Reply to Reviewer 2

General assessment:

This well-written and timely study evaluates the capacity of six widely used ozone deposition models to simulate stomatal O_3/fluxes across various global land cover types. The manuscript contributes to the Tropospheric Ozone Assessment Report (TOAR-II) community effort by assessing model behaviour under standardised conditions. It also explores both inter-model variability and sensitivity to key drivers. The study is particularly relevant for improving global ozone risk assessments and advancing vegetation impact modelling. The integration of FLUXNET and SynFlux observational constraints is commendable, and the structured multi-experiment framework is a strong point of the manuscript. That said, several aspects require clarification, particularly around the interpretation of model differences, treatment of uncertainties, and consistency in terminology and figures.

We gratefully acknowledge the reviewer's positive feedback and have addressed individual comments carefully. You find our answers highlighted in blue and changes in red below.

Major comments:

Clarity on Model-Observation Agreement:

The evaluation of modelled Gst against SynFlux-derived values is informative, but the conclusions could be more precise. It's difficult to assess which model(s) perform best consistently across sites. A summary table with performance metrics for each site and model would strengthen this section.

Consider providing a visual summary (e.g. radar plot or heatmap) comparing model agreement with observations across all evaluated metrics.

Thanks for this valuable suggestion. We prepared scatter plots of hourly modeled stomatal flux and the respective SynFlux stomatal flux (Fig. S3). Furthermore, we have heatmaps of spearman correlation coefficients ($p<0.05$) between the hourly canopy-level modelled Gst using all available data (including SynFlux) in the supplement information (Figures S4). We also created a table showing the normalized mean bias of each model against Synflux (Table S3).

At both sites, all models perform well with correlation between 0.65 - 0.85 whereas best values are reached by the TEMIR and CMAQ_P.

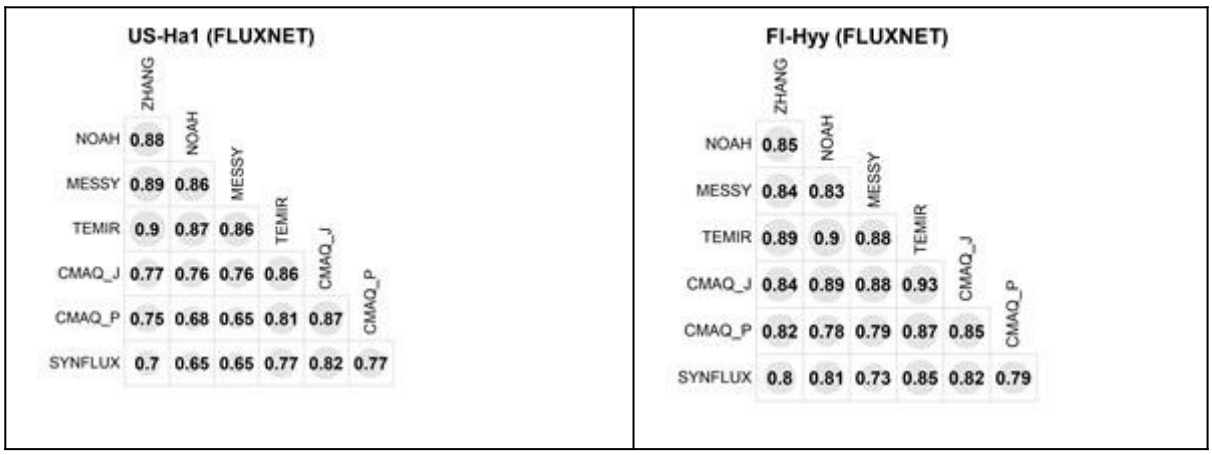
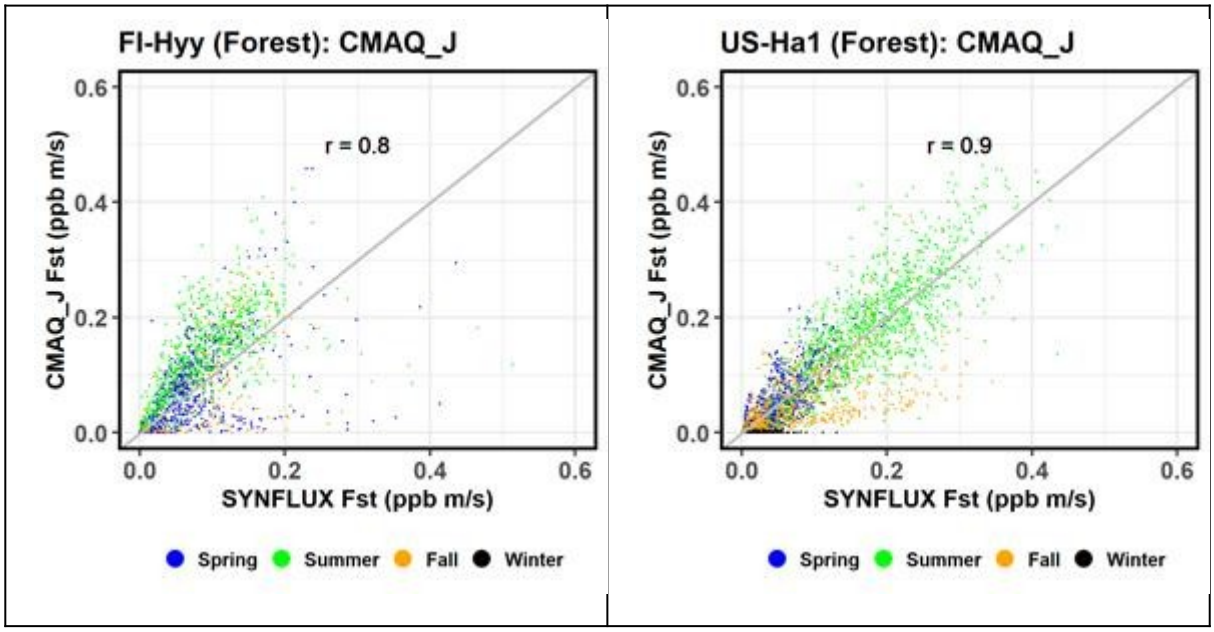
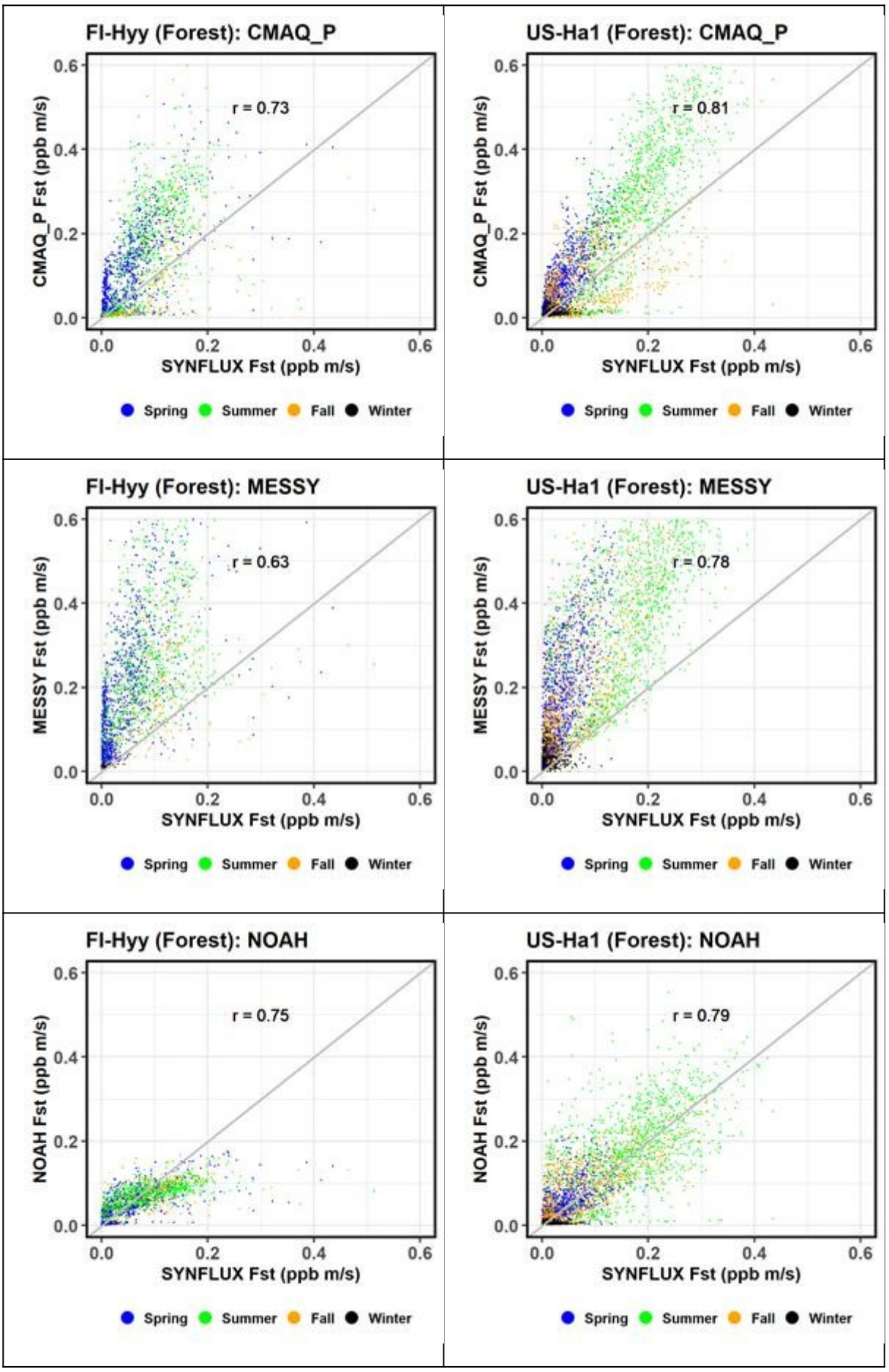


Fig S4. Spearman correlation coefficients ($p<0.05$) between the hourly canopy-level modelled Gst using all available data (including SynFlux). Models were run from the FLUXNET input data.




FI-Hyy (Forest): CMAQ_P
r = 0.73
CMAQ_P Fst (ppb m/s)
SYNFLUX Fst (ppb m/s)
US-Ha1 (Forest): CMAQ_P
r = 0.81
CMAQ_P Fst (ppb m/s)
SYNFLUX Fst (ppb m/s)
FI-Hyy (Forest): MESSY
r = 0.63
MESSY Fst (ppb m/s)
SYNFLUX Fst (ppb m/s)
US-Ha1 (Forest): MESSY
r = 0.78
MESSY Fst (ppb m/s)
SYNFLUX Fst (ppb m/s)
FI-Hyy (Forest): NOAH
r = 0.75
NOAH Fst (ppb m/s)
SYNFLUX Fst (ppb m/s)
US-Ha1 (Forest): NOAH
r = 0.79
NOAH Fst (ppb m/s)
SYNFLUX Fst (ppb m/s)
0.0
0.2
0.4
0.6
Spring
Summer
Fall
Winter

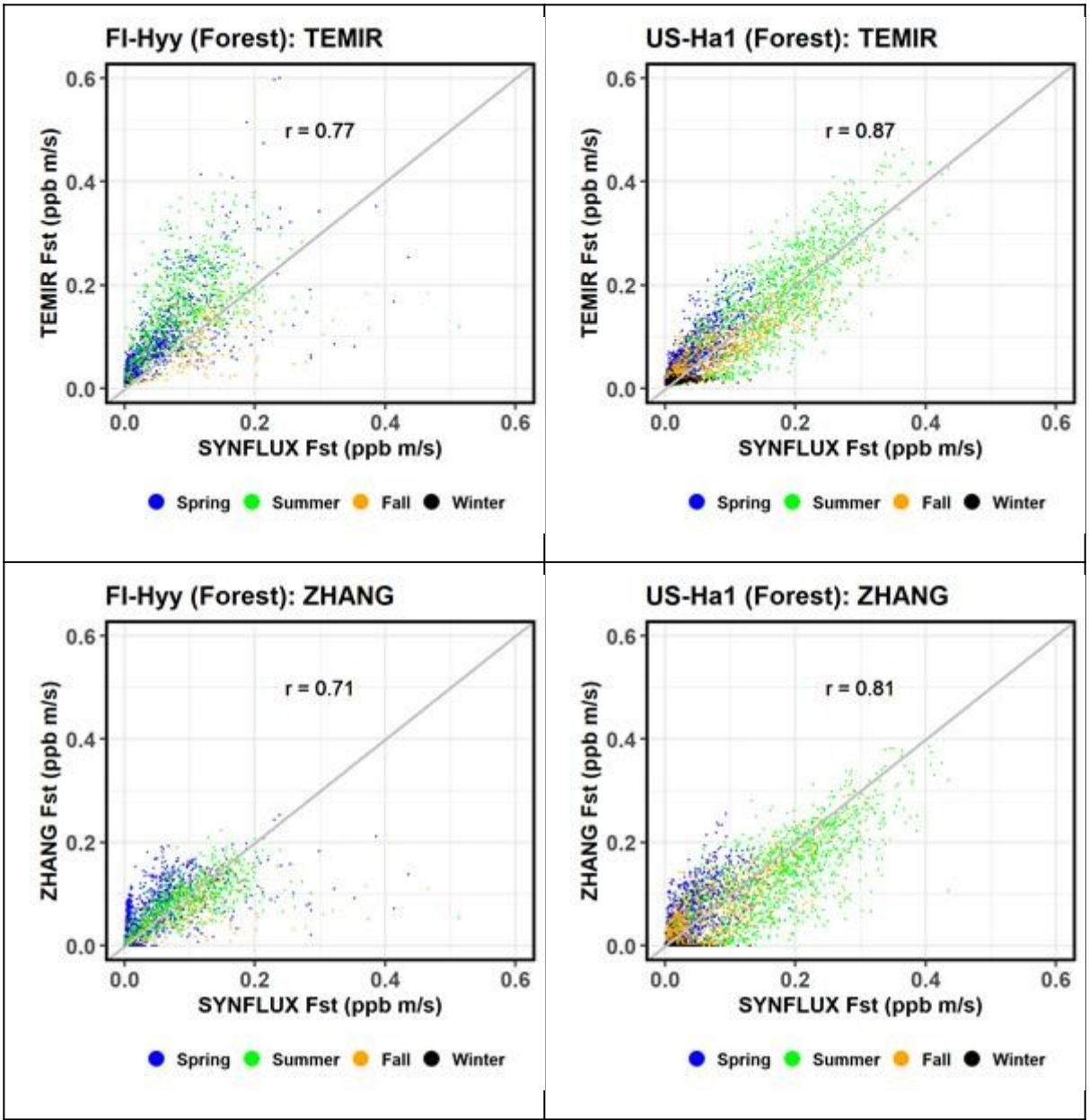


Figure S3. Scatter plots of hourly modeled stomatal flux and the respective SynFlux stomatal flux. 'r' represents Spearman correlations between the hourly fluxes through the entire year ($p < 0.05$).

Table S5. Normalized mean bias (NMB; %) and Spearman correlation coefficient (r) of the model-predicted F_{st} with respect to SynFlux F_{st} . Data for the entire year was used for the calculation.

Site	Model	NMB (%)	r
US-Ha1 (Forest)	CMAQ_J	-7	0.9
	CMAQ_P	54	0.81

	MESSY	**180**	**0.78**
	NOAH	**10**	**0.79**
	TEMIR	**9**	**0.87**
	ZHANG	**-12**	**0.81**
FI-Hyy (Forest)	**CMAQ_J**	**23**	**0.8**
	CMAQ_P	**84**	**0.73**
	MESSY	**222**	**0.63**
	NOAH	**-14**	**0.75**
	TEMIR	**60**	**0.77**
	ZHANG	**4**	**0.71**

Treatment of Uncertainty:

While uncertainty is addressed via sensitivity experiments and ensemble medians, explicit ranges or confidence intervals for key outputs (e.g., PODy estimates) across models would be useful.

Given this variability, how robust are the conclusions regarding PODy differences across land cover types?

Within the scope of the current study we were unable to perform a full sensitivity analysis on the PODy model outputs since this would have taken substantial time and computational power. Also, given the lack of knowledge over the probability distribution of key model parameters for the land cover types explored it would also have suggested a level of knowledge exceeding what we actually have. Therefore, we used the simplified sensitivity assessment to assess which were the key input variables and model parameters that would warrant further study. However, we feel that the PODy differences simulated between land cover types are relatively robust as we do have good evidence to show that key input data (e.g. growing seasons) and variables (e.g. gmax for multiplicative models and Vcmax for photosynthetic based models) are different between land cover groups which will drive broad differences in PODy values. Given the complexity of the deposition models (especially the photosynthesis-based ones), robust confidence intervals could only be computed by

Monte Carlo simulations, which is too computationally expensive since this requires running the models a few thousand times. Also, knowledge over the probability distributions of key model parameters would be required.

PODy Thresholds and Flux-Response Relationships:

The thresholds used for PODy calculation (e.g., 1 nmol m^2s^1 for forests) are stated clearly, but are any species-specific or site-specific adaptations made? The text could benefit from a brief reflection on the limitations of using fixed thresholds across diverse vegetation.

The threshold y is the detoxification capacity, the chosen values are commonly used for crops, forests or grassland (Emberson 2020). In fact, we selected the 6 sites based on their land cover types and we applied the known threshold for the individual sites based on Emberson (2020). We add the following explanation in the respective text: "Studies (Emberson 2020 and references therein) have established thresholds for different land cover types which are used to provide y values for the selected sites with specific land cover types in this study. Some studies suggest that the y threshold for land cover types may vary by global region (e.g. a number of studies suggest higher y values of up to 12 nmol O_3 m^2 s^{-1} is more appropriate for crops and forest tree species in Asia). In this study, which focuses on comparing across models, we maintain consistency and use common y threshold values for each landcover type. However, this is an aspect that would benefit from further study in the future since estimating PODy values with higher thresholds is more challenging for all types of model given the less frequent occurrences of such high O_3 doses. "

Figures and Data Presentation:

Figures 3–6 are central to the conclusions, but they are visually dense due to the number of sites and models. Consider moving some detailed seasonal panels to the Supplement and simplifying the main figures.

We moved the panels showing winter, spring and autumn in Fig. 3 and 4 to the supplement.

Minor Comments:

Ensure consistent use of chemical notation: Use subscript formatting (e.g., O_3, CO_2) where possible. Standardize units throughout the text and figures (e.g., "mmol O_3 m^{-2}" vs "mmol O3 m-2", "cm s^{-1}" vs "cm/s").

We harmonized the units according to Copernicus standards.

In multiple places, "sunlit" is referred to (e.g., Fst, sun, Gsun). Define these variables clearly in the main text, not just in figure captions or formulas.

The definitions can be found in line 332/333 (‘This also helps interpret the modelled stomatal conductance of sunlit leaves (G_{sun}) shown in Fig 4.’) and line 395-397 (‘Figures 5 and 6 show the (SRAD>50 Wm^{-2}) stomatal O_3 flux (F_{st}) and stomatal, sunlit O_3 flux ($F_{st,sun}$) for different models per season at 9 sites representing forest (top), grass (middle), crops (bottom).’

Figures 3, 4: Increase font size in legends and axes for readability.

Done.

Figure 7–9: Consider sorting or grouping sites by land cover or latitude for more straightforward interpretation.

We understand that all figures are dense due to the multiple models used. To make interpretation easier, the same land cover type is displayed in one row.

Table 4: Clarify whether VCmax values refer to standardised temperature conditions (25°C). Also, state if values are per sunlit leaf area or total canopy.

Yes, VCmax refers to 25°C and to the total canopy. We now state that in the table caption.

Avoid overly long sentences, e.g. lines 66–68, which span several embedded clauses. Break these into two sentences for readability.

For this particular sentence (lines 66-68), we don’t see a readability issue. Per the reviewer's suggestion, however, we have reviewed the entire manuscript and revised those long sentences to improve readability.