevant remote sensing data is not clear.	Indeed, particularly in the late dry season we found that cloudiness was a problem, especially for retrievals like NDVI before fire, dNDVI, dNBR, etc. If the scene before or after the fire was cloud-covered, the preceding or successive scene was used with a limit of 14 days before or after the fire. If no cloud-free scene was available in that time window, the fire was removed from the dataset. For the features included in the final models, this was less of an issue given that the meteorological reanalysis data from ERA5-land is not impacted. Fractional tree and non-tree vegetation (MOD44bv006) as well as landcover classification (MCD12Q1C6) are annual while FPAR is based on an 8-day composite meaning the risk of no signal are much lower. When aggregated to 0.25 degree (while using a savanna mask to only take the average of the savanna classified pixels), there were no longer missing values. We added the following sentence to the text (P9 L40): "Further simplification using a subset of features that are not directly correlated, reduced the data dependency

	and computational demands of the model as well as the loss of training data due to cloud cover, without losing much explained variance.”

Table 1: Performance of various regression methods during our initial assessment. This assessment only included our own data. In the variable categories "All data" includes field observations, "Sat data includes Landsat and lower resolution >500m satellite data and "LR satellite data" only includes >500m resolution data.

Method	Variables	MCE		CH_4 EF		N_2O EF	
		RMSE	R^2	RMSE	R^2	RMSE	R^2
Multilinear regressor	All data:	0.03	0.56	1.45	0.66	0.13	0.19
	Sat data:	0.03	0.46	1.36	0.59	0.28	0.18
	LR Sat data:	0.03	0.46	1.43	0.55	0.35	0.16
Decision tree regressor	All data:	0.031	0.70	1.38	0.76	0.05	0.21
	Sat data:	0.023	0.61	1.05	0.71	0.08	0.40
	LR Sat data:	0.022	0.66	1.00	0.70	0.07	0.59
Random forest regressor	**All data:**	**0.026**	**0.80**	**1.16**	**0.87**	**0.04**	**0.47**
	Sat data:	**0.027**	**0.70**	**0.98**	**0.65**	**0.06**	**0.65**
	LR Sat data:	**0.021**	**0.70**	**0.88**	**0.76**	**0.06**	**0.64**
Gradient boosting machine regressor	All data:	0.023	0.78	1.13	0.85	0.04	0.30
	Sat data:	0.028	0.67	1.00	0.67	0.06	0.58
	LR Sat data:	0.021	0.70	0.90	0.75	0.07	0.56
Neural network regressor	All data:	0.035	0.73	1.16	0.84	0.04	0.50
	Sat data:	0.021	0.62	1.09	0.69	0.07	0.59
	LR Sat data:	0.022	0.61	1.05	0.65	0.06	0.60

Table A2. Spearman correlation matrix for the field measurements and the globally available satellite products. Positive correlations are presented in blue while negative correlations are presented in red.