CARBON ATMOSPHERIC TRACER RESEARCH TO IMPROVE NUMERICAL SCHEMES AND EVALUATION

CATRINE

Carbon Atmospheric Tracer Research to Improve Numerics and Evaluation

D8.1 Development of tracer transport metrics

Due date of deliverable	30th April 2026
Submission date	14th May 2026
File Name	CATRINE-D8.1-V4
Work Package /Task	D8.1
Organisation Responsible of Deliverable	ECMWF
Author name(s)	Anna Agusti-Panareda (ECMWF), Frédéric Chevallier , Chiranjit Das (LSCE), Stefan Versick, Achraf Qor-El-Aine (KIT), Joram Hooghiem, Maarten Krol, Wouter Peters (WU)
Revision number	4
Status	ISSUED
Dissemination Level / location	PUBLIC

Funded by the European Union

The CATRINE project (grant agreement No 101135000) is funded by the European Union.
Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the Commission. Neither the European Union nor the granting authority can be held responsible for them.

1 Executive Summary

This deliverable presents the work done in CATRINE towards developing common metrics to evaluate the accuracy of global transport models. It is structured in two parts.

The first part evaluates the skill of four global simulations of CO_2, respectively based on chemical transport models ICON-ART, IFS, LMDZ, and TM5, from the near-surface layer to upper troposphere and lower troposphere regions (UTLS), using both surface and aircraft-based measurements. This evaluation is performed within the framework of the CATRINE TransCom intercomparison, where all participating model simulations used the same prescribed surface fluxes of CO_2 as documented in the previous deliverable D7.1 (Chevallier et al., 2024). Different statistical metrics like Modified Kling-Gupta Efficiency (KGE) and Taylor Skill Score (TSS) are tested as well as a newly developed metric based on the vertical CO_2 difference between adjacent tropospheric layers is used to assess the skill of the model simulations. Results highlighted the fact that some simulations consistently ranked as best-performing across latitude bands and terrain types (coastal, remote marine, and continental), show strong agreement with observations in terms of temporal matching, variability, and mean concentration. However, all model simulations show a systematic underestimation in UTLS CO_2 with the highest disagreement in June-July-August (JJA). Improving transport representation around the UTLS, especially in this period, appears to be critical for reducing transport errors in atmospheric CO_2 inversion systems.

In the second part of the deliverable, a prototype for a scorecard is developed that combines the most relevant metrics for the evaluation of systematic errors in the large-scale 3D distribution of atmospheric tracers, because the seasonal continental model biases have a direct impact on the capability of global inversion systems to provide accurate estimates for the global carbon budget. The core global model in the Copernicus greenhouse gas emission monitoring service – known as the CO2MVS – (IFS) is used to test the application of seasonal and annual global scorecards for CO2 and SF6 simulations using the CATRINE TransCom protocol. The scorecard shows the impact of large-scale seasonal biases in the UTLS and at high latitudes on the bias of the atmospheric column, which has implications for the assimilation of satellite data in the CO2MVS. The scorecard is also used to provide an assessment of the impact of systematic errors introduced by coarse horizontal resolution in the model.

The results emphasize the importance of model resolution on the representation of the large-scale 3D structure of atmospheric tracers, by enhancing both vertical transport from the atmospheric boundary layer to the UTLS and the signal of the surface flux errors in the atmosphere. The scorecard also reveals the importance of increasing the vertical and global coverage in the operational observing system of CO2 and other complementary tracers, which is currently not available.

We anticipate that the metrics and scorecards presented in this deliverable will support the evaluation of the numerical schemes of the global transport models used in the Copernicus Atmosphere Monitoring Service and provide clear guidelines and recommendations for further development of the CO2MVS capacity.

Table of Contents

Contents

2 Introduction

2.1 Background

To support EU countries in achieving the targets, the EU and European Commission (EC) recognise the need to support establishing the new European anthropogenic CO_2 emissions Monitoring and Verification Support capacity (CO2MVS). To support the Commission and the CO2 Task Force with designing and ultimately building the CO2MVS, previous projects have been funded such as: the CO_2 Human Emissions (CHE) project, the CoCO2 project (https://coco2-project.eu/) and recommendations from the VERIFY project (https://verify.lsce.ipsl.fr/). However, some of the recommendations from the CHE project were not available yet at the time of the definition of the CoCO2 project and could therefore not be fully considered.

The Carbon Atmospheric Tracer Research to Improve Numerical schemes and Evaluation (CATRINE) project aims to evaluate and improve the numerical schemes for tracer transport in the new Copernicus anthropogenic CO_2 emissions Monitoring and Verification Support capacity (CO2MVS) and more widely in the Copernicus Atmosphere Monitoring Service (CAMS). The research and development activities in CATRINE will focus on the priorities identified by these previous activities. The CATRINE project will contribute to the further development of the new Copernicus element for the monitoring of anthropogenic CO_2 and CH_4 emissions and sinks.

The main objectives of CATRINE are to improve the methods used to represent resolved tracer transport by the winds, with a particular focus on mass conservation, and to identify other systematic errors associated with unresolved processes represented by parametrizations. The project will define protocols for evaluating tracer transport models at both global and local scales. Test beds that integrate observations from field campaigns with model results will be developed, along with suitable metrics for tracer transport evaluation, utilising a range of tracers and observations at both global and local scales. These metrics will be employed in the operational CO2MVS to evaluate the implementation of new transport model developments, characterise transport accuracy and representativity in data assimilation, and provide a quality control stamp of tracer transport accuracy. Lastly, CATRINE will provide clear recommendations to the CO2MVS and the Carbon Cycle Community that employs atmospheric inversion models for the evaluation and quality assessment of tracer transport models.

2.2 Scope of this deliverable

2.2.1 Objectives of this deliverable

The main objective of this deliverable is to develop common metrics allowing comparisons between models and the evaluation of different model versions. These metrics will support the evaluation of the numerical schemes of the CO2MVS capacity systems from global to regional scales. The development of transport metrics focuses on the detection of systematic errors using a variety of tracers.

2.2.2 Work performed in this deliverable

The accuracy of global transport models has been assessed using a multi-model multi-tracer approach. Four different global chemical transport model simulations have been evaluated by using the CATRINE TransCom protocol developed in WP7. This evaluation focuses on the main tracer, CO_2, by comparing against in-situ observations. Two different metric evaluation approaches are used combining correlation, mean bias, and variability metrics (Taylor Skill

Score and modified Kling-Gupta Efficiency) to systematically rank and compare TransCom simulations. A multi-layer vertical CO_2 difference analysis methodology was also developed to assess transport skill from the boundary layer to the upper troposphere and lower stratosphere, which is less sensitive to prescribed surface fluxes.

Furthermore, a prototype for a seasonal and annual scorecard has also been designed to provide a structured framework for systematically evaluating the model biases across regions and vertical layers using a set of simple metrics that are easy to interpret. These scorecards are particularly useful to assess changes in models and therefore can provide efficient feedback for model development.

*[Editorial note: **bold** text is used in some sections to highlight item/ theme importance.]*

2.2.3 Deviations and counter measures

N/A

2.3 Project partners

Partners	
EUROPEAN CENTRE FOR MEDIUM-RANGE WEATHER FORECASTS	ECMWF
COMMISSARIAT A L ENERGIE ATOMIQUE ET AUX ENERGIES ALTERNATIVES	CEA
METEO-FRANCE	METEO-FRANCE
WAGENINGEN UNIVERSITY	WU
KARLSRUHER INSTITUT FÜR TECHNOLOGIE	KIT
HELSINGIN YLIOPISTO	UH
UNIVERSITE DE REIMS CHAMPAGNE-ARDENNE	URCA
ALBERT-LUDWIGS-UNIVERSITAET FREIBURG	UFR

3 Lessons from the CATRINE TransCom model inter-comparison

3.1 TransCom model simulations

Following the experimental CATRINE TransCom protocol (Chevallier et al., 2024), four global chemical transport models (CTM) contributed CO_2 mole fraction simulations. Table 1 provides an overview of the participating CTM simulations, detailing their institute name, common meteorological forcing datasets (ERA-5), different horizontal and vertical resolutions, and vertical coordinate systems. The model simulations are further categorized as offline or online based on whether the CTM generates their own meteorological variables or not. In offline models (LMDZ, and TM5), tracer transport is driven directly by pre-computed meteorological fields from ERA-5 reanalysis dataset. In contrast, online models (ICON-ART, and IFS) use meteorological fields, including horizontal wind components (zonal and meridional) and, in some cases, temperature and humidity, generated interactively by a general circulation model (GCM). These dynamically computed fields are nudged toward ERA-5 reanalysis data on timescales ranging from minutes to hours, thereby ensuring physically consistent and realistic tracer transport throughout the simulation period. All these model simulations provided three hourly temporal resolution output of CO_2 mole fraction from 1 January 2022 at 00:00 UTC until 31 December 2023 at 24:00UTC.

Table 1. *List of CATRINE TransCom model simulations*

Model simulations	Institute	Category	Meteorology	Horizontal resolution	Vertical Resolution
ICON-ART	KIT	Online	ERA-5	80 km	120η
IFS	ECMWF	Online	ERA-5	28 km	137η
LMDZ_deg	LSCE	Offline	ERA-5	1.4° × 0.7°	79η
LMDZ_km	LSCE	Offline	ERA-5	90 km	79η
LMDZ_era5	LSCE	Offline	ERA-5	90 km	137η
TM5	WU	Offline	ERA-5	1° × 1°	68η

* η represents the hybrid sigma-pressure vertical coordinate system

We have further checked the runtime differences across the CTM simulations, which we requested from the participants through a questionnaire (Figure 1). The results show a difference in runtime between the CTM simulations, ranging from roughly 0.05 hours (LMDZ-km) to over 2 hours (IFS) for a 31-day simulation. The LMDZ family is among the fastest, taking around ~ 3 to 6 minutes, due to the use of the advanced Graphics Processing Units (GPUs) optimised parallel computing facility (Chevallier et al. 2022). ICON-ART stands in an intermediate range, with runtimes of approximately 16 minutes using MPI-based domain decomposition with Central Processing Units (CPU) cores. Differences in runtime can be attributed to varying spatial resolutions, which affect not only the number of grid cells over which calculations are performed but also the model's time step size, in order to ensure the stability of the simulation as resolution increases. This figure has important implications for

correlating the computational investment with the scientific return from model simulations. To do that, we have evaluated the skill of all CTM simulations by comparing their simulated mole fraction of CO_2 with observed mole fraction of CO_2 mole from surface-based monitoring stations and aircraft campaigns.

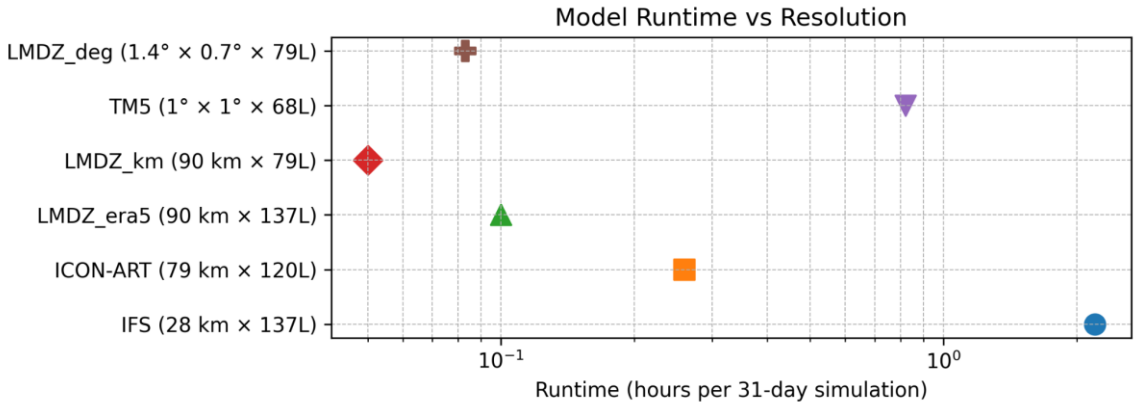

Figure 1: *Model simulations run-time comparison across global chemical transport models (CTMs) along with their respective resolutions shown inside the parenthesis (number with L specifies the number of vertical levels). The x-axis represents the wall-clock runtime in hours required to complete a 31-day simulation, shown on a logarithmic scale for visual simplicity.*

3.2 Observations for evaluation

We used dry air mole fractions of CO2 from 65 surface monitoring sites from the obspack_co2_1_GLOBALVIEWplus_v10.1_2024-11-13 data product distributed by NOAA ESRL, which is on the x2019 calibration scale (Schuldt et al., 2025). Details about the individual monitoring site code and dataset name can be found in table A1 in appendix A. Figure 2(a) shows the spatial distribution of the monitoring sites measuring the mole fraction of CO2 considered for the evaluations. It can be observed that the majority of the monitoring sites are situated in North America and Europe with limited monitoring sites from South America, and South Asia. These sites are chosen based on the condition that all sites have at least one measurement in each month during the period January 2022−December 2023, and the distance between any two sites is at least 1° × 1°, because the lowest resolution of CTM simulations is around 1°. We further classify the sites into four major categories to evaluate performance across different terrain types. These terrain categories are continental (primarily influenced by land fluxes), coastal (influenced by both land and ocean fluxes), remote or marine (dominated by ocean fluxes) and mountains (continental sites located at high elevation).

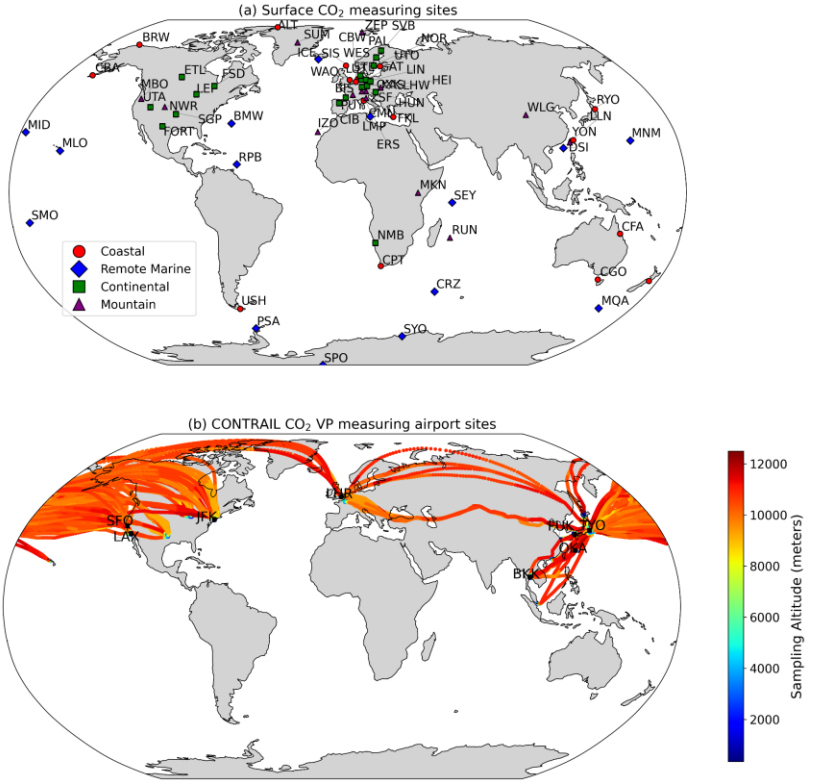

Figure 2: *(a) Spatial distribution of surface based CO_2 monitoring sites used in this study. These sites are categorised as coastal (red circles), remote marine (blue diamonds), continental (green squares), and mountain (purple triangles). (b) CONTRAIL CO_2 sampling location with height information delineated with color maps and also highlighted vertical profile (VP) of CO_2 measuring eight airport sites (BKK, FUK, TYO, OKA, LHR, JFK, LAX, SFO) of the northern hemisphere with black boxes.*

We further use the vertical profile of CO_2 measured around different airports across the globe as part of the Comprehensive Observation Network for TRace gases by AIrLiner (CONTRAIL) aircraft programme to evaluate the CTM simulations from the near-surface to the upper troposphere and lower stratosphere (Machida et al., 2008; Matsueda et al., 2008). These vertical profiles of CO_2 are collected during the aircraft's ascent and descent near airports. We selected eight northern hemispheric airports (BKK, FUK, TYO, OKA, LHR, JFK, LAX, SFO), based on the condition that near all these airports, the aircraft have taken at least 2 months of vertical CO_2 profile in each season for 2022-2023 (Figure 2b). On the other hand, we have selected all CO_2 measurements from the upper troposphere and lower stratosphere (above 8 km) for skill evaluation in UTLS regions. The CONTRAIL based CO_2 dataset is also taken from the obspack_co2_1_GLOBALVIEWplus_v10.1_2024-11-13 data product (https://gml.noaa.gov/ccgg/obspack/data.php).

3.3 Evaluation metrics

The focus of the study is to evaluate the skill of the TransCom simulations based on their performance evaluated against the in-situ observations. For this purpose, we use correlation (r) to evaluate the timing of tracer variability, mean (μ) for systematic error, and standard deviation (σ) for variability error as core metrics, which are then combined into a single statistical metric that can integrate different statistical aspects needed to rank the simulations. We used two independent metrics, the Taylor Skill Score (TSS) and the modified Kling-Gupta Efficiency (KGE), for evaluation, which combine the above-mentioned statistical aspects

(Gupta et al., 2009; Taylor, 2001). Before doing that, 3-hourly TransCom simulations of dry air mole fractions of CO_2 are sampled to the nearest grid point of the sampling time, and locations of surface and aircraft-based CO_2 measurements. Then, the simulations are evaluated directly at the grid-point scale following the aforementioned metric approach.

Pearson's correlation is calculated for observed (o) and modelled (m) time series as follows,

$$r = \frac{\sum(o_i - \bar{o})(m_i - \bar{m})}{\sqrt{\sum(o_i - \bar{o})^2 \sum(m_i - \bar{m})^2}}$$

Mean of model and observation (μ)

$$\mu_m = \frac{1}{N}\sum_{i=1}^{N}(m_i) \text{ or } \mu_o = \frac{1}{N}\sum_{i=1}^{N}(o_i)$$

Taylor skill score (TSS) which combines correlation ($r_{m,o}$), and variability ($\sigma_m \, or \, \sigma_o$) into a single metric, where $r_o = 1$ represents the maximum attainable correlation,

$$\mathrm{TSS} = \frac{4\,(1+r_{m,o})}{\left[\left(\frac{\sigma_m}{\sigma_o}+\frac{\sigma_o}{\sigma_m}\right)^2 (1+r_0)\right]}$$

Another metric for the evaluation of the skill of model simulations is Modified KGE (KGEm), where the first, second and third term under the square root represents the correlation, mean and variability.

$$KGE_m = 1 - \sqrt{(r_{m,o}-1)^2 + \left(\frac{\mu_m}{\mu_o}-1\right)^2 + \left(\frac{\sigma_m/\mu_m}{\sigma_o/\mu_o}-1\right)^2}$$

Both TSS and KGEm are formulated such that a perfect model simulation yields a value of 1. Accordingly, model simulations are ranked based on their metric value close to this ideal value, with the simulation closest to 1 assigned rank 1 and lower ranking assigned to simulations with progressively lower metric values.

For the model evaluation using the CONTRAIL vertical profile of CO_2, we have used the vertical CO_2 difference between the adjacent tropospheric layers. Here, four vertical layers are considered: boundary layer, or BL (near surface to 2 km), lower free troposphere, or lower FT (2 to 5 km), upper free troposphere, or upper FT (5 to 8 km), and upper troposphere and lower stratosphere, or UTLS (above 8 km).Vertical CO_2 difference between any two tropospheric layers is calculated as the difference in CO_2 mole fraction between the bottom layer and the immediate top layer. Since the evaluation is based on the difference in CO_2 between layers rather than the absolute concentration, it is less sensitive to errors in the prescribed surface fluxes and provides an understanding of the differences in vertical transport. This approach is previously used in the model evaluation (Stephens et al., 2007). However, we are limited here to airport sites, with all aircraft emissions assigned to the surface in the inventory, which may affect the simulated vertical gradient. These sites may therefore not be representative of the model quality in the boundary layer.

Our approach is written mathematically in the following manner,

$$\text{vertical } CO_2 \text{ difference } (\Delta CO_2) = CO_{2,\text{bottom}} - CO_{2,\text{top}}$$

3.4 Integrated performance assessment of TransCom model simulations

3.4.1 Performance assessment based on surface-based CO_2 observations

To evaluate the performance of the transport models' simulations at the global scale, we have divided the global latitudes into six latitudinal bands of 30 degrees (90°N-60°N, 60°N-30°N, 30°N-0, 0-30°S, 30°S-60°S, 60°S-90°S). All monitoring sites within each latitude band are used to evaluate the CTM simulations of CO_2 mole fraction, which are sampled at those sites. The evaluation is then performed by assessing how well the simulations reproduce the observed timing, mean concentration, and variability considering the entire time series, using the aforementioned metrics. We also examined the sensitivity of each component of the metrics to understand how much each term contributes to the overall simulation ranking. The analysis shows that correlation and variability play the biggest role in determining the overall ranking of the simulations. Because we observed that the ranking of the simulation based on KGE and TSS is nearly the same, despite the fact that KGE explicitly includes a systematic error term in its formulation while TSS does not. For this reason, we first present model rankings based on correlation and variability within each 30-degree latitude band (Figure 3). We then use the modified KGE and TSS score to summarise the overall model skill for each latitude band while accounting for both correlation and variability.

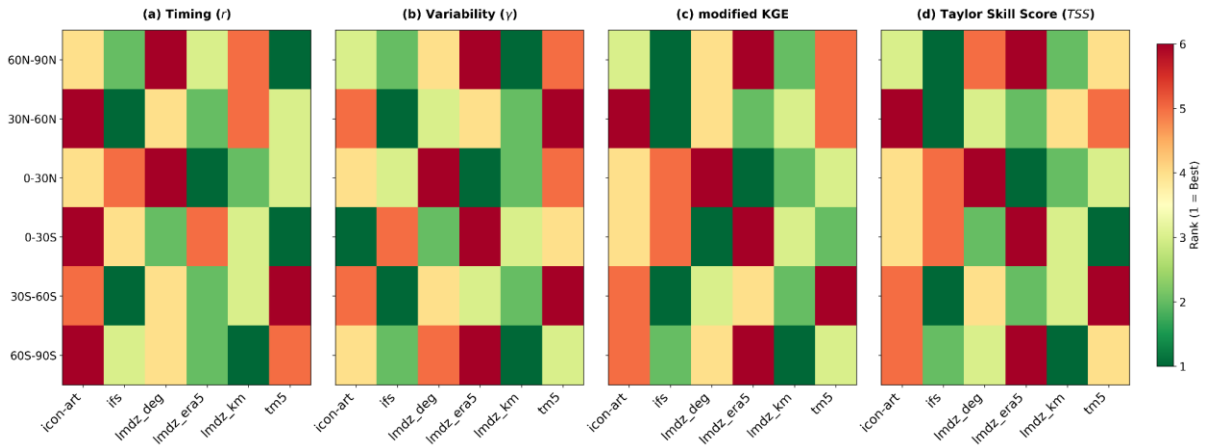

Figure 3: *Ranking of the CTM simulations based on timing, variability and composite metrics (modified KGE and Taylor skill score) in six latitude bands, from the comparison against in-situ CO_2 observations at surface sites.*

This ranking of the CTM simulations, based on in-situ surface CO_2 measurements, shows a clear and consistent performance pattern across latitude bands and skill metrics. The IFS, and LMDZ_km simulations received the best ranks (green, rank 1–2) across most latitude bands and metrics, indicating that these models have reproduced the timing, variability, and overall magnitude of surface CO_2 as captured by the in-situ network (Figure 3c,d). In contrast, the ICON-ART simulation shows relatively low performance in both timing and variability in all latitude bands as compared to other model simulations (Figure 3a,b). The TM5 simulation performed well in capturing the timing of variation, especially in the northern hemisphere and

southern tropics, but it did not show good performance in capturing the observed variability compared to other simulations.

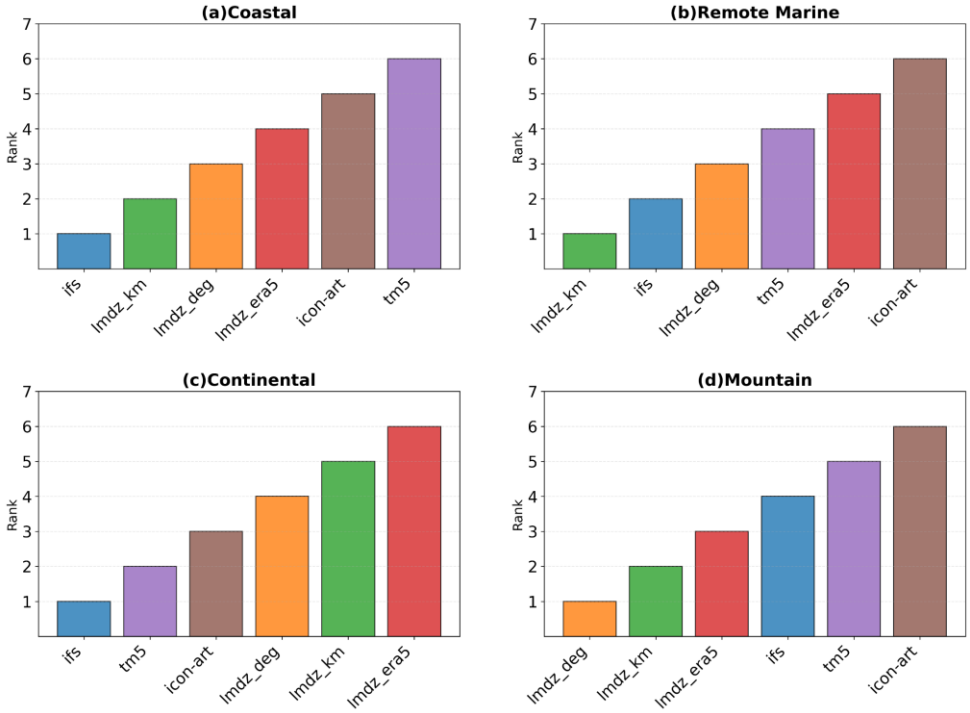

Figure 4: *KGEm score-based ranking of the CTM simulations in different station types, coastal (panel a), remote marine (panel b), continental (panel c), and mountain (panel d).*

We further checked the skill of the CTM simulations based on the KGE metric evaluated using in-situ CO_2 measurements across four site categories, namely coastal, remote marine, continental, and mountain (Figure 4). IFS and LMDZ_km model simulations consistently emerge as best across coastal and remote marine categories. In contrast, ICON-ART, TM5, and LMDZ_ERA5 model simulations showed lower skill over coastal and remote marine terrain. In continental sites, IFS and TM5 simulations performed well, likely linked to their ability to reproduce the diurnal and synoptic transport over land. Further, we noted that the high-resolution model simulation (IFS) performed better than the low-resolution simulations, which likely suggests a better representation of the station environment at higher resolution (Agustí-Panareda et al., 2019). The relatively lower ranking of high resolution models such as IFS at mountain sites may partly reflect a representativity issue arising from the coarse resolution ~1°× 1° of the prescribed surface fluxes. Since high-resolution models resolve mountainous terrain at a finer scale, they tend to place mountain stations at higher elevations than those represented in prescribed surface fluxes. This causes the inconsistency between the model orography and the flux representation at these sites. This will be further explored and quantified in the next phase of the project, with the introduction of a higher-resolution (9 km) IFS simulation into the evaluation.

3.4.2 Performance assessment based on aircraft based CO_2 observations

Since the aforementioned CTM simulations are run with only one set of prescribed surface fluxes, errors in these surface fluxes may accidentally cancel some transport model errors and artificially enhance the corresponding simulation rating. To address this limitation and obtain a more robust assessment based on transport, we extended our evaluation with CONTRAIL aircraft CO_2 measurements, which sample the near-surface to UTLS regions. In this section, we first evaluate model performance using CO_2 differences between tropospheric layers and in the UTLS region. Next, we assess performance across different latitude bands in the UTLS. Finally, we examine the simulation spread, which allows for a more transport-focused assessment that is less sensitive to prescribed surface fluxes.

We have checked the CTM's simulation skill based on vertical CO_2 differences (ΔCO_2) across three atmospheric layers: BL to lower FT, lower FT to upper FT, and upper FT to the UTLS layer. This analysis is conducted using the vertical CO_2 profile from 8 airport sites as mentioned in section 3.2. The dataset is mainly from the Northern Hemisphere mid-latitudes and East Asia. The vertical CO_2 difference across three atmospheric layers reveals distinct CO_2 mixing strength in the vertical direction as well as layer-dependent horizontal transport in the simulations. In the next part of the project, we will also evaluate two different horizontal resolutions of IFS simulations to better understand the sensitivity of mixing strength to resolution.

In the BL to lower FT transition (Figure 5a), all models consistently show positive ΔCO_2 values in all seasons, with the highest inter-model spread in DJF, likely due to the combined effect of stable winter boundary layer conditions and high fossil fuel emissions at these urban airport sites, where shallow mixing traps near-surface CO_2. The observed-ΔCO_2 gradient (grey bar) is generally reproduced by the CTM simulations in this layer, though individual simulations deviate substantially. IFS-ΔCO_2 (steel blue) shows the closest to observed-ΔCO_2 in the winter months. However, ICON-ART, and LMDZ_ERA5 model simulations systematically overestimate ΔCO_2 values in all seasons, suggesting either insufficient vertical mixing in the lower troposphere in these model simulations and/or an error associated with the local emissions at the airports which is enhanced at higher resolution. The JJA season consistently shows the smallest BL–lower FT gradients across all model simulations, likely due to CO_2 removal by active biospheric activity, and deep boundary layer development in summer. Here, we observed that IFS and ICON-ART have the highest ΔCO_2, which could be linked to a weaker vertical mixing strength in these model simulations during JJA. However, it is important to note that the BL-lower FT gradient can also be affected by errors in the local emissions at the airports, which will be enhanced with higher horizontal and vertical resolution in the IFS and ICON-ART simulations. One clue pointing at the effect of local anthropogenic emissions is the relatively large positive gradient during JJA, whereas we would expect negative gradients over large-scale continental regions. In the lower FT– upper FT layer, ΔCO_2 is considerably reduced, within 2 to 3 ppm, which is close to the observed ΔCO_2 likely associated with well-represented large-scale dynamical mixing in the free troposphere and because of the reduced influence by the local surface fluxes at these altitudes. Inter-model agreement improves markedly in the lower FT– upper FT layer, reflecting the well-resolved synoptic-scale transport in the CTM simulations. In JJA, we observed that ΔCO_2 is close to zero for the majority of the model simulations, matching observed ΔCO_2. However, ICON-ART shows a strong positive ΔCO_2 likely linked to misrepresentation of vertical mixing or horizontal transport at this level. Model matches are strongest in the upper FT and UTLS layer, suggesting that large-scale dynamical transport is reasonably captured across model simulations. Furthermore, we observed a negative ΔCO_2 in JJA, which is reproduced in all model simulations except in the ICON-ART simulation.

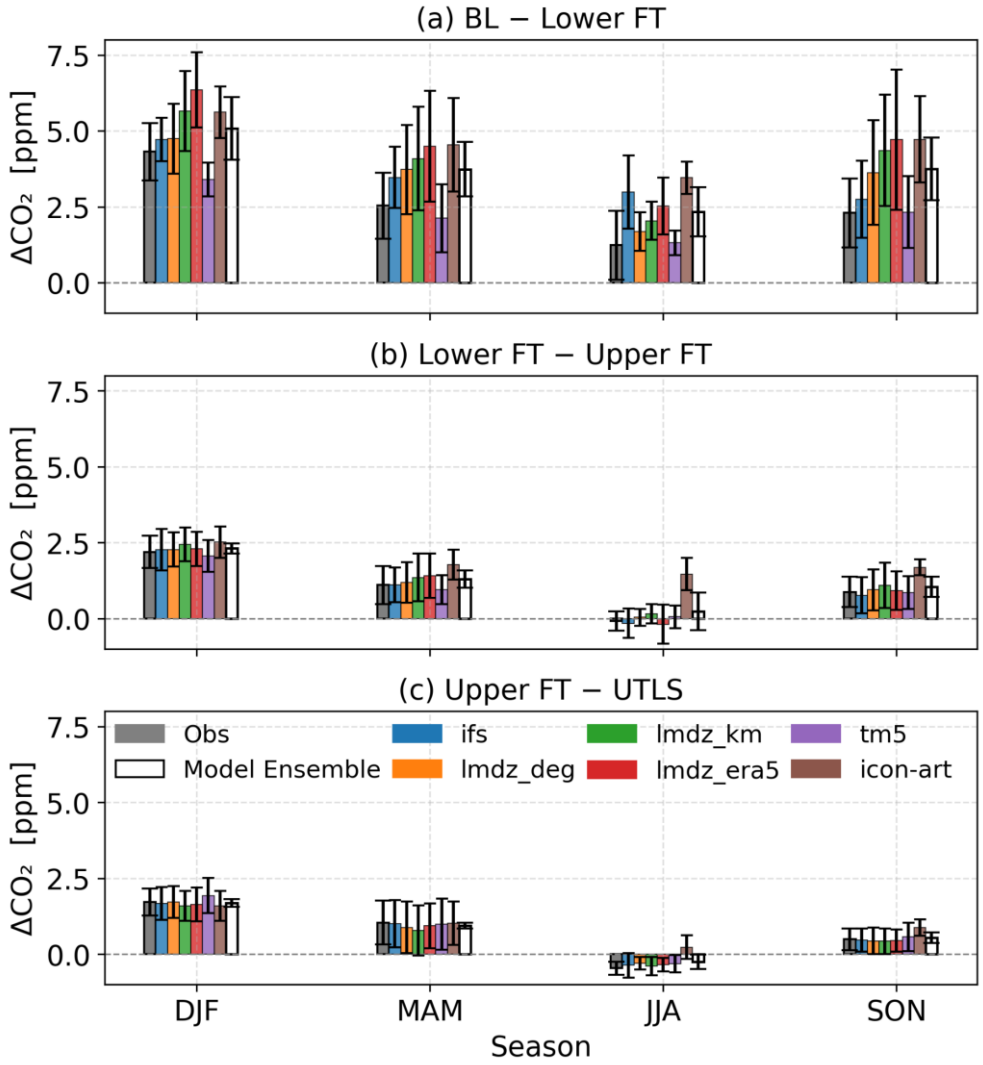

Figure 5: *Seasonal CO_2 difference between the tropospheric layers BL to Lower FT (panel-a), Lower FT to Upper FT (panel-b), and Upper FT to UTLS (panel-c) using the CONTRAIL aircraft CO_2 measurements during both day and night hours near the 8 airports (BKK, FUK, TYO, OKA, LHR, JFK, LAX, SFO) in the northern hemisphere.* The BL ranges from the near-surface to 2 km, the lower FT from 2 to 5 km, the upper FT from 5 to 8 km, and the UTLS above 8 km.

In summary, the differences in CO_2 are likely linked to the inter-model differences in horizontal advection and vertical mixing parametrisations, as well as differences in how the various model simulations represent local emissions and BL mixing at airports. The way in which different modelled processes impact tracer transport is further explored in D6.1. Though part of this disagreement may be linked to the representativeness mismatch between point urban measurements and the coarse model grid. This part will be investigated in the next phase of the project, where a high-resolution 9km IFS simulation will be compared with the 28km IFS simulation to answer these questions. In general, we have seen that the high-resolution IFS performs better than its closest-resolution ICON-ART simulation. This is also tested in the scorecard-based analysis to assess the sensitivity of tracer transport to IFS resolution. Furthermore, this multi-layer seasonal analysis shows that the influence of local surface fluxes is predominant in the near surface which limits a pure transport assessment. Therefore, improving representation near the surface should be a critical priority for reducing errors in atmospheric CO_2 inversions. More aircraft campaigns with extended vertical CO_2 profile

measurements are needed across different environmental settings (away from emission hotspots such as airports) and seasons to support model development at different scales.

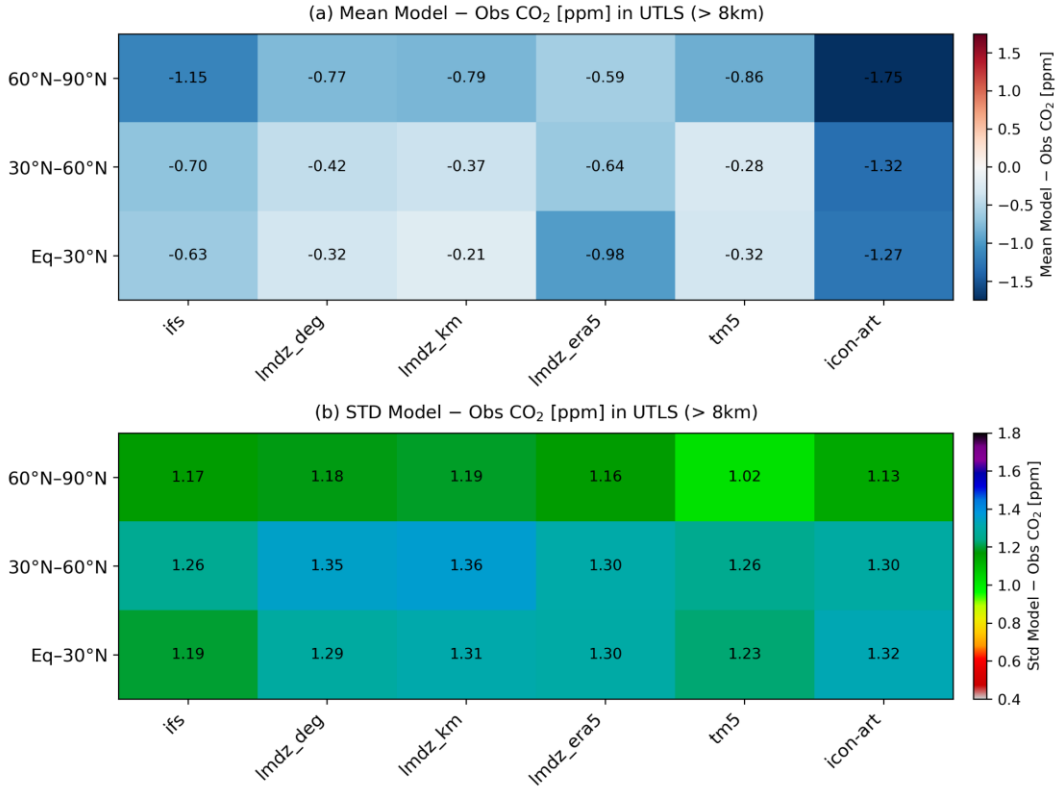

Figure 6: *Mean (panel a) and standard deviation (panel b) in the model and observation difference in CO_2 at the upper troposphere and lower tropospheric layers based on CONTRAIL CO_2 measurements.*

We checked the model simulations performance of CO2 in the UTLS (above 8 km), evaluated against CONTRAIL aircraft measurements across three northern hemispheric latitude bands, tropical (Equator-30°N), mid latitude (30°N-60°N), and polar (60°N–90°N) (Figure 6). The CO2 mean bias in the UTLS shows a systematic negative bias (model simulation underestimation) across nearly all model simulations and latitude bands, indicating that models consistently fail to transport sufficient CO2 into the UTLS. The negative bias is most pronounced in the 60°N-90°N band, where model simulations like ICON-ART show biases of −1.75 ppm. We have also examined the UTLS negative bias using IFS simulations in the scorecard-based analysis in the following section. The standard deviation (STD) of the CO2 error shows that all model simulations have similar random errors, with values ranging from approximately 1.1 to 1.3 ppm. TM5 shows the lowest STD values in polar latitude bands, whereas IFS shows the lowest mid and equatorial latitude bands. The different magnitude of mean model-observation CO2 differences in CTM indicates that their biases are quite different across latitude, likely linked to misrepresented transport in the UTLS. This is further explored with IFS simulation-based tracer transport analysis in the UTLS.

We further checked the seasonal ensemble standard deviation of model-observation CO_2 differences in the UTLS region (above 8 km) within each 5-degree grid box, using CTM simulations and CONTRAIL observations (Figure 7). This highlights critical regions of disagreement between CTM simulations, which are largely due to transport differences, as the UTLS is far from the surface CO_2 sources and sinks and is therefore less sensitive to prescribed surface fluxes. This has important implications in CO_2 inversion systems because

any discrepancy in transport between the CTM simulation can result in different estimates of the optimised surface CO_2 fluxes. The result reveals the spatial pattern of the CTM simulation spread in different seasons (DJF, MAM, JJA, and SON). Across all seasons, the highest spread is observed in JJA along the mid-latitude in east Asian corridor (30°-70°N), which is a region where Asian emissions from the Asian Summer Monsoon Anticyclone (ASMA) come in higher latitudes and finally, mix with the UTLS via Rossby wave breaking (Vogel et al., 2012). The highest standard deviations suggest that model simulations disagree on the Asian outflow into the Pacific storm track in JJA. Note that this is also the period when the biosphere is most active, so any uncertainties in prescribed surface fluxes may be amplified due to additional uncertainty arising from UTLS transport. In contrast, low standard deviation is observed during the DJF and SON seasons, suggesting better agreement in transport than during JJA. MAM also shows a high standard deviation in the Northwest Pacific, extending toward the North American region.

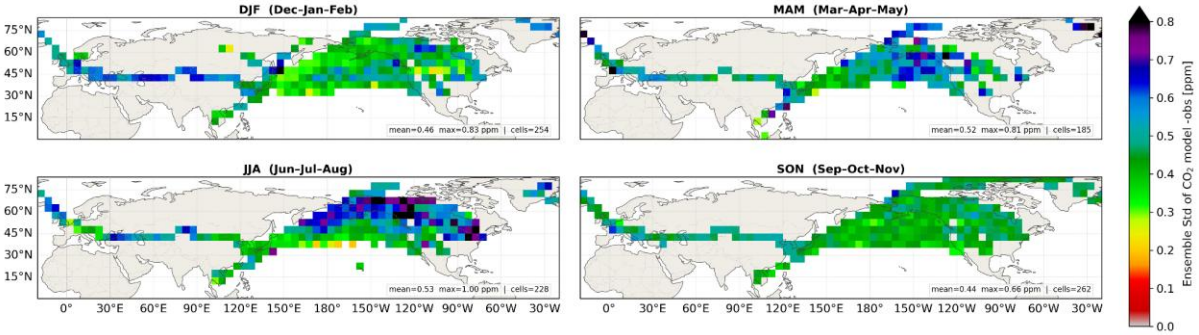

Figure 7: *Seasonal standard deviation of model and observation CO_2 difference in different seasons (DJF, MAM, JJA, SON) using CONTRAIL campaign-based CO_2 measurements in the UTLS (> 8 km) region.*

4 Towards a global tracer transport scorecard

In this section, we present the design of a prototype tracer transport scorecard spanning regional to global scales. Like scorecards in Numerical Weather Prediction (NWP), the tracer transport scorecard should provide a compact and statistically robust visual summary of the transport model performance, enabling rapid assessment and comparison of model performance across the targeted scales. The focus of the tracer transport scorecard presented in this report is on the assessment of the systematic errors at large scales (from continental to global) because these directly translate into surface flux errors relevant to the global carbon budget.

Given the sparseness of observations and the zonal and seasonal consistency of long-lived atmospheric tracers at global scale, the tracer molar fraction has been aggregated with season, 30° latitude bands and 4 vertical layers from the surface to the lower stratosphere (described in section 3.3). This aggregation can be modified and adapted to the observations available, e.g. using monthly timescales instead of seasons. It should also be straightforward to adapt the scorecard to specific continents (e.g. 11 large RECCAP regions). The scorecard should be applied to different tracers in order to disentangle the transport errors from the effect of flux errors in the evaluation metrics. CO_2 and SF_6 are ideal tracers for tracer transport because they are very long-lived inert tracers in the troposphere and stratosphere and they have been widely used to evaluate global transport models, particularly the inter-hemispheric, UTLS and stratospheric transport. Additionally, other tracers such as CH_4, CO and water vapour could also be used as extra tracers. To evaluate the large-scale three-dimensional distribution of systematic errors, the scorecard simultaneously examines horizontal and vertical gradients, as well as the seasonal cycle. Since we are limited to seasonal timescales, we focus on the seasonal amplitude errors. With more observations on monthly timescales, there is also the potential to include a metric for the evaluation of the phase of the seasonal cycle.

With this scorecard prototype we specifically aim to identify and quantify:

1. Where and when the systematic errors are largest (region, vertical layer and season).
2. Discrepancies between near surface and total column systematic errors.
3. Impact of model transport developments on transport accuracy (e.g. model resolution).
4. Implications of the transport biases for flux estimation by using mass budget diagnostics.

4.1 IFS model simulations

For practical purposes the IFS has been used to test and illustrate the global transport scorecard. Different IFS simulations have been conducted to showcase the use of the scorecard for a horizontal resolution sensitivity study on tracer transport model performance (see Table 2). The simulations cover the period 2016-2017, as well as the period of the CATRINE TransCom inter-comparison (2022-2023), in order to capitalize on the wealth of field campaign data during those two periods (see section 3.2). All the simulations use the prescribed fluxes and meteorology from the CATRINE TransCom protocol (Chevallier et al., 2024). Thus, the results from the evaluation of these simulations should also help to explain the transport biases in the TransCom simulations presented in section 2, in particular the sensitivity to the horizontal resolution. For the sake of conciseness, in this report we only show results from the IFS_80km and IFS_28km simulations highlighted in bold in Table 3.1.1. Although in this report we only evaluate CO_2 and SF_6, the IFS simulations include the tracers

specified in the CATRINE TransCom protocol (CO_2, SF_6, ^{222}Rn), as well as the tracers that are part of the CAMS GHG forecast (CH_4, CO and water vapour).

Table 2 *List of IFS numerical experiments used in the tracer transport evaluation*

IFS simulation	Horizontal resolution	Vertical levels	Use of injection height for energy sector
IFS_80km_L60	80km	L60	No
IFS_80km	**80km**	**L137**	**No**
IFS_28km	**28km**	**L137**	**No**
IFS_28km_IH	28km	L137	Yes
IFS_9km_IH	9km	L137	Yes

4.2 Observations for evaluation

In order to evaluate global atmospheric transport models, it is paramount to have a reference based on observations of the 3D atmospheric distribution of the tracers covering the different seasons. A compilation of various airborne observations from aircrafts and AirCore balloons over the period of the IFS simulations can be found in Table 3. Many of these observations come from field campaigns that aim to study specific research topics, such as troposphere-stratosphere exchange, and therefore, they are also used in the CATRINE UTLS scorecards (see D6.1). NOAA aircrafts provide vertical profiles in the troposphere at specific sites, approximately twice a month, mostly in North America. There are also operational commercial aircrafts measuring tracers, such as CONTRAIL and IAGOS, which provide vertical profiles at airports and flight track data between airports across the globe. In this report, we have selected the ATom dataset to design the first prototype, because it provides the largest 3D global coverage for CO_2 and SF_6, as well as all the other tracers in the IFS simulations. It also covers three seasons over the period of simulations described in section 4.1. Figure 8 shows the spatial coverage of the 3 ATom campaigns used for the prototype. In the next phase of the project we will use as many of the other observations listed in Table 3 as possible to improve the zonal mean reference of the global scorecard metrics described in section 4.3. Some of these observations are also used in the UTLS test bed scorecards (D6.1) and for the CATRINE TransCom evaluation in section 3.

Table 3 *List of datasets with vertical profiles of various tracers included in the IFS simulations (2016-2017 and 2022-2023). ATom field campaigns are highlighted in bold because they provide the backbone reference for the evaluation in the scorecard prototype presented in this report.*

Dataset	Period	Species	Reference
IAGOS	Jan-Feb 2022	CO_2, CH_4, CO,q	https://www.iagos.org/iagos-data/, Petzold et al. (2015)
CONTRAIL CME	2016-17, 2022-23	CO_2	https://www.cger.nies.go.jp/contrail/, Machida et al. (2008), Matsueda et al. (2008)
CONTRAIL flask	2016-17, 2022-23	CO_2, CH_4, SF_6	https://www.cger.nies.go.jp/contrail/, Machida et al. (2008), Niwa et al. (2011)

Dataset	Period	Species	Reference
AirCore	2016-17, 2022-23	CO_2, CH_4, CO, q	https://gml.noaa.gov/ccgg/aircore/, https://aircore.aeris-data.fr/, https://space.fmi.fi/2018/07/05/high-altitude-weather-balloons-help-to-discover-greenhouse-gases-in-the-atmosphere/, Baier et al. (2021)
NOAA aircraft	2016-17, 2022-23	CO_2, CH_4, CO, SF_6	Sweeney et al. (2015)
ATom1	Jul-Aug 2016	CO_2, CH_4, CO, SF_6, q	https://earth.gsfc.nasa.gov/acd/campaigns/atom, Wofsy et al. (2018) Elkins, J. W., Hintsa, E. J., and Moore, F. L. (2019)
ATom2	Jan-Feb 2017	CO_2, CH_4, CO, SF_6, q	
ATom3	Sep-Oct 2017	CO_2, CH_4, CO, SF_6, q	
ORCAS	Jan-Mar 2016	CO_2	https://www.eol.ucar.edu/field_projects/orcas, Stephens et al. (2018)
ABove	Apr-Nov 2017	CO_2	https://above.nasa.gov/airborne_2017.html
WISE	Aug-Oct 2017	CO_2, CH_4, CO, SF_6, q	https://halo-research.de/science/previous-missions/wise/
StratoClim	Jul-Sep 2016	CO_2, CH_4, CO, SF_6, q	https://www.stratoclim.org/
ACT-America	Jul-Aug 2016, Jan-Mar 2017, Oct-Nov 2017	CO_2, CH_4, CO, SF_6, q	https://science.larc.nasa.gov/act-america/
DCOTSS	May-Jul 2022	CO_2, CH_4, CO, SF_6, q	https://dcotss.org/
ACCLIP	Jul-Sep 2022	CO_2, SF_6, CO, CH_4, q	https://www.eol.ucar.edu/field_projects/acclip
PHILEAS	Jul-Sep 2023	CO_2, SF_6, CO, CH_4, q	https://www.imkasf.kit.edu/english/4256.php

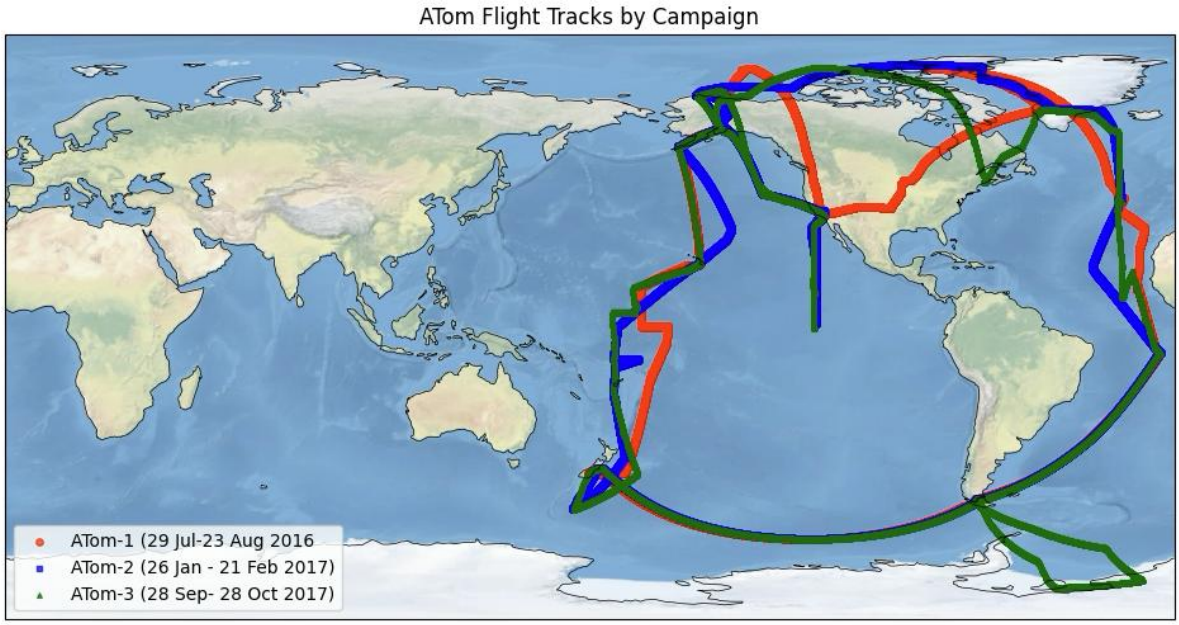

Figure 8: *Map with spatial coverage of the ATom observations used in the global scorecard prototype.*

4.3 Evaluation metrics

The basic evaluation metric used in the scorecard is the systematic error or bias. After sampling the model simulations at the time and location of the ATom observations, the observations and simulations are aggregated in six latitude bands representing high latitudes (60°-90°S and 60°-90°N), mid-latitudes (30°-60°S and 30°-60°N) and tropics (0°-30°N and 30°S-0°); four seasons (DJF, MAM, JJA, SON); and four vertical layers (0-2km as a proxy for the atmospheric boundary layer (PBL), 2-5km representing the lower free troposphere (FT), 5km-tropopause corresponding to the upper FT, and finally, the stratospheric layer above the tropopause). Work has been carried out to test different tropopause definitions based on potential vorticity, temperature, stability, and humidity, in collaboration with the CAMEO project (https://www.cameo-project.eu/). We have found that a hybrid definition combining stability and humidity provides the best results in terms of consistency and smoothness of the tropopause gradients. This new definition has been implemented in IFS CY50R1 and it is used in the scorecard to identify the tropospheric and stratospheric partial columns. The aggregated data from the model (m) and the observations (o) in each latitude (l), height (z) and seasonal (t) bin are then used to compute the regional bias for each layer and season:

$$bias_{l,z,t} = \overline{m(l,z,t) - o(l,z,t)}$$

Additional total column and global annual integrals are computed with a pressure-weighted average of all the vertical layers, and an area weighted average of the latitude bands respectively. These are useful to link the scorecard to the satellite total column observations of CO_2, and the global growth rate computed from the global annual mean values.

In addition to the regional bias, the scorecard also includes an evaluation of the systematic errors in the horizontal and vertical gradients. Horizontal and vertical gradients are computed between adjacent latitude bands and adjacent vertical layers respectively. In order to avoid confusion in the interpretation of the bias associated with the gradients when they change sign (as it occurs for CO_2 in summer), the bias of the model gradient G_m is computed as a percentage of the observed gradient G_O:

$$gradient\ relative\ bias_{l,z,t} = 100 \times \frac{\overline{G_m(l,z,t) - G_o(l,z,t)}}{\overline{G_o(l,z,t)}}$$

A positive relative bias in the gradient indicates the magnitude of the gradient in the model is overestimated and a negative value with a magnitude of 100% or less implies the gradient in the model is too weak. A negative relative bias with a magnitude larger than 100% would reveal that the sign of the gradient in the model is wrong. As relative biases can have very large magnitudes when the observed gradients are very weak, we have included the observed gradient within brackets to support the interpretation of the gradient errors.

Another way to convey systematic errors in the gradients is to compute the standard deviation (σ) of the bias across the different layers ($\sigma_{l,t}$) and latitude bands ($\sigma_{z,t}$). This approach is widely used by the remote sensing community to evaluate satellite retrievals. Here we use the standard deviation of the biases across different vertical layers to provide a bulk error estimate of the vertical gradient for each region:

$$STD_{vert_{l,t}} = \sqrt{\overline{(bias_{l,z,t} - bias_{l,t})^2}} = \sigma_{l,t}(bias_{l,z,t})$$

Where $bias_{l,t}$ is the bias of the total column in each region l and season t. The bulk relative bias of the vertical gradient is computed with respect to the standard deviation of the observed mean molar fraction at each region l and season t:

$$relative\ STD_{vert_{l,t}} = 100 \times \frac{\sigma_{l,t}(bias_{l,z,t})}{\sigma_{l,t}(o_{l,t,z})}$$

and the standard deviation across different latitude bands as a bulk error of the large-scale meridional gradients for each vertical layer:

$$STD_{Lat_{z,t}} = \sqrt{\overline{(bias_{l,z,t} - bias_{z,t})^2}} = \sigma_{z,t}(bias_{l,z,t})$$

Where $bias_{z,t}$ is the global bias of a specific vertical layer z. The bulk relative bias of the meridional gradient is computed with respect to the standard deviation of the observed mean molar fraction at each layer z and season t:

$$relative\ STD_{Lat_{z,t}} = 100 \times \frac{\sigma_{z,t}(bias_{l,z,t})}{\sigma_{z,t}(o_{l,t,z})}$$

In the next phase of the project we will also explore the inclusion of mass budget diagnostics in the scorecard in order to provide additional information on the role of transport versus fluxes in explaining the atmospheric tracer mass distribution. To facilitate the link between the concentrations and mass which can be directly compared to the surface fluxes, we have included a bulk conversion of the biases from molar fraction to mass (shown as PgC for CO_2 and kton for SF_6 within brackets next to the molar fraction value). This is only an approximate estimate as we apply the conversion to the regional bias using a mean bulk air mass estimate in each layer obtained from the observed minimum and maximum pressure values. The purpose of providing the regional biases in units of mass is to be able to compare these atmospheric biases to the surface fluxes and thus, assess the implications of these biases on the regional/global carbon budget.

4.4 Model performance scorecards

4.4.1 Seasonal regional scorecard

Transport biases are seasonal and can often have opposite signs in different seasons. The seasonal scorecard provides details on the latitude bands and vertical layers that have the largest errors for a specific season, as well as the systematic errors in the vertical and horizontal large-scale gradients. An example of this composite seasonal scorecard over the period December to February (DJF) is shown in Figures 9 and 10 for the IFS 80km simulation of CO_2 and SF_6 respectively. We proceed to describe the main features depicted by the various sections of the scorecard.

Section A, focuses on the large-scale systematic errors in dry mole fraction, allowing to compare the bias in different layers in the atmosphere with the total column bias:

- The CO_2 scorecard shows a consistent positive bias in the lower troposphere and negative bias in the UTLS in all latitude bands, consistent with a systematic **underestimation of the vertical transport from lower to upper troposphere at global scale**. The SF_6 scorecard also has a relatively large negative bias in the UTLS, but the positive bias in the PBL is smaller and limited to the tropics. This probably indicates that the ATom data is not representative of the global zonal mean, because within the PBL, SF_6 is characterized by localised hotspots over the continents close to its sources and these are likely not well sampled by the ATom flight track. In order to ascertain this, we recommend an assessment of the observation representativeness as a preliminary step to the computation of the evaluation metrics in the scorecard. The model simulations are ideal for this type of assessment, as they contain the full 3D spatial and temporal information, allowing the comparison of the model sampled at the observation location and time with its true zonal mean value at global scale.
- The global vertical dipole in the bias results in very low biases in the total column (except for high latitudes which are discussed below). This highlights the inconsistencies between the surface and total column evaluation, and therefore, the importance of having an integrated three-dimensional evaluation approach. It also illustrates the partial constraint that total column observations provide in atmospheric inversions given the complex vertical structure of the model error.
- The largest biases are at **Northern Hemisphere (NH) high latitudes**. This is consistent with the biases in vertical transport at lower model resolutions. At coarser resolutions, more tracer remains in the lower troposphere which is transported meridionally towards high latitudes, and less tracer is being transported to the UTLS and zonally along the westerly upper-level jet. It is also important to note that there are no total column observations from satellites in winter at high latitudes, which will affect the capability of the atmospheric inversions to constrain fluxes and likely result in larger errors in the CAMS optimized CO_2 fluxes used in the CATRINE TransCom protocol.

Section B provides an evaluation of the vertical gradients:

- The relative bias in the vertical gradients across the different layers is overall positive, which means the vertical gradients are too strong in the model, reflecting the inefficiency of the model transport. This is consistent with the underestimation of vertical transport both in the NH extratropics and tropics. An exception is the vertical gradients in Southern Hemisphere (SH) mid-latitudes which are largely controlled by the inter-hemispheric transport from the NH in the upper troposphere. This suggests that there might be an **underestimation of the inter-hemispheric transport.** The relative error in SF_6 vertical gradients in the SH tropics – also related to inter-hemispheric transport – is consistent with that of CO_2 in the SH mid-latitudes, i.e. the

vertical gradient is too weak throughout the column. SF_6 also shows very large negative biases over NH mid-latitudes and tropics in the free troposphere – and surprisingly not in the PBL – because ATom is sampling the large-scale plume from SE Asia in the free troposphere, as indicated by the positive observed gradient in the observations shown by the numbers within brackets in Figure 10. This is in agreement with the known underestimation of SF_6 emissions from Asia in EDGAR.

- The CO_2 vertical gradient across the tropopause is underestimated in the NH extratropics, where it is strongest. In the tropics, where the tropopause vertical gradients are weaker, they are overestimated. While for SF_6 the tropopause gradient is underestimated in all latitude bands.

Section C scores the horizontal gradients:

- The largest relative error in the large-scale horizontal gradients is between mid-latitudes and high-latitudes in both hemispheres. The horizontal gradients are too steep in the lower troposphere and too weak in the upper troposphere and lower stratosphere, which implies the UTLS gradient associated with the **polar vortex is too weak**.

The fact that the pattern of the systematic errors in SF_6 and CO_2 are not always consistent suggests that the errors in the surface fluxes dominate the systematic errors in most regions of the atmosphere, i.e. the model transports the signal of the surface flux errors from the PBL to other parts of the atmosphere. However, from the scorecard we can also see some consistency in the large-scale distribution of the errors which is backed up by the results from sensitivity studies of model resolution.

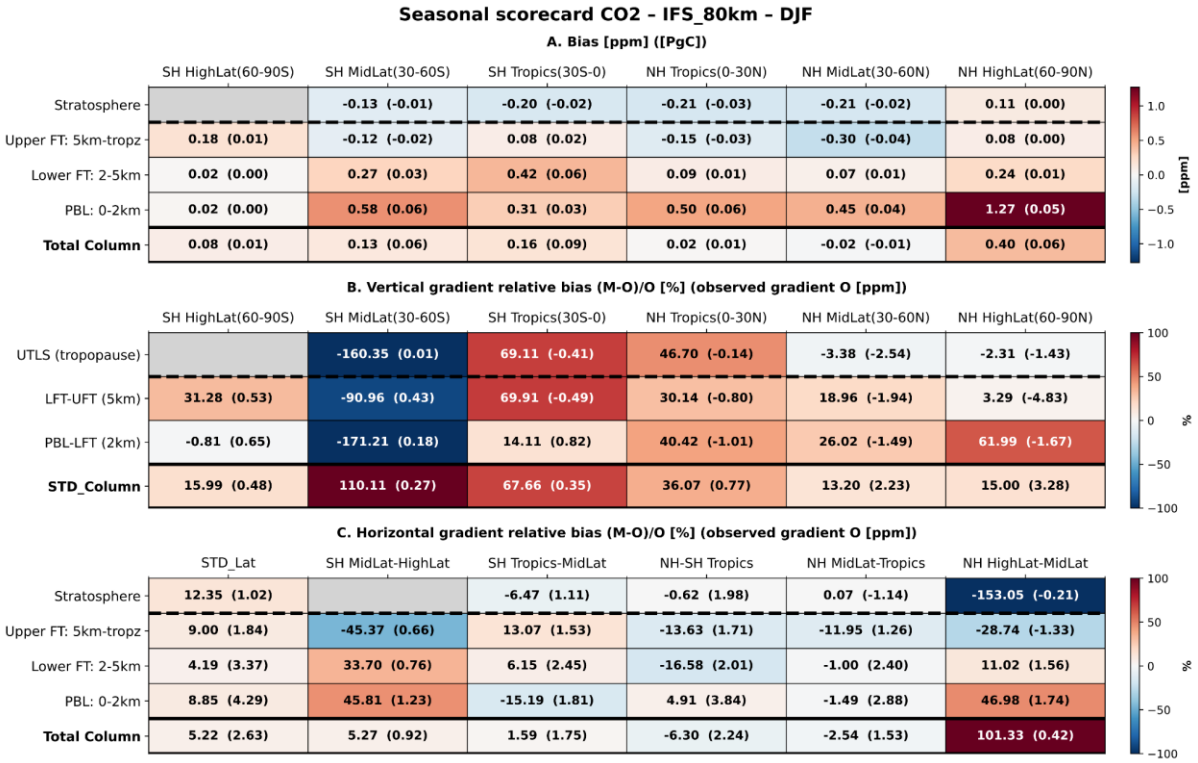

Figure 9: *Example of seasonal regional scorecard prototype for 80km IFS simulation for CO_2 in NH winter. Grey shading indicates missing data. Numbers in brackets correspond to regional biases in PgC in section A and observed gradients in ppm in sections B and C. Grey shading indicates missing data.*

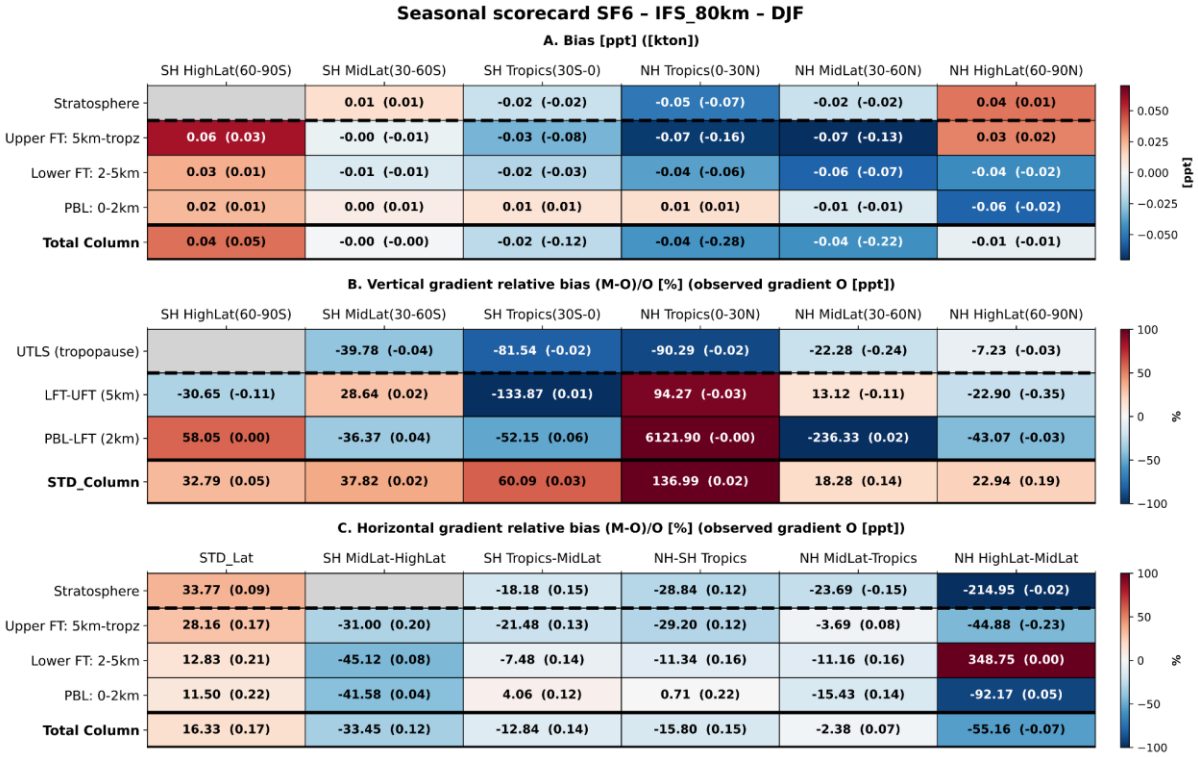

Figure 10: *Example of seasonal regional scorecard prototype for 80km IFS simulation for SF_6 in NH winter. Grey shading indicates missing data. Numbers in brackets correspond to regional biases in kton in section A and observed gradients in ppt in sections B and C. Grey shading indicates missing data.*

4.4.2 Annual global scorecard

The annual global scorecard focuses on the systematic errors in the seasonal cycle at global scale for the total column and various layers in the atmosphere. We have selected the vertical layers which have more observations in the operational observing system and have the largest transport biases. The specific values of the layers can be adapted and revised in the future according to observation availability. The errors in the vertical and horizontal gradients are summarised at the bottom of the scorecard using the standard deviation of the regional systematic errors in the vertical and horizontal dimensions (as in the seasonal regional scorecard). Figures 11 and 12 show an example of the scorecard summarising the systematic errors in the IFS 80km simulation over different seasons with the available data from the ATom field campaign. A summary of the key features in the annual global scorecard is provided below.

Seasonal cycle:

- For CO_2, it is clear that there are seasonal differences in the sign of the systematic errors associated with transport because of the change in the gradients associated with surface fluxes when they switch from being a source in the winter to a sink in the summer.
- For SF_6, the differences between seasons partly reflect the spin down from initial conditions provided to the model and the timing of the ATom field campaign data. The JJA data from ATom1 in 2016 samples the model after just 6 months of simulation, and therefore, it is partly influenced by the initial conditions. It is characterized by large positive bias in the PBL and small negative bias in the UTLS, with the exception of the

UTLS over the south pole that has a large positive bias associated with the well-known problem of SF_6 in the polar vortex. This is due to the omission of the SF_6 loss advected from the upper stratosphere/lower thermosphere. After 1 year of simulation, in DJF and SON (ATom-2 and ATom-3 respectively in 2017), the bias in the simulation has changed to much smaller negative values in the PBL and negative biases in the UTLS. SON shows larger negative biases in the UTLS because most SF_6 observations from ATom-3 are located in the Pacific ocean, sampling the large negative biases in the Asian plume, and therefore are not representative of the zonal mean in the NH extratropics and tropics.

Inter-hemispheric gradient:

- After one year of simulation (DJF) we see a negative SF_6 bias in the NH and positive SF_6 bias in SH, while for CO_2 the positive bias in the NH is stronger than the positive bias in the SH. This inconsistency points to the errors in the inter-hemispheric gradient being more influenced by errors in surface fluxes than inter-hemispheric transport. There is also the question of representativity of observations that need to be properly checked.

Overall systematic errors in the horizontal and vertical gradients:

- The standard deviation of the bias in the horizontal and vertical dimensions is comparable: between 0.1 and 0.4 ppm for CO_2 and 0.02 to 0.07 ppt for SF_6, varying with season and height. The largest common errors in the CO_2 and SF_6 **horizontal gradients** occur **in the stratosphere in SON**; while the largest errors in the **vertical gradients occur in the NH extratropics in summer** for both CO_2 and SF_6. This gives us confidence in attributing these errors to atmospheric transport. We would therefore recommend to focus on evaluating the model vertical transport in summer using the process-based scorecards developed in WP6 (see D6.1).

Global systematic errors:

- The global bias in the total column is an indicator of the error in the global growth rate of the fluxes of CO_2 and SF_6, if we neglect the chemical source of CO_2 in the troposphere and SF_6 loss in the upper atmosphere, and assume the observations are representative of the global annual mean. In this case we know that we are missing the values from NH spring (MAM) and the observations are probably not fully representative of the zonal mean of the full atmospheric column (e.g. the stratosphere is largely unsampled by the ATom observations). Notwithstanding this, the last column in the scorecard is a useful synthesis of persistent systematic errors at annual scales. There is a persistent feature, namely a consistent underestimation of the vertical transport of tracer from the PBL to the UTLS, with overestimation of lower tropospheric tracer and underestimation of tracer in the upper levels in the extratropics for both SF_6 and CO_2. Tropical UTLS regions have some inconsistent biases in CO_2 and SF_6 in JJA and SON because of the reversal of the CO_2 vertical gradient in summer.

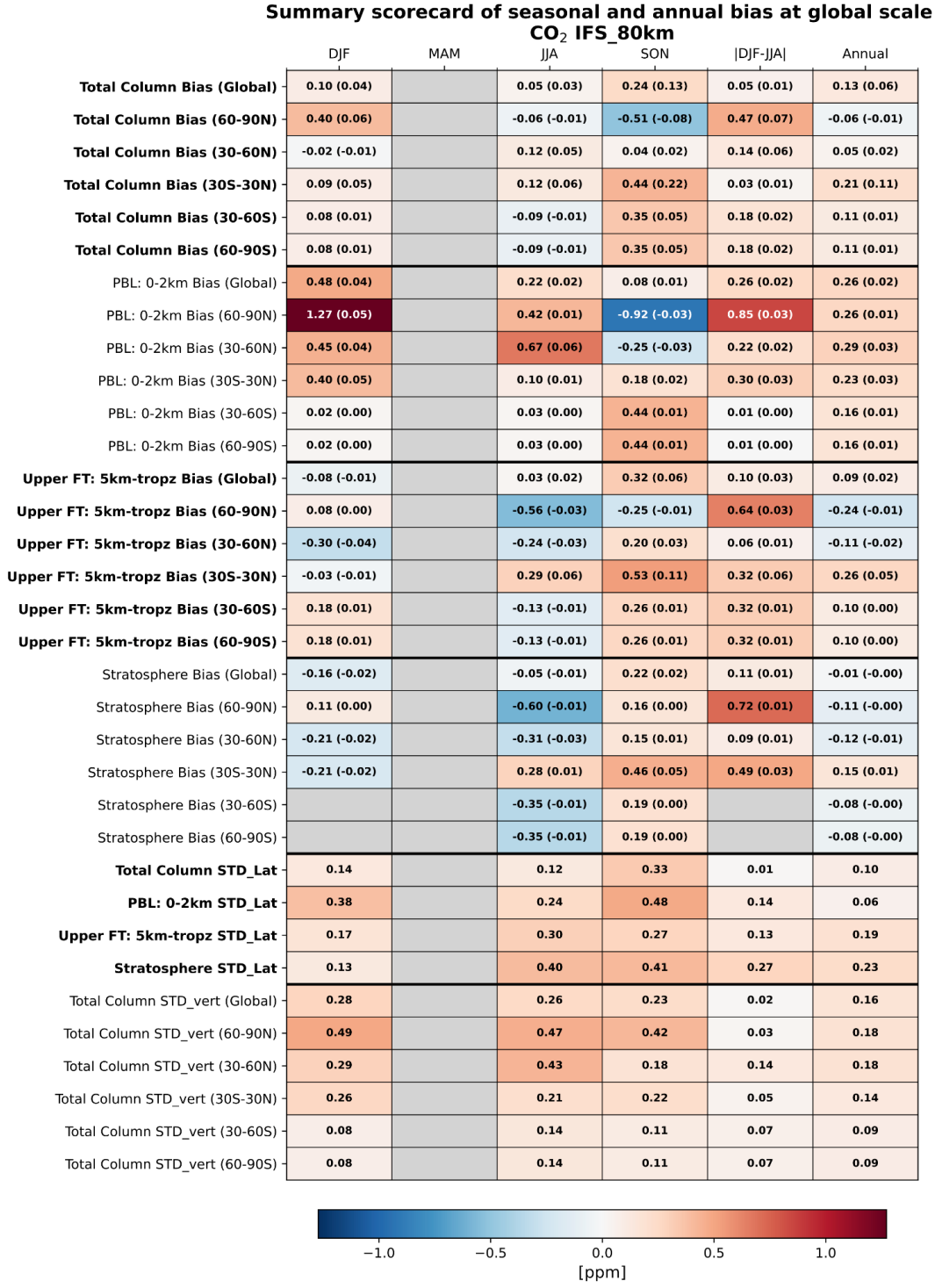

Figure 11: *Example of CO_2 annual global scorecard prototype for the IFS 80 km simulation. Columns: bias associated with the seasonal mean, the absolute difference between winter and summer and the annual mean. Rows: bias associated with regional mean for different latitude bands and vertical layers, as well as the standard deviation of the bias across latitudinal bands (STD_Lat) and across vertical layers (STD_vert). Units are ppm. The regional bias has also been provided in PgC within brackets. Grey shading indicates missing data.*

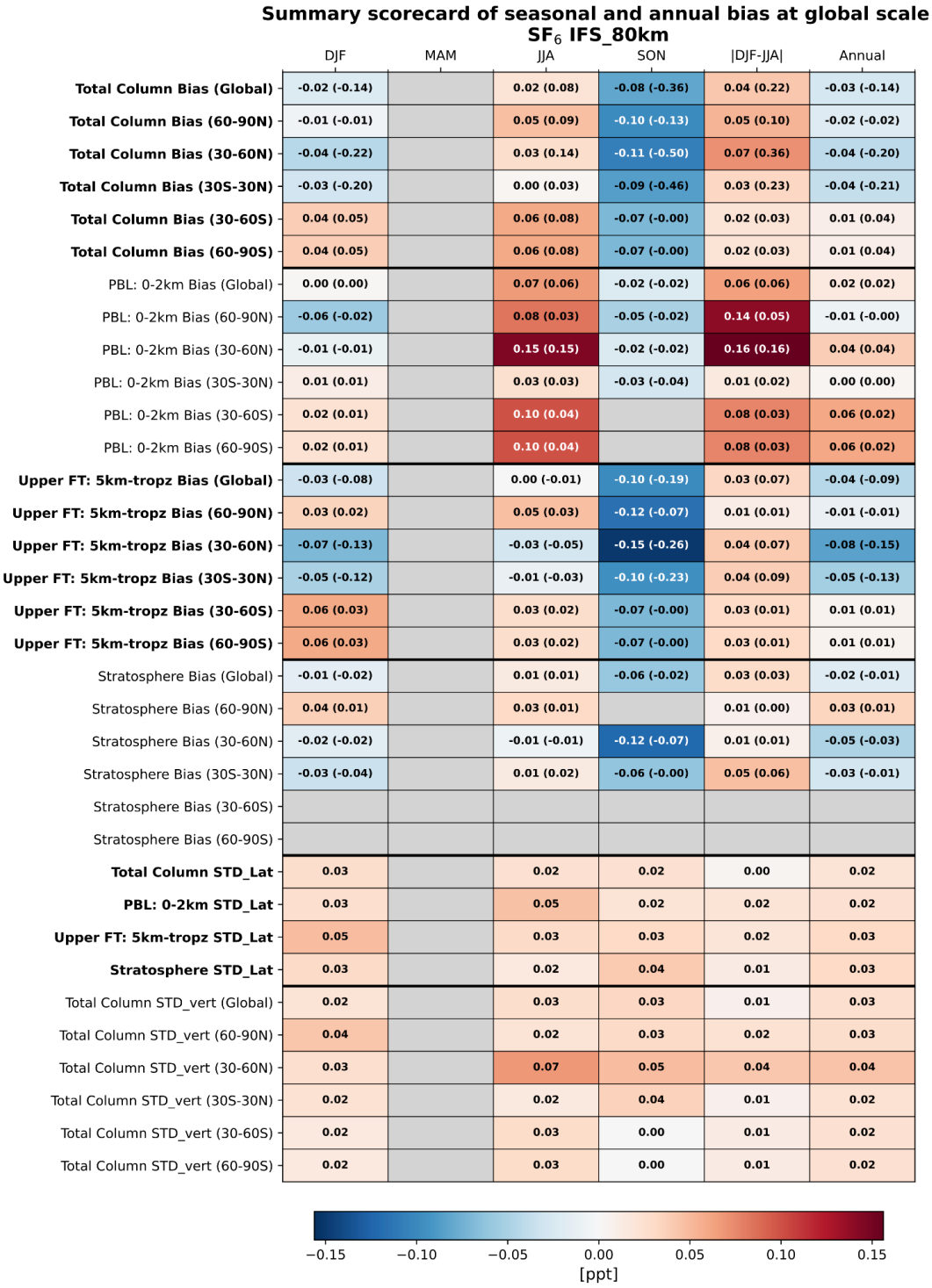

Figure 12: *Example of SF_6 annual global scorecard prototype for the IFS 80 km simulation. Columns: bias associated with the seasonal mean, the absolute difference between winter and summer and the annual mean. Rows: bias associated with regional mean for different latitude bands and vertical layers, as well as the standard deviation of the bias across latitudinal bands (STD_Lat) and across vertical layers (STD_vert). Units are ppt. The regional bias has also been provided in kton within brackets. Grey shading indicates missing data.*

4.5 Demonstration of scorecards to assess impact of model resolution

One of the main purposes of the scorecards is to be able to provide a quick integrated assessment of model developments and/or model inter-comparisons, summarising large amounts of information regarding different aspects of the model performance. In Figures 13 and 14 we use the annual global scorecard to evaluate the impact of increasing the IFS model resolution from 80km (tl255 grid) as used in the CAMS re-analysis (eac4 and egg4) to 28 km (tco399 grid) currently used in the operational CAMS GHG analysis.

For CO_2 in Figure 13, we see mostly improvements (i.e. a reduction of the absolute bias) as there are more blue inverted triangles than red triangles (see Annual column for a summary of mean impact). If we focus on the areas where the bias reduction is 0.1 ppm or larger, we see that the largest improvements are in:

- **NH high latitudes region in the layer between 0 and 2km in winter** (DJF) and in the seasonal cycle amplitude (|DJF-JJA|). This is associated with a reduction in the atmospheric transport from mid latitudes to high latitudes in the lower troposphere when the resolution is increased. This improvement is also seen in the seasonal cycle of the total column at high latitudes.
- **NH mid-latitude region between 0 and 2km in summer** (JJA) and an improvement of the seasonal cycle amplitude, reflected also in improved model performance for the total column.
- **NH mid-latitude region in the UTLS in winter** (DJF) which is probably associated with an increase in the vertical transport to the upper troposphere with higher resolution.
- **The tropical region** also shows improvements between 0 and 2km, particularly in DJF and SON; and in the UTLS in JJA and SON.
- **Vertical gradients** are improved in the NH winter with a reduction of the standard deviation of the biases across different vertical layers, particularly in the polar region.
- **Horizontal gradients in UTLS during NH summer and autumn (JJA and SON)** with a reduction in the standard deviation of biases across different latitude bands.

We also see some degradations:

- **NH high latitude region between 0 and 2 km in the summer** (JJA). This large increase in the bias (0.61 ppm) is likely due to **a positive bias in the surface fluxes**. This positive CO_2 bias is amplified with resolution because it is less compensated by the transport error bringing too much low CO_2 from mid-latitudes.
- **NH mid latitude region between 0 and 2km in autumn** (SON). The negative CO_2 bias already present at 80km is further increased with the 28km resolution. We hypothesise that this could be due to the source of CO_2 not being strong enough to compensate for the increased vertical transport with resolution.
- **Extratropical UTLS in winter (DJF in NH and JJA in SH).** We hypothesize that biases in the stratospheric polar vortex could have a larger influence on the UTLS at high resolution because of changes in the downwelling branch of the Brewer-Dobson circulation, increasing the influence of the biases in the upper stratosphere (present in the initial conditions) to the UTLS. The high-resolution simulation will also have stratospheric intrusions penetrating further down, transporting more effectively these stratospheric biases to the upper troposphere.
- The bias increase below 2km and in the UTLS have both an impact on the bias increase in the total column in summer and winter respectively.

The interpretation of the scorecard for SF_6 in Figure 14 is more complicated due to the combination of more biased initial conditions, different sampling for different seasons, and larger errors in the emissions. In JJA (during the ATom1 field campaign) there are

observations from the Pacific, Atlantic and American continent (see Figure 8). This makes the data more representative of the latitudinal band than in DJF and SON, when there are no observations over the American continent, nor over the Atlantic in the lower to mid-tropospheric layers in SON. This is particularly important for SF_6 because the SF_6 emissions are overestimated in North America and underestimated in Asia. Thus, the impact of resolution on SF_6 shows:

- Overall improvement in the bias with increased resolution during JJA, when observations from the North American continent are included. It is likely that these improvements are associated with enhanced vertical transport which lowers the positive bias in the continental PBL over North America.
- Overall degradation with increased resolution in DJF and SON, when the errors are strongly influenced by the Asian plume. This apparent degradation occurs because at high resolution, the emission hotspots are more intense and the vertical transport is also enhanced, resulting in an increased sensitivity of the 3D tracer distribution to flux errors with model resolution.

In summary, it is clear from the scorecard that the large-scale three-dimensional distribution of CO_2 and SF_6 is sensitive to model resolution. The largest impact of resolution can be found in polar regions and in the UTLS. This emphasizes the importance of having additional information over those regions, particularly because it is in those regions where satellite observations are not able to provide a robust constraint.

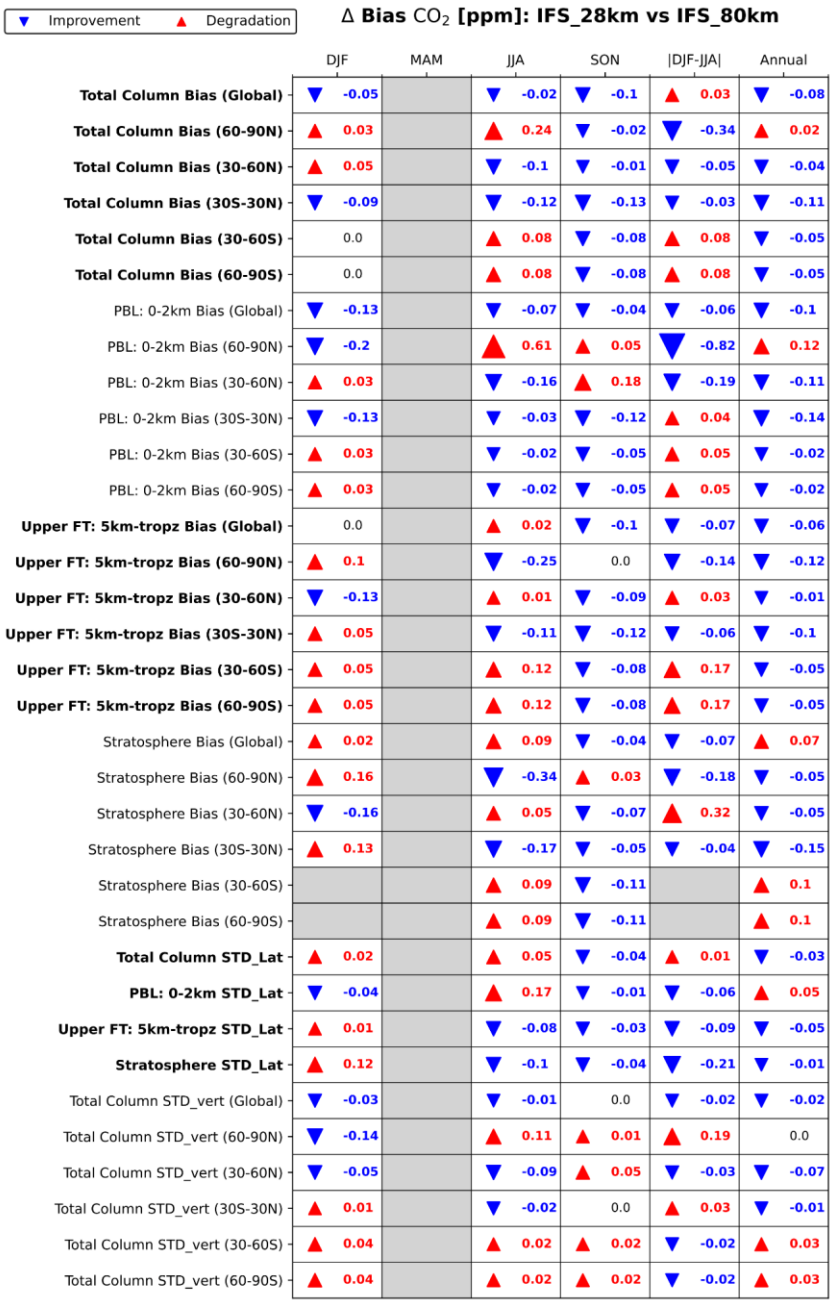

Figure 13: *Impact assessment of increasing the IFS model horizontal resolution (from 80km to 28km) on atmospheric CO_2 using the annual global scorecard. Red/blue triangles represent an increase/decrease in the systematic error over different latitude bands/vertical layers and in the standard deviation of the bias across vertical layers (STD_vert) and latitude bands (STD_Lat); grey shading indicates missing data.*

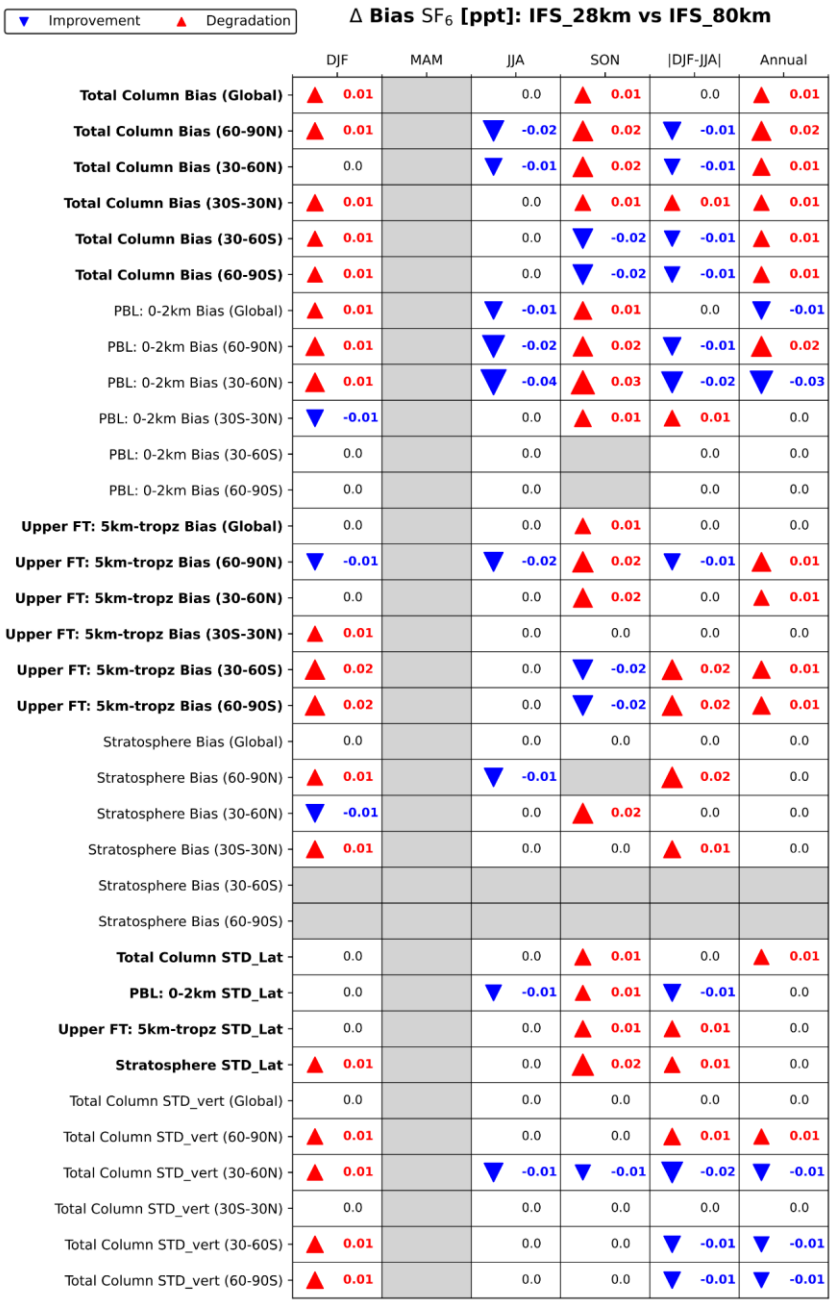

Figure 14: *Impact assessment of increasing the IFS model horizontal resolution (from 80km to 28km) on atmospheric SF_6 using the annual global scorecard. Red/blue triangles represent an increase/decrease in the systematic error over different latitude bands/vertical layers and in the standard deviation of the bias across vertical layers (STD_vert) and latitude bands (STD_Lat); grey shading indicates missing data.*

5. Conclusions and recommendations

In this deliverable, a robust assessment of TransCom model simulations is performed by comparing model simulations of CO_2 with in-situ surface and aircraft-based CO_2 observations. Alongside, a prototype scorecard framework is developed for evaluating the CO_2 and SF_6 tracers from different resolutions of IFS to provide an integrated assessment of transport and surface flux errors. Together, these two complementary approaches help to understand the skill of model simulations, their runtimes, and the zones of major transport disagreement.

A summary of the main findings and recommendations is listed below:

1. While some TransCom model simulations of CO_2 consistently performed well across different latitude bands or station categories, others consistently underperformed as compared to top-ranked simulations. The result has important implications for model development. By further comparing the mass-budget based diagnostic fields (vertically integrated total mass, vertically integrated zonal and meridional fluxes, total surface fluxes) from the best-performing model simulations with those from underperforming model simulations, we can better diagnose the causes of the differences, which can be directly linked to flux errors at continental and seasonal scales relevant to CO_2 inversions. This analysis will be carried out in the next phase of the study.

2. While higher-resolution TransCom model simulations with longer runtimes generally show improved performance, this is not for all model simulations, as some high-resolution simulations with shorter runtimes underperform, indicating that computational investment does not necessarily translate into scientific return.

3. TransCom simulations suggest the highest inter-model disagreement is in the boundary layer to lower free troposphere transition during the winter months. Inter-model disagreements decrease with altitude, from the lower free troposphere through the upper free troposphere to the UTLS. This is also reiterated in D6.1, where representative issues of boundary layer and altitude dependent biases are observed in model simulations.

4. In all TransCom simulations, systematic underestimation of CO_2 concentrations is observed in the UTLS, likely due to reduced transport of CO_2 into the UTLS from the lower troposphere. Further, the highest inter-model spread is observed in the UTLS during JJA, which follows the dynamical active pathways over the east Asian outflow towards the northern Pacific driven by the Asian summer monsoon anticyclone. This high spread is likely linked to the vertical transport mismatches among model simulations. This is also the period when the biosphere is most active, so any errors in prescribed surface fluxes may be amplified due to additional uncertainty arising from UTLS transport. The UTLS testbed scorecard developed in D6.1, uses similar aircraft campaign datasets to systematically diagnose such transport errors. This highlights the value of aircraft campaigns to identify these zones of disagreement and strongly motivates towards expanded vertical coverage measurements during JJA.

5. The scorecard of IFS CO_2 and SF_6 simulations shows that the pattern of the systematic errors is not always consistent between the two tracers. This indicates that the errors in the surface fluxes probably dominate the systematic errors in most regions of the atmosphere, i.e. there is a long-range transport of the flux error from near-surface to other parts of the atmosphere. However, the scorecard also shows some consistency in the large-scale distribution of the systematic errors, for certain seasons, regions and vertical layers, which is backed up by the results from sensitivity studies of model resolution.

6. With the scorecards, we have identified areas where the large-scale transport errors are largest and therefore deserve particular attention from modellers and also would require additional/enhanced observations:

   a) Global underestimation of vertical transport from PBL to upper troposphere resulting in negative bias in the UTLS, which is the focus of deliverables D6.1 and D6.2.
   b) Extratropics: Overestimation of meridional transport from mid to high latitudes in the lower troposphere and underestimation of vertical transport associated with warm conveyor belts and stratospheric intrusions. The result is an accumulation of tracers in the lower troposphere over the polar region.
   c) Tropics: strength of Hadley cell needs to be investigated further, as the scorecard and resolution sensitivity experiment indicate that the vertical transport in the tropics is underestimated at low resolution
   d) High latitudes in the UTLS: polar vortex appears to be too weak. Errors in the horizontal gradients in the UTLS will affect the gradients in the total column measured by satellites. This is important for all tracers, but even more so for tracers that have stronger gradients in the tropopause such as CH_4 and CO.

7. The errors in the vertical profile produce inconsistencies between the evaluation of tracers near the surface (below 2km) and the total column. The key to explain this inconsistency is the substantial bias found in the UTLS layer, largely associated with vertical transport errors. This has implications for the atmospheric inversions assimilating total column and surface observations.

8. Model resolution has a significant impact on large-scale transport. Higher resolution leads to an enhancement in the deep vertical transport to the UTLS.

9. In the next phase of the study, we will analyse the SF6 and 222Rn tracer from TransCom simulations to further disentangle the surface flux and transport related errors for the model evaluations. We will further analyse the IFS 9km horizontal resolution simulation to assess the impact of horizontal resolution on the simulation performance.

10. Representativity errors associated with observations need to be assessed. This can be addressed using model simulations. In the next phase of the project, the scorecard will be expanded to use various field campaigns available over the period of the simulations and information on their representativity of the global zonal mean will be included in the scorecard. Information on the statistical significance of the scores will also be provided.

11. The links between transport errors across spatial and temporal scales will be discussed in the next CATRINE General Assembly to bridge the scorecards from the PBL and UTLS test beds (D6.1) and the global scorecards presented here.

12. This initial scorecard prototype will be further refined in the future. Within CATRINE, it will be used to assess various transport model developments in WP2. There is also the scope to adapt the global scorecard to continental scales, such as the RECCAP regions.

13. We recommend that the scorecard prototype presented in this deliverable is further developed in the future with: operational observations from satellites and the TCCON network for the total column, vertical profile data from commercial/research aircrafts, and AirCore balloons sampling the UTLS and stratosphere and the in situ network of flasks and continuous observations from tall towers to sample the boundary layer. Although in this report we have focused on CO_2 and SF6 as long-lived transport tracers, the scorecard prototype presented here can be produced for other tracers. In particular, it will be

straightforward to extend it to CH4, CO and water vapour as these species are all available from the field campaign observations and the IFS simulations. Having more tracers will strengthen our ability to disentangle the transport errors from the errors coming from the surface fluxes and other atmospheric sources/sinks for water vapour and reactive species. It would also have the added benefit of engaging with the atmospheric air quality and NWP communities.

6. Acknowledgements

We acknowledge the individual PIs and contributors contributing to the NOAA ObsPack data product (obspack_co2_1_GLOBALVIEWplus_v10.1_2024-11-13) used in this report.

We also thank the technical and operational team of Japan Airlines, JAL Foundation, and JAMCO Tokyo for their support for the CONTRAIL aircraft campaign.

The contributions from the following people and institutions are thankfully acknowledged: B. Paplawsky, E. Gloor, E. Kort, F. Apadula, M. Kumar Sha, M. De Mazière, P. Trisolino, S. Walker, S. Piper and T. Biermann; A. Giorgio di Sarra and S. Piacentino (ENEA); A. Vermeulen (LU); A. Manning (METOFFICE); A. Beyersdorf (CSUSB); A. Zahn, F. Obersteiner, H. Boenisch and T. Gehrlein (KIT/IMK-IFU); A. Manning, G. Forster and R. A. F. de Souza (UEA); A. Karion (NIST); A. Hoheisel, I. Levin, J. Della Coletta and S. Hammer (UHEI-IUP); A. Leskinen, J. Hatakka, K. Lehtinen, O. Peltola and T. Aalto (FMI); A. Hensen, A. Frumau and P. van den Bulk (ECN); A. Andrews, B. Baier, C. Sweeney, E. Hintsa, J. Peischl, J. B. Miller, J. Mund, K. McKain, K. Aikin, K. N. Schuldt, K. Thoning, P. Tans, S. Montzka and X. Lan (NOAA); A. Jordan, C. Gerbig, H. Moossen, J. Lavric, M. Heimann, S. Zaehle and W. A. Brand (MPI-BGC); A. Colomb and J. Marc Pichon (OPGC); B. Scheeren, H. Meijer and H. Chen (RUG); B. Law and C. Hanson (OSU); B. Munger, M. Sargent and S. Wofsy (HU); B. Viner (SRNL); B. Stephens (NCAR); C. Labuschagne (SAWS); C. Lund Myhre, C. René Lunder and S. Matthew Platt (NILU); C. Couret (UBA); C. E. Miller (NASA-JPL); C. Plass-Duelmer, D. Kubistin, M. Schumacher, M. Lindauer and T. Kneuer (DWD); D. Jaffe (UofWA); D. Heltai (RSE); D. Bowling, J. Lin and L. Mitchell (U-ATAQ); D. Munro (NOAA - CIRES); D. Young, J. Pitt and S. O'Doherty (UNIVBRIS); D. Worthy (ECCC); E. Kozlova (CEDA); E. Cuevas, E. Reyes-Sanchez and P. P. Rivas (AEMET); E. Morgan, J. Kim, L. Merchant, R. Keeling, R. Weiss and S. Clark (SIO); F. Meinhardt (UBA-SCHAU); G. Vitkova, K. Kominkova and M. V. Marek (CAS); G. Chen and M. Shook (NASA LaRC); G. A. Martins (FDB); G. Manca and P. Bergamaschi (JRC); G. Brailsford and S. Nichol (NIWA); H. Riris, J. Brice Abshire and S. Randolph Kawa (NASA-GSFC); H. Matsueda (MRI); I. Lehner and M. Heliasz (LUND-CEC); I. Mammarella, J. Levula, P. Kolari and P. Keronen (UHELS); J. Necki, L. Chmura, M. Galkowski and M. Zimnoch (AGH); J. Müller-Williams (HPB); J. Turnbull (GNS); J. Morgui, R. Curcoll and S. Climadat (ICTA-UAB); J. P. DiGangi (NASA-LaRC); J. Holst and M. Mölder (LUND-NATEKO); K. Davis, N. Miles, S. Richardson and T. Lauvaux (PSU); L. Lotte Sørensen (AU); L. V. Gatti (INPE); L. Emmenegger (EMPA); L. Haszpra (RCAES); M. Delmotte, M. Schmidt, M. Ramonet, M. Lopez and V. Kazan (LSCE); M. Torn (LBNL); M. Leuenberger (KUP); M. Steinbacher (empa); M. Sasakawa, T. Machida and Y. Niwa (NIES); O. Laurent (ICOS-ATC); P. Cristofanelli (CNR-ISAC); P. Krummel, R. Langenfelds and Z. Loh (CSIRO); P. Shepson (PU); S. Newman (CIT); S. C. Biraud (LBNL-ARM); S. Morimoto (TU); S. Fang (CMA); S. De Wekker (UofVA); S. Conil (Andra); T. Schuck (IAU); T. Griffis (uminn); and V. Ivakhov (MGO).

We thank the ATom Science Team and the flight crew and support staff of the NASA DC-8 and ATom. For providing CO2 measurement support, we thank K. McKain, C. Sweeney, T. Newberger, F. Moore, and G. Diskin for NOAA Picarro CO_2 measurements; and J.W. Elkins, E.J. Hintsa, and F.L. Moore for the UAS Chromatograph for Atmospheric Trace Species (UCATS) SF6 measurements.

We are grateful to B. Stephens for his feedback on the scorecard, and for the support provided by Luke Jones and Jonathan Wilkinson on the use of flight_ver code to sample flight track from IFS simulations. We would like to acknowledge the collaboration with Noureddine Semane and the CAMEO project on the testing and implementation of the hybrid tropopause definition used in the CATRINE scorecard.

7. References

Agustí-Panareda, A., Diamantakis, M., Massart, S., Chevallier, F., Muñoz-Sabater, J., Barré, J., Curcoll, R., Engelen, R., Langerock, B., Law, R. M., Loh, Z., Morguí, J. A., Parrington, M., Peuch, V.-H., Ramonet, M., Roehl, C., Vermeulen, A. T., Warneke, T., and Wunch, D. (2019): Modelling CO_2 weather – why horizontal resolution matters, Atmos. Chem. Phys., 19, 7347–7376, https://doi.org/10.5194/acp-19-7347-2019.

Baier, B., Sweeney, C., Newberger, T., Higgs, J., Wolter, S., & NOAA Global Monitoring Laboratory. (2021). NOAA AirCore atmospheric sampling system profiles (Version 20240909) [Data set]. NOAA GML. https://doi.org/10.15138/6AV0-MY81.

Chevallier, F., Agustí-Panareda, A., Krol, M., Peters, W. and Versick, S. (2024): D7.1 Design of protocol for preliminary global model intercomparisons, CATRINE HE project deliverable, https://www.catrine-project.eu/sites/default/files/2024-06/CATRINE-D7.1-V1.pdf.

Chevallier, F., Lloret, Z., Cozic, A., Takache, S., & Remaud, M. (2023). Toward high-resolution global atmospheric inverse modeling using graphics accelerators. Geophysical Research Letters, 50, e2022GL102135. https://doi.org/10.1029/2022GL102135

Degen, J., Baier, B. C., Jöckel, P., Menken, J. M., Schuck, T. J., Sweeney, C., and Engel, A.: CO2 variability and seasonal cycle in the UTLS: insights from EMAC model and AirCore observational data, Atmos. Chem. Phys., 25, 15741–15763, https://doi.org/10.5194/acp-25-15741-2025, 2025.

Elkins, J. W., Hintsa, E. J., and Moore, F. L. (2019): *ATom: Measurements from the UAS Chromatograph for Atmospheric Trace Species (UCATS)*, Version 1, ORNL Distributed Active Archive Center, https://doi.org/10.3334/ORNLDAAC/1750. Accessed 2026-05-14.

Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. Journal of Hydrology, 377(1–2), 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003

Machida, T., Matsueda, H., Sawa, Y., Nakagawa, Y., Hirotani, K., Kondo, N., et al. (2008). Worldwide measurements of atmospheric CO2 and other trace gas species using commercial airlines. Journal of Atmospheric and Oceanic Technology, 25(10), 1744–1754. https://doi.org/10.1175/2008JTECHA1082.1

Matsueda, H., Machida, T., Sawa, Y., Nakagawa, Y., Hirotani, K., Ikeda, H., et al. (2008). Evaluation of atmospheric CO2 measurements from new flask air sampling of JAL airliner observations. Papers in Meteorology and Geophysics, 59, 1–17. https://doi.org/10.2467/mripapers.59.1

Niwa, Y., Patra, P. K., Sawa, Y., Machida, T., Matsueda, H., Belikov, D., Maki, T., Ikegami, M., Imasu, R., Maksyutov, S., Oda, T., Satoh, M., and Takigawa, M.: Three-dimensional variations of atmospheric CO2: aircraft measurements and multi-transport model simulations, Atmos. Chem. Phys., 11, 13359–13375, https://doi.org/10.5194/acp-11-13359-2011, 2011.

Petzold, A., et al. (2015). Global-Scale Atmosphere Monitoring by In-Service Aircraft – Current Achievements and Future Prospects of the European Research Infrastructure IAGOS. *Tellus Series B: Chemical and Physical Meteorology*, 67, 28452, https://doi.org/10.3402/tellusb.v67.28452.

Schuldt, K. N., Mund, J., Aalto, T., Abshire, J. B., Aikin, K., Allen, G., et al. (2024): Multi-laboratory compilation of atmospheric carbon dioxide data for the period 1957-2024; obspack_co2_1_GLOBALVIEWplus_v11.0_2026-01-12; NOAA Earth System Research Laboratory, Global Monitoring Laboratory. http://doi.org/10.25925/20250801

Stephens, B. B., Gurney, K. R., Tans, P. P., Sweeney, C., Peters, W., Bruhwiler, L., Ciais, P., Ramonet, M., Bousquet, P., Nakazawa, T., Aoki, S., Machida, T., Inoue, G., Vinnichenko, N., Lloyd, J., Jordan, A., Heimann, M., Shibistova, O., Langenfelds, R. L., Steele, L. P., Francey, R. J., & Denning, A. S. (2007). Weak northern and strong tropical land carbon uptake from vertical profiles of atmospheric CO2. Science, 316(5832), 1732–1735. https://doi.org/10.1126/science.1137004

Stephens, B.B., M.C. Long, R.F. Keeling, E.A. Kort, C. Sweeney, E.C. Apel, E.L. Atlas, S. Beaton, J.D. Bent, N.J. Blake, J.F. Bresch, J. Casey, B.C. Daube, M. Diao, E. Diaz, H. Dierssen, V. Donets, B. Gao, M. Gierach, R. Green, J. Haag, M. Hayman, A.J. Hills, M.S. Hoecker-Martínez, S.B. Honomichl, R.S. Hornbrook, J.B. Jensen, R. Li, I. McCubbin, K. McKain, E.J. Morgan, S. Nolte, J.G. Powers, B. Rainwater, K. Randolph, M. Reeves, S.M. Schauffler, K. Smith, M. Smith, J. Stith, G. Stossmeister, D.W. Toohey, and A.S. Watt (2018): The O2/N2 Ratio and CO2 Airborne Southern Ocean Study. Bull. Amer. Meteor. Soc., 99, 381–402, https://doi.org/10.1175/BAMS-D-16-0206.1

Sweeney, C., A. Karion, S. Wolter, T. Newberger, D. Guenther, J. A. Higgs, A. E. Andrews, P. M. Lang, D. Neff, E. Dlugokencky, J. B. Miller, S. A. Montzka, B. R. Miller, K. A. Masarie, S. C. Biraud, P. C. Novelli, M. Crotwell, A. M. Crotwell, K. Thoning, and P. P. Tans (2015): Seasonal climatology of CO2 across North America from aircraft measurements in the NOAA/ESRL Global Greenhouse Gas Reference Network, Journal of Geophysical Research: Atmospheres, 120, 10, doi:10.1002/2014JD022591.

Taylor, K.E. (2001) Summarizing multiple aspects of model performance in a single diagram. Journal of Geophysical Research, 106, 7183–7192.

Vogel, B., Günther, G., Müller, R., Grooß, J.-U., Afchine, A., Bozem, H., Hoor, P., Krämer, M., Müller, S., Riese, M., Rolf, C., Spelten, N., Stiller, G. P., Ungermann, J., and Zahn, A. (2016): Long-range transport pathways of tropospheric source gases originating in Asia into the northern lower stratosphere during the Asian monsoon season 2012, Atmos. Chem. Phys., 16, 15301–15325, https://doi.org/10.5194/acp-16-15301-2016.

Wagenhäuser, T., Jesswein, M., Keber, T., Schuck, T., and Engel, A.: Mean age from observations in the lowermost stratosphere: an improved method and interhemispheric differences, Atmos. Chem. Phys., 23, 3887–3903, https://doi.org/10.5194/acp-23-3887-2023, 2023.

Wofsy, S.C., S. Afshar, H.M. Allen, E. Apel, E.C. Asher, B. Barletta, J. Bent, H. Bian, B.C. Biggs, D.R. Blake, N. Blake, I. Bourgeois, C.A. Brock, W.H. Brune, J.W. Budney, T.P. Bui, A. Butler, P. Campuzano-Jost, C.S. Chang, M. Chin, R. Commane, G. Correa, J.D. Crounse, P. D. Cullis, B.C. Daube, D.A. Day, J.M. Dean-Day, J.E. Dibb, J.P. DiGangi, G.S. Diskin, M. Dollner, J.W. Elkins, F. Erdesz, A.M. Fiore, C.M. Flynn, K. Froyd, D.W. Gesler, S.R. Hall, T.F. Hanisco, R.A. Hannun, A.J. Hills, E.J. Hintsa, A. Hoffman, R.S. Hornbrook, L.G. Huey, S. Hughes, J.L. Jimenez, B.J. Johnson, J.M. Katich, R. Keeling, M.J. Kim, A. Kupc, L.R. Lait, J.-F. Lamarque, J. Liu, K. McKain, R.J. Mclaughlin, S. Meinardi, D.O. Miller, S.A. Montzka, F.L. Moore, E.J. Morgan, D.M. Murphy, L.T. Murray, B.A. Nault, J.A. Neuman, P.A. Newman, J.M. Nicely, X. Pan, W. Paplawsky, J. Peischl, M.J. Prather, D.J. Price, E. Ray, J.M. Reeves, M.

Richardson, A.W. Rollins, K.H. Rosenlof, T.B. Ryerson, E. Scheuer, G.P. Schill, J.C. Schroder, J.P. Schwarz, J.M. St.Clair, S.D. Steenrod, B.B. Stephens, S.A. Strode, C. Sweeney, D. Tanner, A.P. Teng, A.B. Thames, C.R. Thompson, K. Ullmann, P.R. Veres, N. Vieznor, N.L. Wagner, A. Watt, R. Weber, B. Weinzierl, P. Wennberg, C.J. Williamson, J.C. Wilson, G.M. Wolfe, C.T. Woods, and L.H. Zeng, 2018. ATom: Merged Atmospheric Chemistry, Trace Gases, and Aerosols. ORNL DAAC, Oak Ridge, Tennessee, USA. https://doi.org/10.3334/ORNLDAAC/1581

8. Appendix A: Obspack surface based measurement sites

Table A1 shows a table containing the list of surface based monitoring sites code and corresponding dataset names.

Table A1: List of site code and dataset used from the ObsPack data product.

site_code	Dataset
BIS	co2_bis_surface-insitu_11_allvalid
CIF	co2_cfa_surface-flask_2_representative
PUY	co2_puy_surface-insitu_11_allvalid
SGP	co2_sgp_surface-insitu_64_allvalid-60magl
RPB	co2_rpb_surface-flask_1_representative
STE	co2_ste_surface-flask_147_allvalid-252magl
CBA	co2_cba_surface-flask_1_representative
WES	co2_wes_surface-insitu_25_allvalid
YON	co2_yon_surface-insitu_19_representative
CGO	co2_cgo_surface-insitu_2_allvalid
CPT	co2_cpt_surface-insitu_36_marine
GAT	co2_gat_surface-flask_147_allvalid-341magl
BHD	co2_bhd_surface-insitu_15_baseline
MKN	co2_mkn_surface-insitu_701_allvalid
RUN	co2_run_surface-insitu_472_allvalid
SUM	co2_sum_surface-flask_1_representative
SMO	co2_smo_surface-insitu_1_allvalid
OXK	co2_oxk_surface-flask_45_representative
SYO	co2_syo_surface-insitu_8_allvalid

ZSF	co2_zsf_surface-insitu_25_allvalid
UTO	co2_uto_surface-insitu_30_allvalid
NOR	co2_nor_surface-flask_424_allvalid-100magl
LUT	co2_lut_surface-insitu_44_allvalid
CBW	co2_cbw_surface-flask_445_allvalid
SIS	co2_sis_surface-flask_45_representative
IMP	co2_lmp_surface-insitu_28_allvalid
BRW	co2_brw_surface-insitu_1_allvalid
RYO	co2_ryo_surface-insitu_19_representative
PSA	co2_psa_surface-flask_1_representative
CIB	co2_cib_surface-flask_1_representative
MBO	co2_mbo_surface-insitu_1_allvalid-11magl
SPO	co2_spo_surface-insitu_1_allvalid
NMB	co2_nmb_surface-flask_1_representative
ZEP	co2_zep_surface-insitu_56_allvalid
SEY	co2_sey_surface-flask_1_representative
IZO	co2_izo_surface-insitu_27_allvalid
USH	co2_ush_surface-flask_1_representative
KAS	co2_kas_surface-insitu_53_allvalid
WAO	co2_wao_surface-insitu_13_allvalid
ALT	co2_alt_surface-insitu_6_allvalid
BMW	co2_bmw_surface-flask_1_representative
PAL	co2_pal_surface-insitu_30_allvalid
UTA	co2_uta_surface-flask_1_representative
LEF	co2_lef_surface-flask_1_representative
MID	co2_mid_surface-flask_1_representative
CRZ	co2_crz_surface-flask_1_representative

HEI	co2_hei_surface-insitu_22_allvalid
MQA	co2_mqa_surface-flask_2_representative
WLG	co2_wlg_surface-flask_1_representative
ETL	co2_etl_surface-insitu_6_allvalid
DSI	co2_dsi_surface-flask_1_representative
MNM	co2_mnm_surface-insitu_19_representative
ERS	co2_ers_surface-insitu_11_allvalid
ICE	co2_ice_surface-flask_1_representative
CMN	co2_cmn_surface-insitu_443_allvalid
FKL	co2_fkl_surface-insitu_11_allvalid
HUN	co2_hun_surface-flask_1_representative
LHW	co2_lhw_surface-insitu_5_allvalid
MLO	co2_mlo_surface-flask_1_representative
LIN	co2_lin_surface-flask_147_allvalid-98magl
SVD	co2_svb_surface-flask_440_allvalid-150magl
NWR	co2_nwr_surface-flask_1_representative
LLN	co2_lln_surface-flask_1_representative
FORT	co2_fort_surface-insitu_60_allvalid-128magl
FSD	co2_fsd_surface-insitu_6_allvalid

Document History

Version	Author(s)	Date	Changes
1	Anna Agusti-Panareda	17-03-2026	Initial document created
2	Anna Agusti-Panareda, Chiranjit Das, Frederic Chevallier	29-04-2026	First complete draft available for internal review
3	Input from co-authors	11-05-2026	Inclusion of TM5 in Fig. 3.1 and correction of typos and clarifications of explanations.
4	Anna Agusti-Panareda, Chiranjit Das, Frederic Chevallier	13-05-2026	All comments and feedback from review addressed. Version ready for submission

Internal Review History

Internal Reviewers	Date	Comments
Jordi Vila (WUR)	2026-05-02	Comments and suggestions to emphasize links with D6.1 and D6.2

This publication reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.