CARBON ATMOSPHERIC TRACER RESEARCH TO IMPROVE NUMERICAL SCHEMES AND EVALUATION

CATRINE

Carbon Atmospheric Tracer Research to Improve Numerics and Evaluation

D5-2 Test-bed realisation

Due date of deliverable	30/6/2025
Submission date	26/6/2025
File Name	CATRINE-D5.2-V1.0
Work Package /Task	WP5 / Task 5.2
Organisation Responsible of Deliverable	KIT
Author name(s)	Stefan Versick, Jordi Vila, Anna Agusti-Panareda, Achraf Qor-El-Aine, Vincent de Feiter, Alessandro Savazzi, Mary Rose Mangan, Wouter Mol
Revision number	1.0
Status	Issued
Dissemination Level / location	Public

Funded by the European Union

The CATRINE project (grant agreement No 101135000) is funded by the European Union.

1 Executive summary

This report details the work performed to build testbeds and to evaluate the performance of greenhouse gas (GHG) transport in models, specifically DALES, the Integrated Forecasting System (IFS) and ICON-ART (ICOsahedral Nonhydrostatic model with Aerosols and Reactive Trace gases). Two targeted test cases were developed focusing on contrasting ecosystems crucial for the European and global carbon cycle: the Amazon rainforest and the grassland–temperate forest region of the Netherlands. The primary objective was to assess the models' accuracy in simulating GHG transport processes, particularly those influenced by turbulence and cloud dynamics, which are vital for the vertical redistribution of carbon dioxide (CO_2) and other trace gases.

The evaluation involved a combination of surface observations, upper-atmosphere in-situ measurements (e.g., aircraft campaigns or tall towers), and dedicated large-eddy simulations (LES). LES served as a high-fidelity benchmark due to their ability to explicitly resolve fine-scale turbulent and convective processes. The selected sites were chosen for their high-quality and comprehensive observational datasets, allowing for the study of deep convective transport in tropical forests (Amazon) and shallow convection and boundary-layer mixing in mid-latitude conditions (Netherlands). The TestBed approach encompassed statistical analysis, selection of key transport metrics (e.g., atmospheric boundary layer height, mass flux), and diagnostics of main GHG transport processes.

Key findings from the atmospheric boundary layer (ABL) testbeds indicate that while models capture overall trends, there are sensitivities to choices of model resolution and physics. Higher resolution simulations (e.g., IFS 4.4 km) generally showed improved agreement with observations in terms of correlation and variability. Focusing on the Amazonia Testbed, initial results indicate that the radiation budget is captured satisfactorily, though not perfectly. The evaporative fraction—reflecting the partitioning of net available radiation into sensible and latent heat fluxes—is within the correct order of magnitude, as is the net ecosystem exchange (NEE). However, noticeable discrepancies appear in the diurnal cycle, particularly during the morning and afternoon transitions.

An analysis of the Diurnal Carbon Range reveals a clear bias between observations and simulations across the three IFS resolutions, notably in the timing and amplitude of the daily CO_2 maximum. This bias is likely linked to limitations in the representation of the stable nocturnal boundary layer and may point to the need for a more detailed multi-layer canopy description. Further investigation will be carried out using targeted diagnostics. Finally, the atmospheric boundary layer height—a key integrative variable connecting surface and free-tropospheric processes—compares well with observations, but it is slightly over estimated probably related to the shallow cumulus formation and deepening. All of these initial evaluations will be further analysed through diagnostics of the individual terms in the governing equations for carbon dioxide.

For the Upper Troposphere Lower Stratosphere (UTLS) test cases, data from the Atmospheric Tomography Mission (ATom), Dynamics and Chemistry of the Summer Stratosphere (DCOTSS), STRATOCLIM (The Stratospheric and upper tropospheric processes for better climate predictions), WISE (Wave-driven Isentropic Exchange), and In-service Aircraft for a Global Observing System (IAGOS) campaigns were utilized. ATom data provided a robust benchmark, showing that both IFS and ICON models have considerable skill in simulating global CO_2 and SF6 tracer distributions. However, challenges remain in correctly representing the interplay of different transport processes, including vertical exchange. Comparison to the DCOTSS campaign highlighted that while both ICON-ART and IFS reproduce broad vertical structure and statistical characteristics, they struggle to reflect the complete complexity of

vertical transport, especially during overshooting convection events, often resulting in smoother vertical profiles and underestimation of vertical gradients and temporal variability. Preliminary DCOTSS results for CO_2 vertical profiles showed IFS generally performing better across most metrics compared to ICON. Comparison to IAGOS data revealed consistent underestimation of variability in the ICON model for Northern Hemisphere regions and generally mediocre agreement in correlation coefficients for vertical gradients across all regions. For STRATOCLIM, which focused on the Asian monsoon anticyclonic circulation, both ICON-ART and IFS models showed biases in CO_2 mixing ratios—ICON with a positive bias and IFS with a negative bias in the troposphere. Both models struggled to capture the observed steep negative vertical gradients, suggesting an over-diffusive representation of vertical transport. While both models performed well statistically in stable, high-altitude conditions, IFS demonstrated better statistical performance during dynamic ascent and descent phases, indicating a more realistic representation of vertical mixing and advection compared to ICON. Simulations of the WISE campaign, which investigated UTLS composition and dynamics, show that both models consistently underestimated observed CO_2 variability and generally showed a positive bias. Notably, ICON produced a remarkably flat vertical profile, capturing virtually none of the observed vertical structure, while IFS showed better, though still limited, skill in reproducing vertical CO_2 gradients and structure.

Challenges encountered include memory errors in ICON high-resolution simulations and a concerning CO_2 drift in the ICON model for UTLS test cases. Additionally, the IFS model setup exhibited a consistent high bias in CO concentrations, which propagated into ICON runs by using boundary conditions from IFS. These issues are actively being investigated. Despite these challenges, the project has successfully demonstrated the feasibility of integrating intensive field campaigns, numerical experiments, and model evaluation workflows, yielding valuable insights into greenhouse gas transport.

Table of Contents

2 Introduction

2.1 Background

As part of the Copernicus Atmosphere Monitoring Service (CAMS), a new service will be established to monitor emissions of CO_2, CH_4 and relevant air pollutants, referred to as the CO_2 Monitoring and Verification Support (CO2MVS) capacity. The CAMS CO2MVS capacity is targeted for operational status in 2026 to provide support to the 2028 Global Stocktake using observations from the CO2M satellite constellation as well as other satellite sensors and in-situ networks. The CATRINE project follows in the footsteps of previous and current H2020 and Horizon Europe projects that were set up to scope, design, develop, and implement prototype systems for the future operational CO2MVS (CHE, VERIFY, CoCO2 and CORSO). CATRINE follows the recommendations from the CHE project to provide improvements and quality control metrics for modelling tracer transport in the CO2MVS which will be crucial for the reliable use of the satellite observations in the operational system.

Uncertainties and errors in the transport of greenhouse gases (GHGs) are often related to the inaccurate representation of unresolved processes, namely the sub-grid processes occurring at smaller spatiotemporal scales than the grid (Schuh and Jacobson, 2023; Yu et al., 2018). These sub-grid processes require the use of representations that approximate their physics in the form of parametrisation schemes. These processes occur and act at spatiotemporal scales that are smaller compared to the resolved circulation. Representative examples of these parametrisations are the transport driven by dry and moist convective turbulence (mainly clouds)

To quantify these uncertainties and systematic errors we have designed testbeds with the aim to systematically identify errors (days-week comparison), and statistically evaluate models at the process level (seasonal comparison). In this deliverable, the main purpose is to identify large-scale challenges that are related to transport parametrisation schemes.

The testbed research strategy is divided in two parts (i) a comprehensive comparison of short periods (up to 15 days) with a systematic comparison with numerous observations from field campaigns, operational observing networks and large-eddy simulations (LES), and (ii) intercomparison of representative metrics such as atmospheric boundary layer height and transport driven variables like flux divergence to identify systematic errors.

Two regions of the atmosphere have been selected as the focus of the testbeds: the boundary layer (BL) including the exchange with the free troposphere, and the upper troposphere lower stratosphere (UTLS). These have been identified as areas of priority for the diagnostics of systematic errors as they are subject to large uncertainties and they play a very important role in the vertical transport of tracers (Stephens et al., 2007; Gerbig et al. 2008; Gaubert et al., 2019) across two transport barriers in the atmospheric column, i.e, the boundary layer top (Kretschmer et al., 2012) and the tropopause (Deng et al. 2015), as well as the long-range transport and the inter-hemispheric gradient (Schuh et al., 2019).

Due to the availability of high quality observations, two ecosystems have been selected for the BL testbeds: the Amazonian rainforest and grasslands and forests in temperate climate conditions in the Netherlands. For the upper troposphere, lower stratosphere (UTLS) testbeds around the globe and in different seasons have been chosen, e.g. the ATom aircraft campaign which spans almost all latitudes in three different seasons or the DCOTSS aircraft campaign which measured during an overshooting event. More details and further examples can be found in chapter 2.4. They will be used to assess models' skills in vertical transport and long-range transport. One common challenge for the transport schemes near the ground and in the UTLS are the large vertical gradients of the trace gases. The strategy for the UTLS testbed is looking into metrics determined from trace gas distributions.

As outlined below, simulations will be performed with DALES (Dutch Atmospheric Large-Eddy Simulation), ICON-ART (ICOsahedral Non-hydrostatic modell - Aerosol and Reactive Trace gases) and the IFS models (Integrated Forecasting System). The IFS model will be the core global model of the CO2MVS, and the ICON-ART model will be used operationally by DWD and EMPA to monitor the national GHG emissions. The DALES model has been used previously as part of the Amazon testbed to evaluate Numerical Weather Prediction (NWP) models (Vilà-Guerau de Arellano et al., 2022).

2.2 Scope and objectives of this deliverable

This deliverable focuses on the design of a testbed, integrating a comprehensive in-situ and remote sensing, dedicated DALES experiments and the IFS (three resolutions) and ICONART. Based on this Testbed, we perform a systematic the evaluation of global transport model simulations using the testbeds developed in Task 5.1. Our primary objectives are:

- To calculate model performance score metrics for global transport models, collaborating with WP7 and WP8.
- To evaluate model performance within the atmospheric boundary layer (ABL) by:
  - Intercomparing the temporal evolution and spatial distribution of mean thermodynamic states, tracer variables, and fluxes.
  - Defining and calculating advanced metrics, such as mass fluxes and turbulent exchange coefficients, under various conditions (convective, stable ABL, cloudy boundary layer).
- To evaluate model performance in the upper troposphere and lower stratosphere (UTLS) by utilizing tracer-tracer correlations to differentiate between transport and chemistry errors.

This evaluation will leverage the established utility of the testbeds for identifying systematic and random errors, ultimately contributing to the operational implementation of these testbeds for atmospheric tracer transport evaluation within the CO2MVS.

2.2.1 Work performed in this deliverable

The work performed in this deliverable is as per the Description of Action.

2.3 Testbeds for atmospheric boundary layer

To evaluate the performance of GHG transport in the global models IFS (Integrated Forecasting System) and ICON-ART (ICOsahedral Nonhydrostatic model with Aerosols and Reactive Trace gases), two targeted test cases have been developed. These testbeds focus on contrasting ecosystems that are critical for the European and global carbon cycle: the Amazon rainforest and the grassland–temperate forest region of the Netherlands. The tests are seen as complete pilot studies to be extended in other ecosystems and long-term periods (one or more years).

The main objective is to assess how accurate these models simulate GHG transport processes, particularly those influenced by turbulence and cloud dynamics, which play a key role in the vertical redistribution of CO_2 and other trace gases. These processes are especially important in forested, grassland and crop environments, where canopy structure, surface heterogeneity, and land–atmosphere interactions can strongly modulate local and regional atmospheric composition. The urban environment is studied in Work Packages 3 and 4, with plans for a dedicated simulation that links both work packages through a coordinated case study. This case will include Rotterdam (urban), Cabauw (grassland), and Loobos (forest), leveraging their contrasting surface types. Comprehensive observations for these sites were collected during the Ruisdael campaign RITA, conducted in August–September 2022.

The model results will be evaluated using a combination of surface observations, upper-atmosphere in-situ measurements (e.g. aircraft campaigns or tall towers), and dedicated large-eddy simulations (LES). LES are used as a reference because they explicitly resolve fine-scale turbulent and convective processes, providing a high-fidelity benchmark that is not achievable in coarser-scale global models.

The two sites were selected based on the availability of high-quality and comprehensive observational datasets. The Amazon site provides a unique opportunity to study deep convective transport in a tropical forest setting, while the Dutch site—characterised by grasslands and temperate forests—offers insights into shallow convection and boundary-layer mixing under mid-latitude conditions. In all cases, the testbed delivers a systematic comparison with the following variables: radiation balance, surface energy balance including the carbon exchange fluxes, atmospheric boundary layer values of the state variables (temperature, wind, specific humidity) and CO_2.

The specific time periods for the simulations are chosen to coincide with intensive observation campaigns and well-documented meteorological conditions, ensuring robust comparison between models, LES, and observations.

The specific periods under investigation are:

- Rainforest Testbed: 10-18 August 2022. Dry season in the Amazon Basin

The case study corresponds to a shallow cumulus situation observed over the Amazon Basin during the dry season, as part of the CloudRoots 2022 campaign *(Vila-Guerau de Arellano et al., 2024*).

- Grassland-Temperate forest Testbed:17-18 May 2023 Growing season with the possibility to extend to the entire year

Within the testbed our approach consists of three main parts: (a) complete evaluation of the variables using statistical analysis, (b) selecting key metrics of the transport such as atmospheric boundary layer height and mass flux and (c) diagnostic of the main processes for the transport of greenhouse gases. Below we provide a short description of the current status of the three parts.

a. Statistical evaluation

Table 1 provides a representative example of the main findings of the testbed of Grassland-temperate forest. The systematic comparison enables us to identify which variables are comparing satisfactorily against observations (obs). The models in the testbed are large-eddy simulations: mhh-js (MicroHH with Jarvis Stewart CO_2 exchange at the surface), mhh_ags (van Heerwaarden et al., 2017) using a photosynthesis-stomatal aperture representation (Ags), dales_knmi (Heu et al., 2015) (only meteo), dales_co2 (using the Ags version) and IFS (with high resolution of 4.5 x 4.5 km^2) (Boussetta et al., 2013). I In general, the intercomparison shows satisfactory agreement; however, a systematic, variable-by-variable evaluation is currently underway and will be discussed during the workshop on July 2–3 2025.

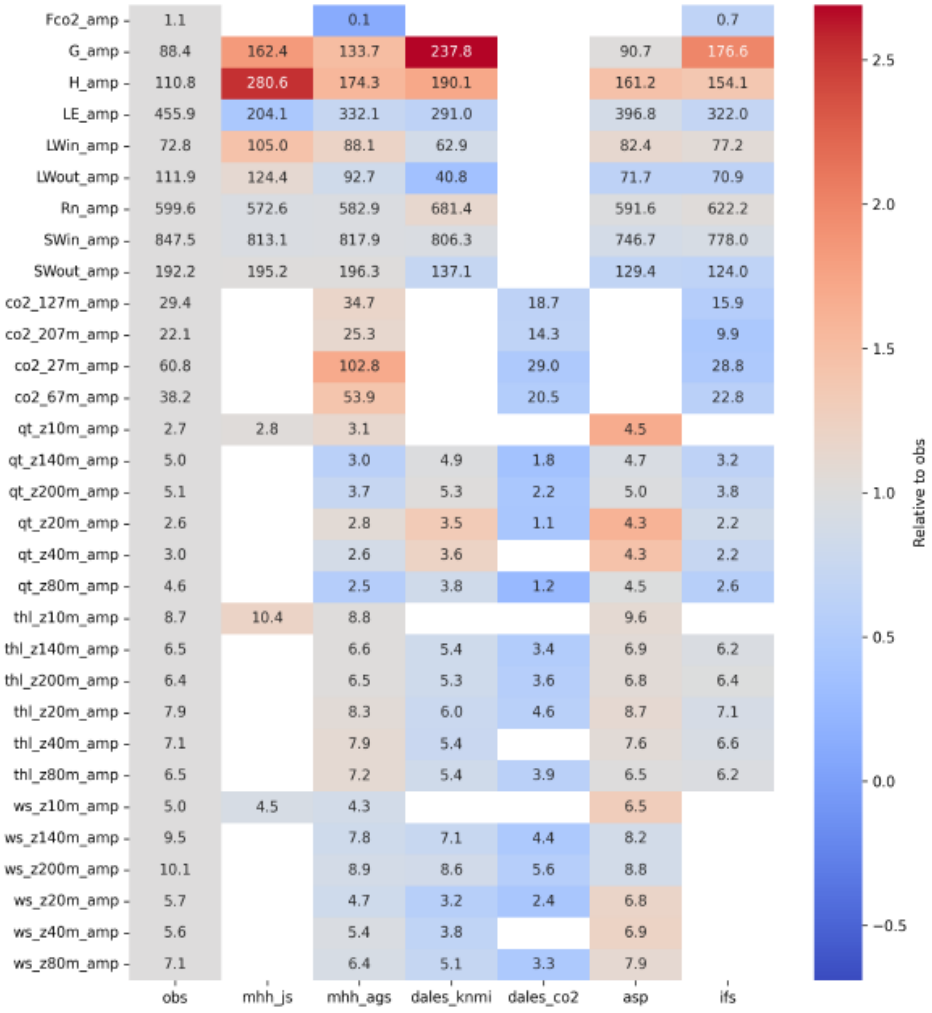
b. Selecting key variables: Example ABL height

Figure 1 presents an evaluation of the ABL height, a key variable governing the exchange between the ABL and the free troposphere (FT) observed and calculated in the Rainfores testbed. The figure includes observational estimates (OBS) inferred from frequent high-resolution radiosonde soundings, as well as outputs from LES and the Integrated Forecasting System (IFS) run at three horizontal resolutions: 25 km, 9 km, and 4.5 km.

Given the central role of ABL height in regulating vertical mixing and the near-surface concentration of GHGs, its accurate representation is critical—not only in terms of its mean magnitude but also its diurnal and seasonal evolution. Misrepresentation of the ABL height can lead to errors in the vertical distribution and transport of trace gases, ultimately affecting model-derived estimates of surface-atmosphere exchange.

This evaluation forms part of a collaborative study that integrates observational constraints with process-level simulations to identify the sensitivity of ABL height to model resolution and physics choices. The results provide essential benchmarks for improving parameterisations of boundary layer processes in large-scale models.

***Table 1**. The average daily amplitude for observations (first column) and the models (other columns) against several observed variables at the Cabauw towe during the period 17-18 May 2022r. The colours are magnitude of the modelled amplitude with respect to the observed amplitude. Red colours mean that the model had a larger daily amplitude than the observations, and the blue colours mean that the model had a smaller daily amplitude compared to the observations. Fco2: flux carbon dioxide, G; ground heat flux, H: sensible heatflux, LE: evaporation, LWin longwave in, LWout longwave out, Rn available radiation, SW shortwave, CO_2 mole fraction at different heights, qt: specific humidity at different heights, thl: liquid potential temperature at different heights and ws: wind speed at different heights, more specifically 10, 140 and 200 neters.*



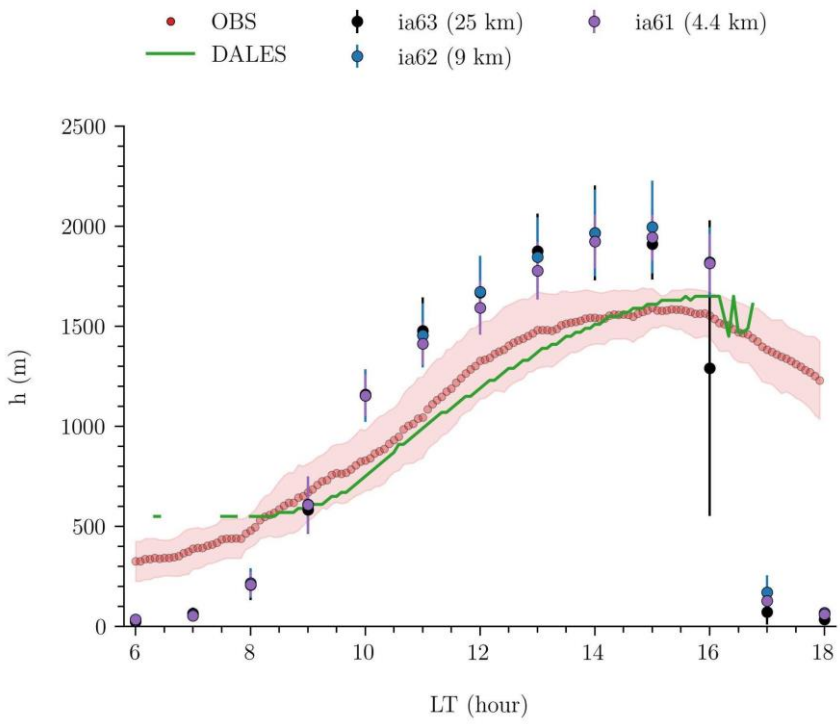


Figure 1. *Diurnal evolution of the atmospheric boundary layer height (h) during a representative case study at one of the testbed supersites. The observations are a 6-day aggregate of shallow cumulus observed and anlzyzed in the .Rainforest testbed Red circles indicate observational estimates (OBS) derived from frequent radiosonde soundings, with the shaded area representing the interquartile range. The green line shows results from the Large-Eddy Simulation (DALES). Colored markers represent simulations from the IFS model at different horizontal resolutions: 25 km (black, ia63), 9 km (blue, ia62), and 4.4 km (purple, ia61). Vertical error bars denote the spread among ensemble members or temporal variability.*

This analysis is complemented using Taylor diagrams, which provide a concise statistical summary of model performance in representing ABL height and its governing variables. As an example, Figure 2 (left) presents both a scatter plot of observed versus modelled ABL height and a Taylor diagram (right) summarising the performance of the IFS model at three resolutions: 25 km, 9 km, and 4.4 km.

The figure highlights the impact of horizontal resolution: the highest resolution simulation (IFS 4.4 km) shows improved agreement with observations, both in terms of correlation and variability, compared to the coarser-scale runs.

These results underscore the importance of resolving fine-scale processes in the representation of boundary layer dynamics and support the continued development of convection-permitting models for GHG transport studies.

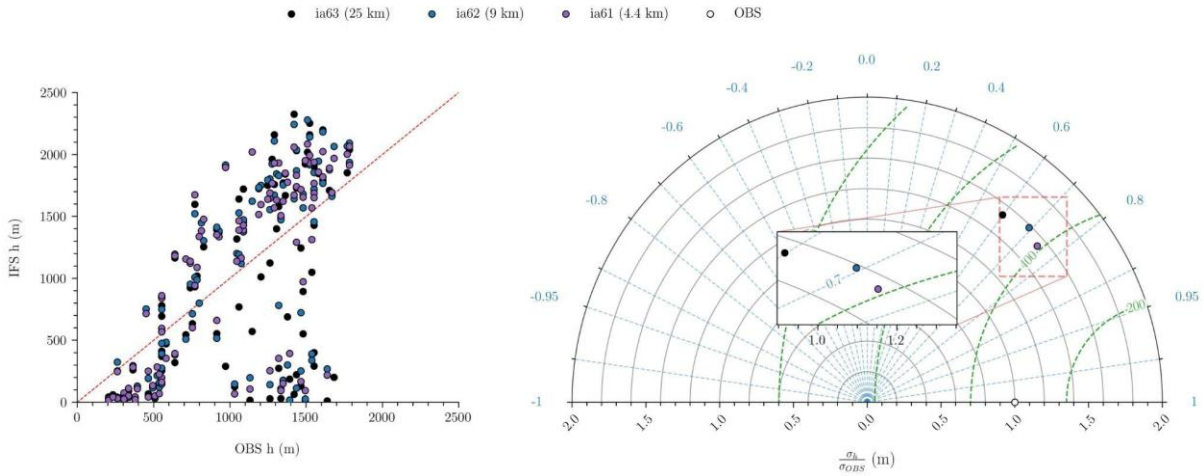


Figure 2 Left: *Scatter plot comparing observed atmospheric boundary layer (ABL) heights - derived from frequent radiosonde soundings collected from 8-22nd August 2022 - with ABL heights simulated by the Integrated Forecasting System (IFS) at three horizontal resolutions: 25 km (ia63, black), 9 km (ia62, blue), and 4.4 km (ia61, purple). The 1:1 line (red dashed) indicates perfect agreement.* ***Right:*** *Taylor diagram summarizing the statistical performance of the IFS simulations for the same case. The diagram shows the correlation coefficient, standard deviation (normalised by observations), and centred root-mean-square error (RMSE) for each resolution. Higher resolution (4.4 km) simulations show improved agreement with observations, indicating better representation of boundary layer processes. The case corresponds to a shallow cumulus day over the Amazon Basin during the CloudRoots 2022 campaign (Vila-Guerau de Arellano et al., 2024).*

c. Diagnostic of the transport GHG

Closely connected to this evaluation and as an outcome of the Rainforest and Grassland-Temeparte forest testbed analysis the following deliverable will be to determine the diagnostic of the processes. This work has already started and it will be discussed in depth during the CATRINE workshop in Huly 2-3. In other words, we decompose the transport of GHG in the main components: mass flux ventilation by clouds and entertainment. In doing so, we assess if the process can be improved and how. This part will be completed in a dedicated workshop entitled: Metrics to evaluate the transport processes in global tracer transport models to be held in Wageningen 2-3 July 2025. The outcome of the workshop, where the main results will be reported and summarized as deliverable, will be a protocol in the determination of the diagnostic of the budget CO_2 and a final assessment of the transport between the atmospheric boundary layer and the free troposphere.

2.4 Testbeds for the Upper Troposphere Lower Stratosphere (UTLS)

Model simulations were performed following the TransCom modelling protocol proposed in WP7. Following the protocol, IFS and ICON have used the same emissions and the same dynamics ("nudged" to ERA5). The protocol focusses on the transport of CO_2, CH_4, CO, and SF6. For CO and CH_4 the model setups have been slightly different: Instead of passive tracers both models used their own simplified chemistry. CO and CH_4 in ICON were initialized for specific periods with the mixing ratios from the IFS simulations, which show a significant bias. In the following sections, we will present results of the model evaluation against ATom (2.4.1), The Dynamics and Chemistry of the Summer Stratosphere (DCOTSS) campaign (2.4.2) In-service Aircraft for a Global Observing System (IAGOS) (2.4.3), Stratospheric and upper tropospheric processes for better climate predictions (Stratoclim) (2.4.4), and Wave-driven Isentropic Exchange (WISE) (2.4.5).

For ICON the transport scheme ("hadv52aero") that was found to be the best within D1.1 was used for all the simulations.

2.4.1 ATom

The Atmospheric Tomography Mission (ATom, Wofsy et al., 2018) was a NASA Earth Venture Suborbital-2 mission focused on understanding the impact of human-produced air pollution on greenhouse gases and chemically reactive gases in the atmosphere. The mission utilized the NASA DC-8 aircraft, deploying an extensive gas and aerosol payload for systematic, global-scale atmospheric sampling. The aircraft continuously profiled the atmosphere from 0.2 to 12 km altitude. Flights were conducted in each of the four seasons from 2016 to 2018. These flights originated from the Armstrong Flight Research Center in Palmdale, California, and followed a global circumnavigation route, flying north to the western Arctic, south to the South Pacific, east to the Atlantic, north to Greenland, and returning to California across central North America. The ATom mission established a single, contiguous, global-scale dataset, providing critical information for validating models.

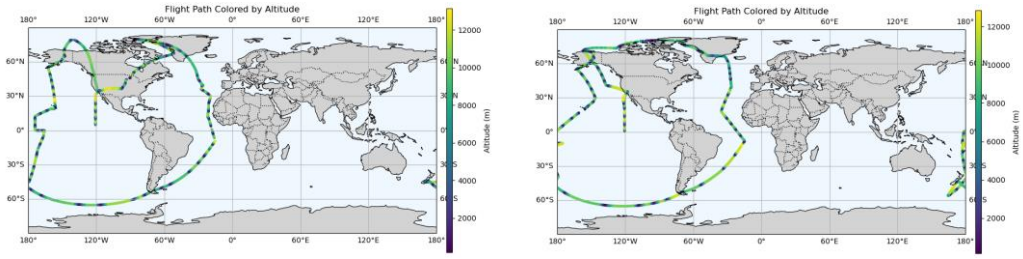


Figure 3: Flight pattern of ATom campaign. Left: ATom1; Right: ATom2

The four ATom flight campaigns and their date ranges are:

- ATom-1: July 29-August 23, 2016
- ATom-2: January 26-February 21, 2017
- ATom-3: September 28-October 28, 2017
- ATom-4: April 24-May 21, 2018

For the purpose of CATRINE, only ATom-1, ATom-2, and ATom-3 fall within the relevant timeframe of the simulations. An extension to the time period of ATom-4 is under discussion. We are utilizing the MER10 dataset for our comparisons, which provides merged measurement data at 10-second intervals across all instruments.

Several instruments onboard the NASA DC-8 measured CO_2, CO, CH_4, and/or SF6:

- AO2 (NCAR Airborne Oxygen Instrument): Measures CO_2.
- Medusa (Medusa Whole Air Sampler): Measures CO_2.
- NOAA Picarro: Measures CO_2, CH_4, and CO.
- PANTHER (PAN and Trace Hydrohalocarbon Experiment): Measures CH_4, CO, and SF6.
- PFP (Programmable Flask Package Whole Air Sampler): Measures SF6, CO_2, CH_4, and CO.
- QCLS (Quantum Cascade Laser System): Measures CO_2, CO, and CH_4.
- UCATS (UAS Chromatograph for Atmospheric Trace Species): Measures SF6, CH_4, and CO.

The main differences between these instruments lie in their measurement techniques and specific gas suites. For example, AO2 and NOAA Picarro are spectrometers, with NOAA Picarro specifically being an in-situ spectrometer. Medusa and PFP are whole air samplers, which typically involve collecting air samples for later analysis. PANTHER and UCATS both employ gas chromatography, which separates different chemical components in a sample. QCLS utilizes laser absorption for its measurements. These varied methodologies offer different sensitivities, temporal resolutions, and accuracies, contributing to the comprehensive nature of the ATom dataset.

For the ATom campaign, we compared both IFS and ICON model-simulated CO_2 and SF6 with observational data. We analyzed the data by dividing the campaign into its three distinct periods. For spatial binning, we used 20-degree latitudinal bins and 1 km vertical bins.

To assess average differences, we calculated model_value – observation_value for each observation point (in ppm) and then averaged all values within each bin. This quantity effectively highlights model biases, and changes in these biases with altitude can indicate issues with vertical transport.

Additionally, we calculated the correlation coefficient between the model and observations within each bin. This metric primarily reveals skill to represent the finer structural details of the gas distribution.

To assess overall behaviour in the vertical, we also produced Taylor diagrams for CO_2 for the 20-degree latitudinal bins and the complete column for ATom1. Those diagrams will highlight the model skill to represent the vertical structure of the profiles.

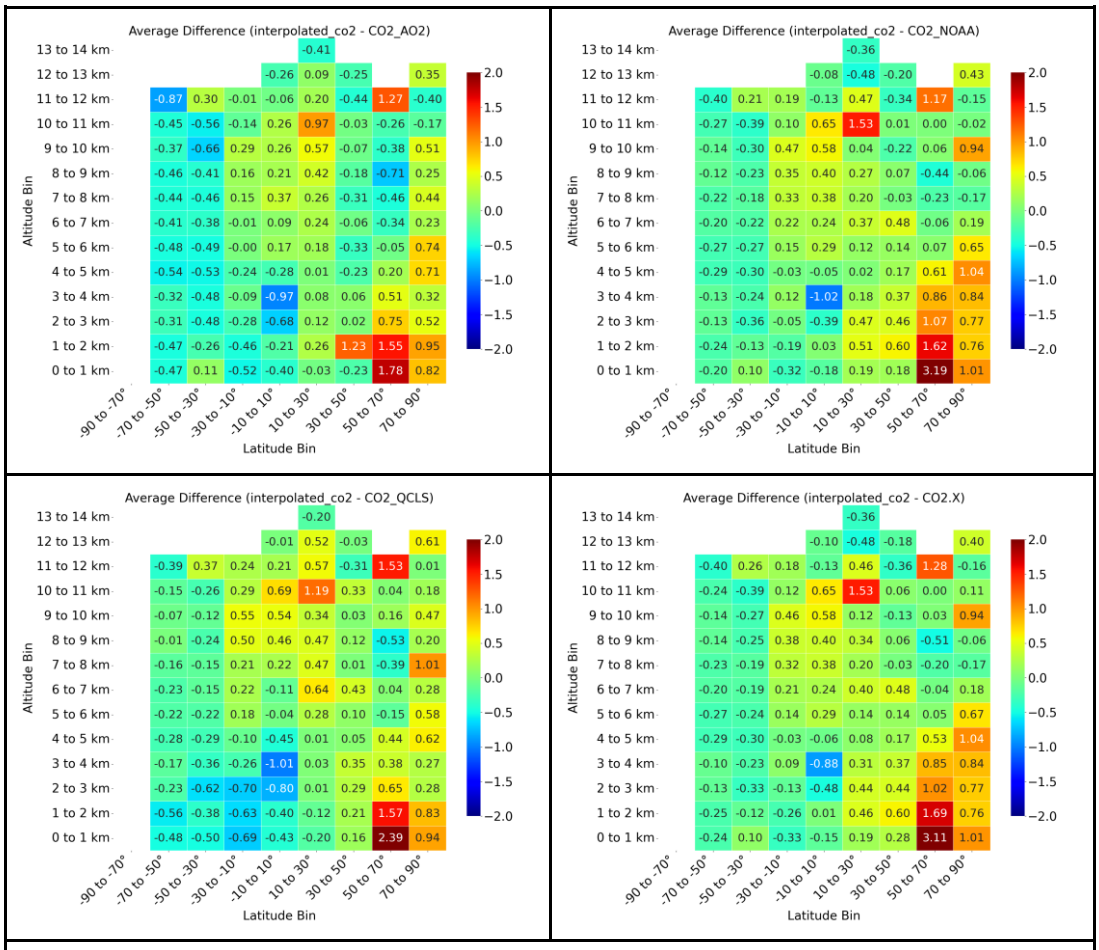


Figure 4: Average Difference (model-observation in ppm) from CO2 for ATom1 for the IFS mode compared to A02 instrument (top left), to NOAA instrument (top right), QCLS instrument (bottom left) and combined CO2 (bottom left)

The CO_2 simulations from the IFS model were compared against various instrument datasets obtained during the ATom1 campaign. These datasets included measurements from the NCAR Airborne Oxygen Instrument (AO2), the NOAA Picarro instrument, the Quantum Cascade Laser System (QCLS), and a combined CO2 product.

Regarding average differences (Figure 4), the IFS model generally exhibited a positive bias for CO_2 when compared to all instruments in the northern high latitudes, particularly near the surface. In contrast, a small negative bias was observed in the southern hemisphere, also near the surface.

Concerning correlation coefficients (Figure 5), the IFS model demonstrated strong correlations with all CO_2 observational datasets. The highest correlation was found in the low northern latitudes. Conversely, the correlations were weakest in the high southern latitudes within the lower troposphere, and similarly, in the inner tropics within the lower troposphere. A slight decrease in correlation was also noted in the high northern latitudes from approximately 3 km altitude upwards. In the southern hemisphere there is a small negative bias near the surface.

For a better understanding of those values one has to look into the profiles (see figure 6). As example we have chosen the latitudinal band between 10°S and 10°N. This shows a lower correlation near the ground which partly can be attributed to lower gradients for individual profiles there.

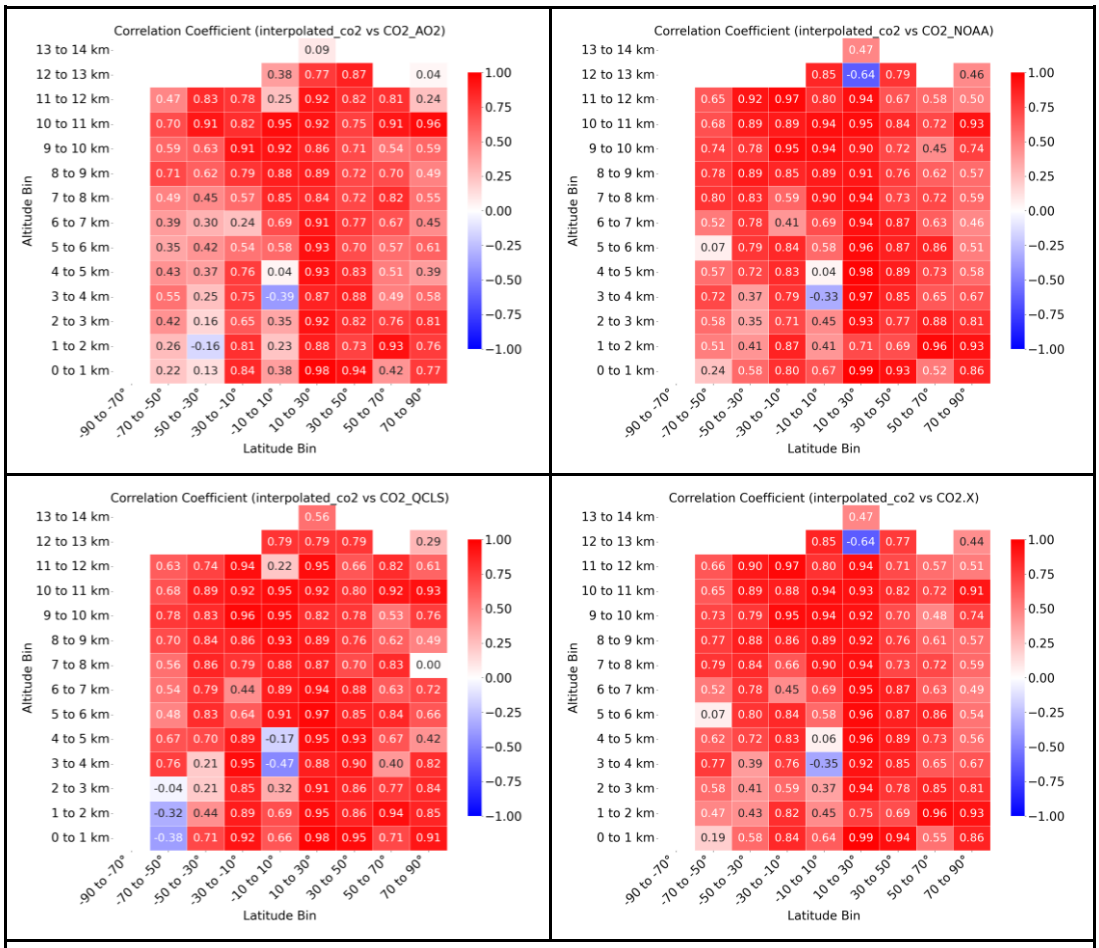


Figure 5: Same as figure 4, but correlation coefficient instead of difference

The very low correlation coefficient between approximately 3 and 5 km reveals a different problem: Both models and also the observations show distinct deviations in the CO2 mixing ratio from the otherwise quite flat profiles in the lower troposphere. The worse correlation there mainly comes from a displacement in altitude. While all three show higher mixing ratios between 2 and 3 km, ATom also has some higher values from 3 to 5 km where IFS and ICON mainly show a flat profile or in the case of IFS even some lower values.

Going higher up where the correlation coefficient is quite good both models and the observations show some profiles within this latitude band that show decreasing mixing ratios. This well captured vertical structure is the main reason for the good correlation coefficient.

We want to point out that part of the correlation is also coming from the datapoints that are coming from different ascents and descents within the latitude bands. Overall this comparison in combination with the profiles show that the metrics that are used within the Taylor diagram are useful to investigate the vertical structure of profiles and clearly indicate where problems can be found.

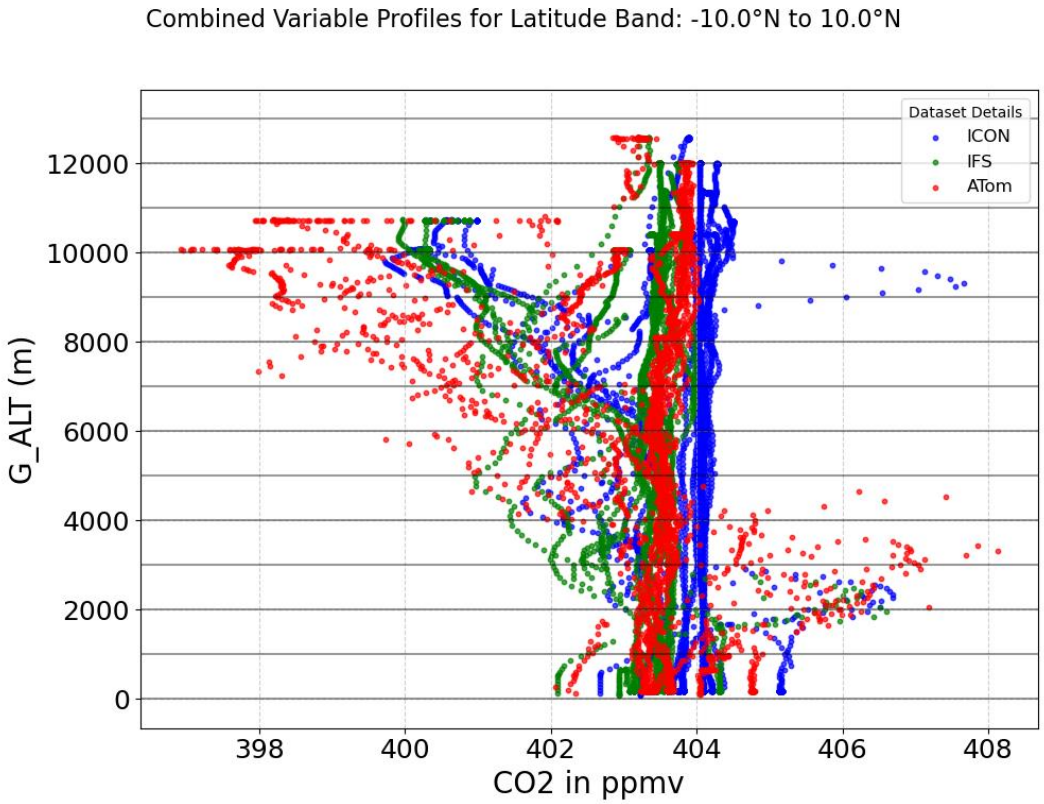


Figure 6: Profiles for CO2 (CO2.X in the case of ATom) for the latitude band between 10°S and 10°N for ICON (blue), IFS (green) and ATom1 (red)

When comparing the ICON model to the combined CO_2 observational product from ATom1 (Figure 7), ICON exhibited a slightly larger average positive difference than the IFS model. This suggests that, on average, ICON tends to overestimate CO2 concentrations more than IFS when compared to this specific combined dataset. This behaviour is strange as the same fluxes were used in ICON and IFS. Also tests done in D1.1 did not reveal any problems with mass balance due to transport in ICON. The reason is unknown and currently under investigation.

The correlation coefficient for ICON with the combined CO_2 product was very similar to that of IFS, indicating that both models capture the observed CO_2 variability to a comparable degree.

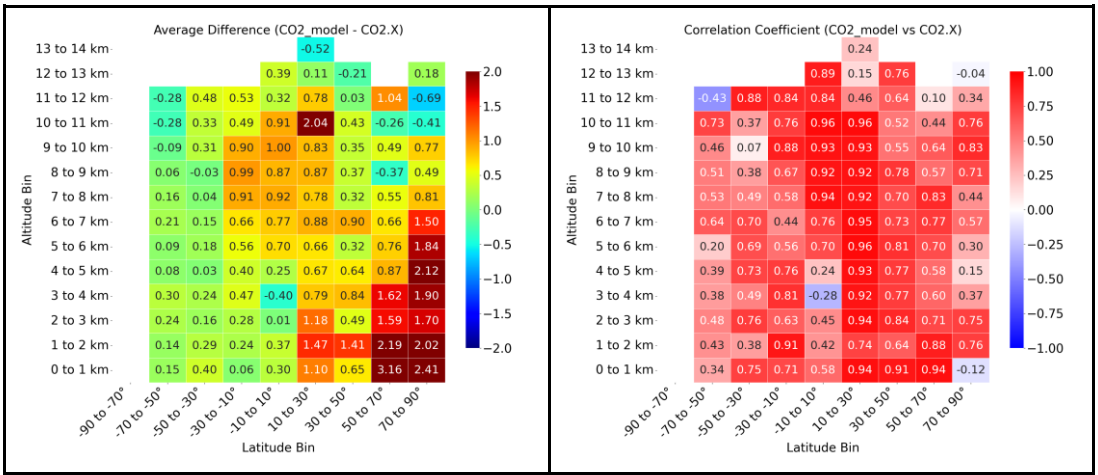


Figure 7: Comparison of ICON to the combined CO2 product of ATom1. Left: Average Difference in ppmv; Right: Correlation coefficient

The evaluation of SF6, a long-lived tracer, provides further insights into the models' transport characteristics.

For the IFS model (Figure 8), the average difference for SF6 compared to ATom1 data was relatively small, suggesting good overall agreement in mean concentrations. However, the correlation coefficient for SF6 is significantly lower than that observed for CO2.

The ICON model (Figure 8) also exhibited a small average difference for SF6, although it was higher than that of the IFS model. As for CO2 the reason is unknown and currently under investigation. Its correlation coefficient for SF6 was similar to that of IFS. The highest differences for ICON were observed in the northern high latitudes, where its correlation coefficient became negative while that of IFS remained positive.

The patterns of the average difference for SF6 in both models are broadly similar to their respective CO_2 patterns. This similarity suggests that the models' ability to simulate the transport of these two tracers is roughly equivalent at the scales resolved by ATom. However, the correlation coefficients for SF6 and CO_2 differ quite significantly for both models. This discrepancy is unlikely to be solely due to transport issues, as SF6 and CO_2 should be transported similarly. The precise reason for the worse correlation coefficient of SF6 is currently unknown and will be further investigated. The different emission pattern (e.g. no sinks) most likely contribute to a more homogenous contribution which will result in a worse correlation coefficient even for small deviations.

The CO_2/SF6 ratio (Figure 9) offers insights into the relative transport and distribution of these gases, which possess different source/sink profiles. For ATom1, both IFS and ICON showed moderate correlation coefficients for this ratio. Generally, the correlation coefficient increases with altitude. This suggests that vertical and long-range transport are captured quite well in the models, whereas the lower troposphere is more influenced by the interplay of sources, sinks, and differential transport pathways within the boundary layer. Distinguishing between transport issues in the boundary layer and problems with sources and sinks remains a challenge. The patterns for ICON and IFS are very similar in this case. The current guess for this similarity is the same meteorology/transport. This behavior is currently under investigation.

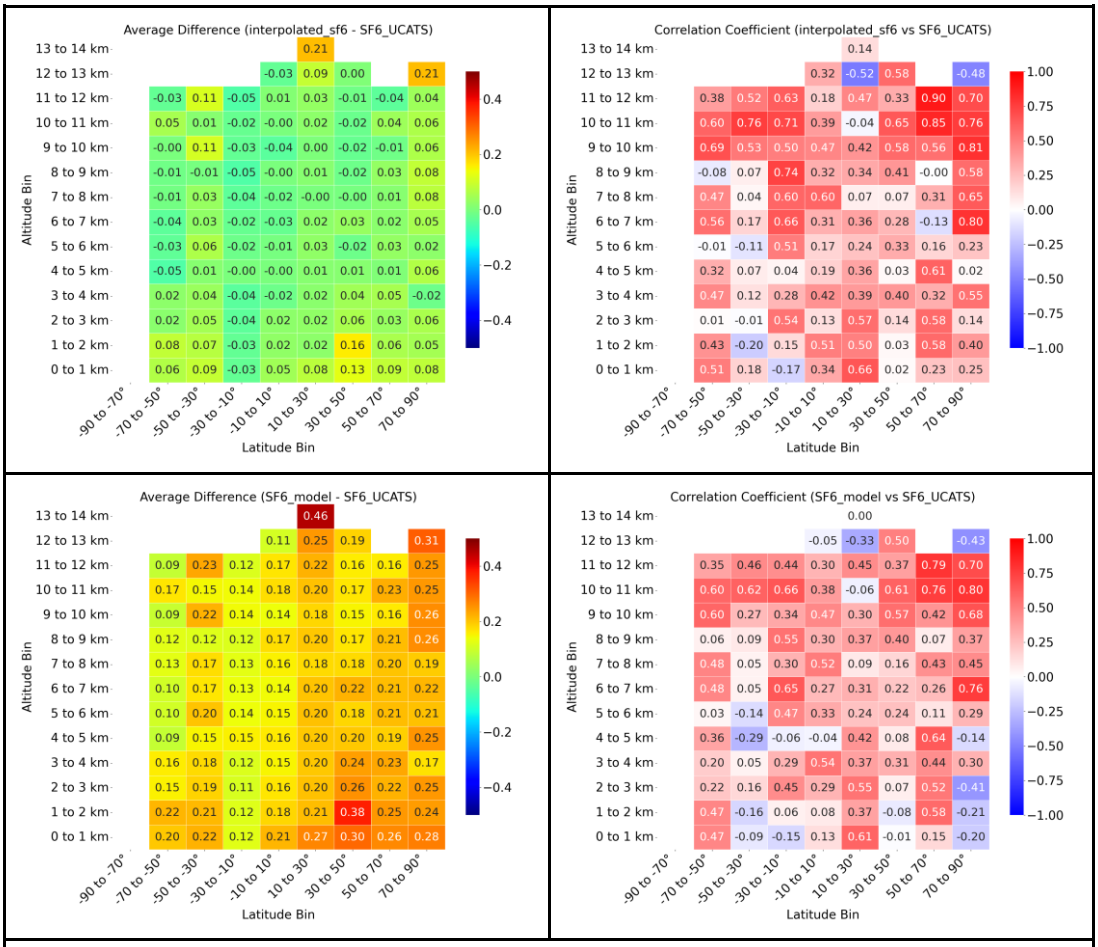


Figure 8: Same as figure 6, but for SF6 and IFS (top row) and ICON (bottom row)

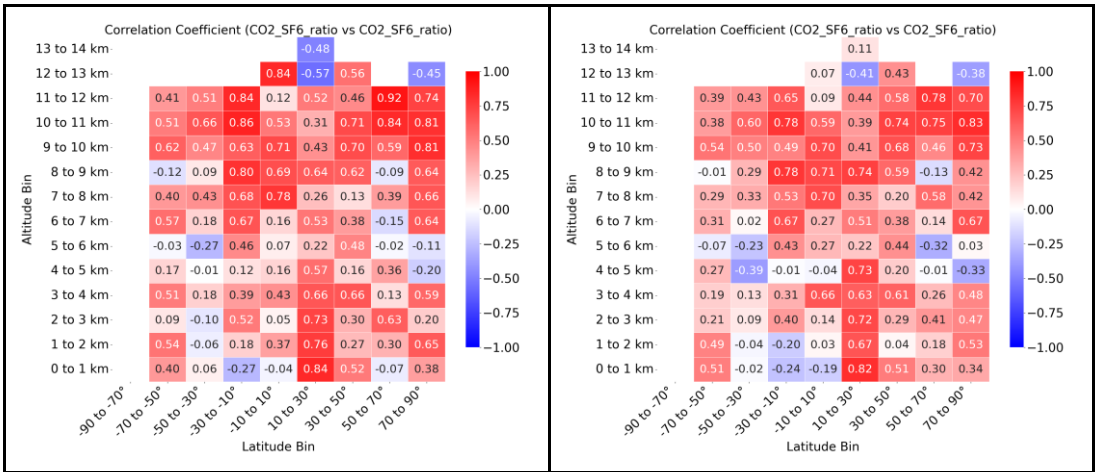


Figure 9: Correlation coefficients of the ratio CO2/SF6 for IFS to ATom1 (left) and for ICON to ATom1 (right)

In conclusion, the ATom data provides a robust benchmark for global model evaluation. Both IFS and ICON show considerable skill in simulating the global distribution of CO2 and SF6. However, differences in biases and the somewhat weaker performance for tracer ratios suggest ongoing challenges in correctly representing the interplay of different transport

processes, including vertical exchange between the boundary layer, free troposphere, and the Upper Troposphere Lower Stratosphere (UTLS) region.

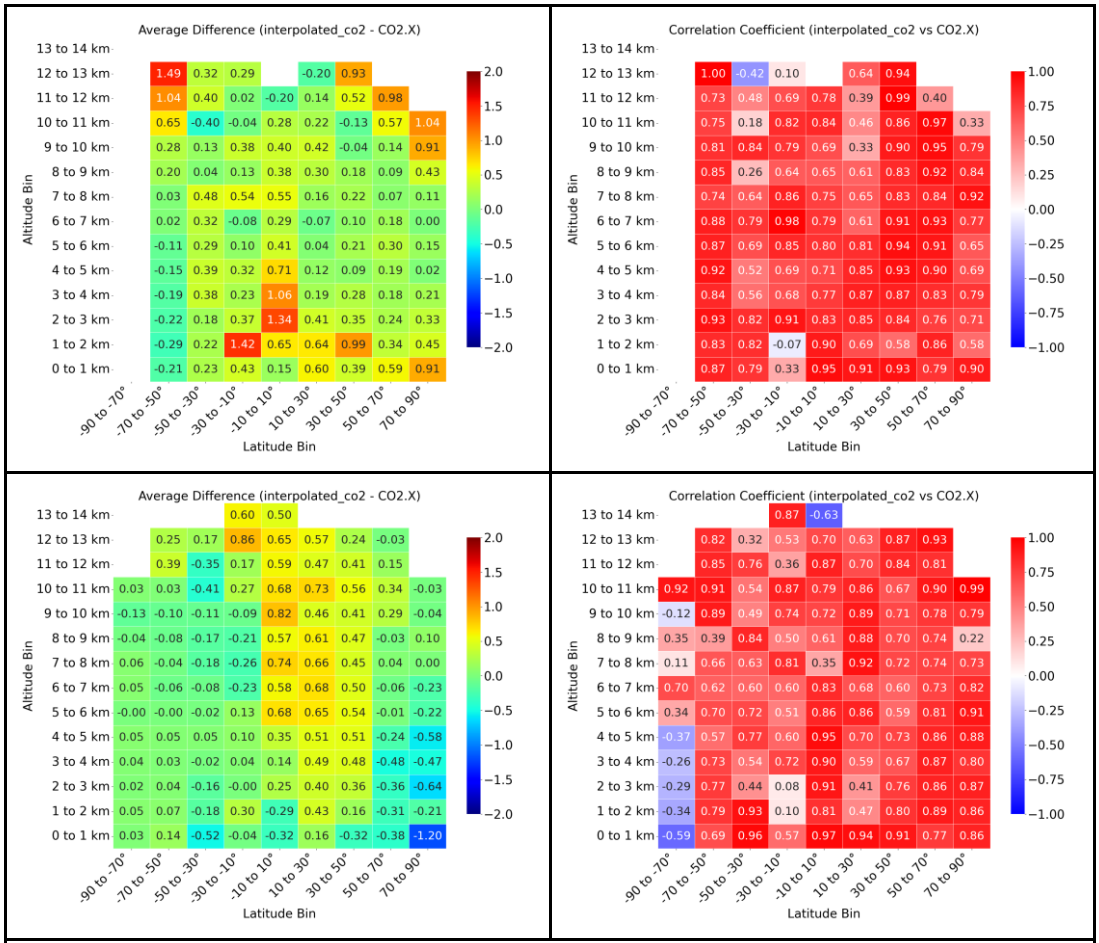


Figure 10: CO2 Average Difference in ppmv (left) and Correlation coefficients (right) to combined CO2 for IFS. Top for ATom2 and bottom for ATom3

The comparison of CO_2 simulations for ATom2 and ATom3 (Figure 10) reveals slightly different results for the IFS model, particularly regarding the bias in the boundary layer. For ATom2, the differences are now consistently positive across all regions except for the high southern latitudes, and the magnitude of the bias in high northern latitudes has decreased. In contrast, for ATom3, IFS shows smaller CO2 mixing ratios everywhere except for the high southern latitudes. This change in the sign of the bias indicates a seasonal dependency. The generally lower values close to the surface might suggest issues with emissions rather than transport.

For both ATom1 (August) and ATom2 (February), IFS exhibits a higher bias at the upper altitudes of the observations at the respective winter pole compared to the summer pole. This pattern might indicate the influence of subsiding stratospheric air, suggesting that this bias is most likely attributable to problems in the stratosphere and/or stratosphere-troposphere exchange.

The correlation coefficients for ATom2 appear similar to those for ATom1. However, for ATom3, a significant difference is observed in the high southern latitudes, where the correlation coefficient turned negative. This is caused by almost no vertical structure in the CO2 profile (not shown) besides almost perfect agreement.

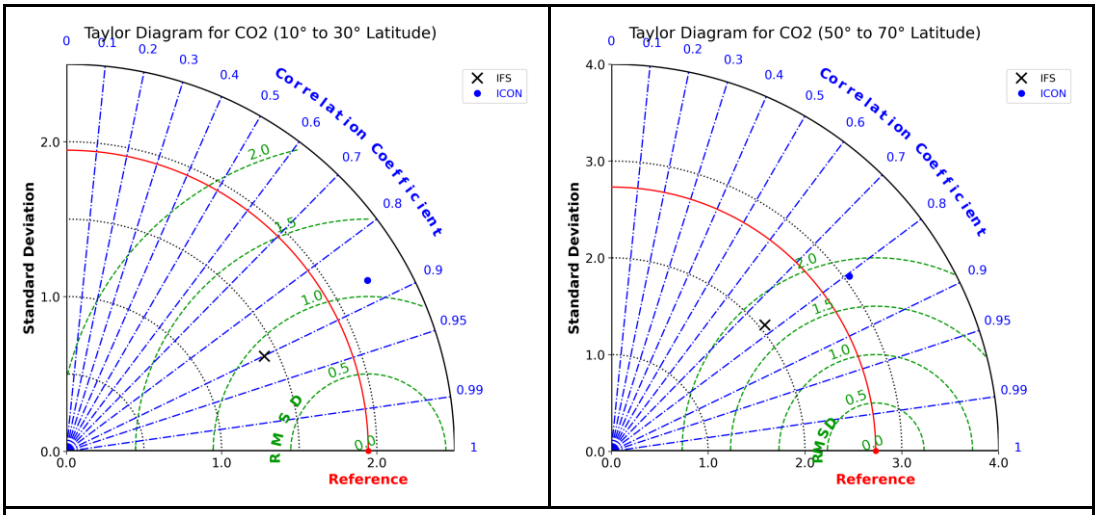


Figure 11: Taylor diagrams for CO2 in the latitudinal bin 10 to 30 degrees (left) and 50 to 70 degrees (right) for ATom1

The comparison of Taylor diagrams (see Figure 11 for two examples) reveals distinct differences between the IFS and ICON models.

ICON consistently exhibits a higher standard deviation (indicating greater variability) than IFS across all latitudinal bins. In most bins, ICON overestimates the observed variability, whereas IFS generally underestimates it. Furthermore, the root mean square deviation (RMSD) from the observations is almost always higher for ICON than for IFS.

The correlation coefficient, which indicates the fidelity of the vertical structure, is generally good for both ICON and IFS in most cases. However, there are notable exceptions. For instance, in the -50 to -30 degree latitudinal bin, both models perform less effectively, with ICON's correlation coefficient being particularly low, around 0.3. In this same bin, ICON's RMSD and standard deviation are also considerably off, while IFS performs quite well. This strongly suggests significant issues within the ICON model for this specific latitude range.

Another bin with a notably low correlation is the -70 to -50 degree range. Generally, the correlation coefficient is lower across the Southern Hemisphere, indicating greater challenges in accurately representing the vertical structure of atmospheric components in this region. Looking at the profiles (see Figure 12) indicate a lot of different small deviations in the vertical structure. But those also can be found in other latitudes. In combination with an in general flatter profile on high southern latitudes this leads to a worse correlation coefficient.

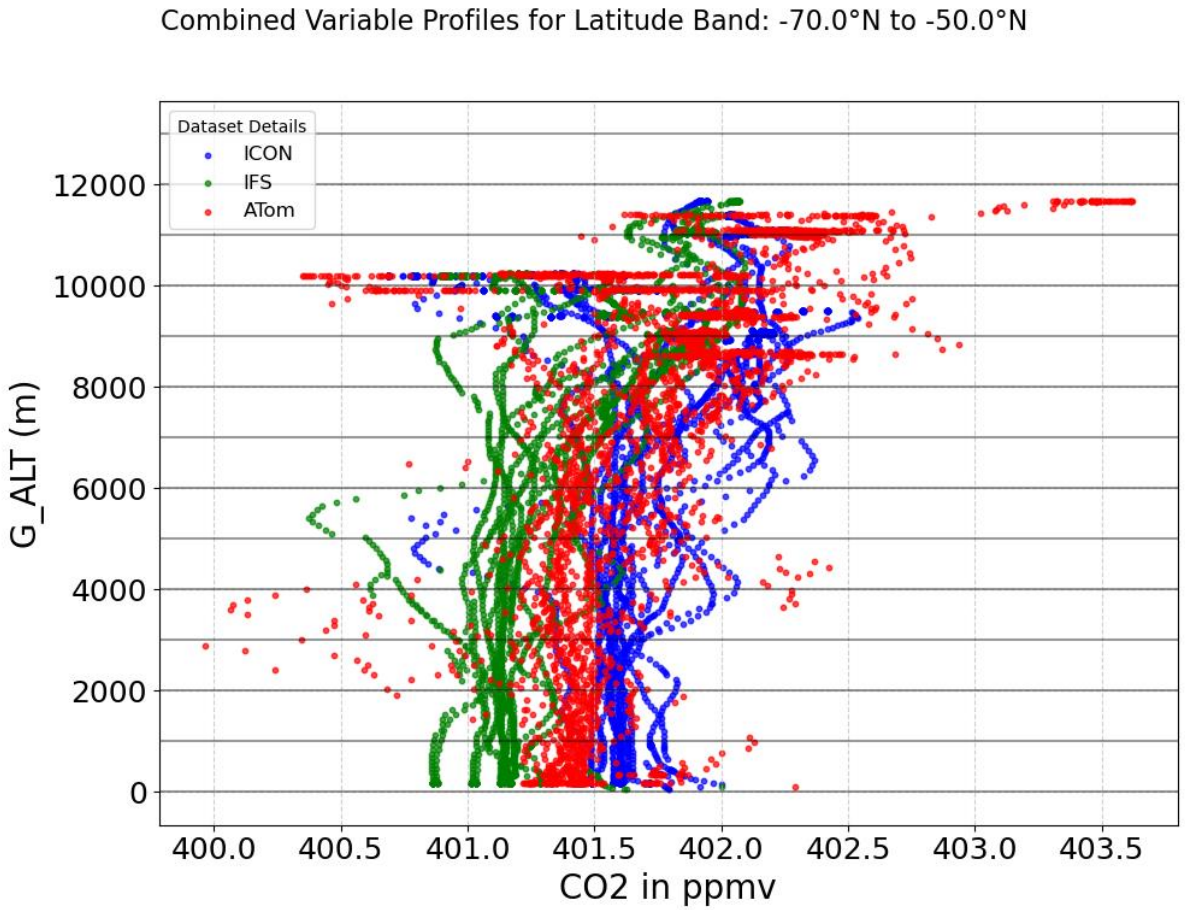


Figure 12: Profiles for CO2 (CO2.X in the case of ATom) for the latitude band between 70°S and 50°S for ICON (blue), IFS (green) and ATom1 (red)

2.4.2 DCOTSS

The Dynamics and Chemistry of the Summer Stratosphere (DCOTSS) is a NASA Earth Venture Suborbital research project dedicated to investigating the impacts of intense thunderstorms over the U.S. on the summertime stratosphere. Strong convective storms during the summer in North America can overshoot the tropopause, injecting water and pollutants from the troposphere into the typically dry stratosphere. This transport can significantly influence radiative and chemical processes, including stratospheric ozone.

During the summers of 2021 and 2022, DCOTSS utilized the NASA ER-2 high-altitude research aircraft to conduct 25 research flights. These flights, based in Salina, KS, and Palmdale, CA, were strategically located for sampling convective plumes in the stratosphere. The ER-2 aircraft is equipped with an extensive suite of instruments for measuring trace gases and aerosol properties and can operate at altitudes up to 70,000 feet (~ 21 km). The DCOTSS team successfully intercepted outflow plumes from overshooting storms.

The DCOTSS-Aircraft-Data product features data collected by various instruments onboard the NASA ER-2 aircraft. For this testbed we used the Harvard University Picarro Cavity Ringdown Spectrometer (HUPCRS). This instrument measured the carbon tracers CO2, CO and CH4.

In the beginning of the project, various skill metrics were calculated for the DCOTSS (Dynamics and Chemistry of the Overshooting Tropopause Sampling Strategy) campaign. This comparison was specifically performed against CO2 observations collected during the

DCOTSS flight campaign. It is important to note that the model runs from ICON and IFS used for this preliminary analysis were subsequently improved for the other results presented in this report. The skill metrics evaluated were Spearman's rank correlation coefficient, Kling-Gupta Efficiency (KGE), Nash-Sutcliffe Efficiency (NSE), and Normalized Root Mean Square Error (NRMSE).

Example Results for the 2022/05/31 Overshooting Event (CO2 Vertical Profiles):

For the 2022/05/31 overshooting event, the following preliminary skill metrics were obtained when comparing model outputs to DCOTSS CO2 vertical profile observations:

Table 2: Different skill metrics for overshooting event at 2022/05/31. For a definition of the metrics, see Section 3.2.

Model	KGE	Spearman	NSE	NRMSE
ICON	0.559	0.860	-0.533	0.347
IFS	0.632	0.642	0.680	0.159

Based on these preliminary results shown in Table 2 for CO_2 vertical profiles, IFS generally shows better performance across most metrics for this specific event. IFS exhibits a higher KGE (0.632 vs. 0.559), indicating a better overall agreement. While ICON has a notably higher Spearman correlation (0.860 vs. 0.642), suggesting a stronger monotonic relationship, its NSE is negative (-0.533), implying that the mean of the observed data would be a better predictor than the ICON model for this particular case. In contrast, IFS has a positive and relatively high NSE (0.680), indicating good predictive power. Furthermore, IFS shows a significantly lower NRMSE (0.159 vs. 0.347), signifying smaller normalized errors.

Ultimately, while these individual metrics provided valuable insights into the model's ability to reproduce CO_2 vertical profiles, for a more comprehensive and visual representation of model performance, it was decided to primarily utilize Taylor Diagrams for the final results presented in this report. Taylor Diagrams effectively summarize the correlation, RMS error, and standard deviation of simulated fields in comparison to observations.

In the analysis we focused on flights during 2022-05-26, 2022-05-29, 2022-05-31, and 2022-06-02, with a specific emphasis on the flight on 2022-05-31, during which the aircraft sampled air influenced by a strong overshooting convective event over western Oklahoma. On 2022-05-31, a significant overshooting convective system developed over Oklahoma began near 22:00 UTC on 31 May 2022 and dissipated near 06:30 UTC on 1 June 2022, leading to deep convection that penetrated the tropopause.

Comparing the Taylor diagrams for CO_2 mixing ratios across the selected days (Figure 12), we see that both ICON-ART and IFS capture some level of correlation with the DCOTSS observations, typically ranging from around 0.8 to 0.95 for ICON (12min) and often slightly lower or comparable for IFS. On most days, ICON (12min) has a standard deviation closer to the reference (DCOTSS, implicitly at the (1,0) point) and a lower RMSD (Root Mean Square Difference) than IFS (e.g., 20220529, 20220531, 20220602), indicating better agreement in terms of variability and overall difference. The comparison between ICON (12min) and ICON_3hr output demonstrates the importance of temporal resolution. While the differences in the Taylor diagrams are not always dramatic, ICON (12min) appears to be slightly better

positioned relative to the reference on some days (e.g., 20220531, 20220602), implying that the higher frequency output from ICON-ART captures more of the observed temporal variability, leading to improved statistical metrics in some cases. However, the 3-hour output retains a fair amount of skill. Therefore, for future comparison we decided to not use such high output frequencies as the results are only slightly better and lead to too much data.

Examination of the CO_2 vertical profiles indicates the models' ability to reproduce the observed vertical structure. DCOTSS profiles demonstrate the expected decline in CO_2 mixing ratios with altitude, especially at higher altitudes where stratospheric air has lower CO_2 concentrations. However, the DCOTSS profiles show complicated layering and inversions on different days, indicating different air masses and mixing processes. Both ICON-ART and IFS generally reproduce the overall pattern of decreasing CO_2 with height. ICON-ART's profile frequently follows the observed structures more closely than IFS, capturing some of the shape variability. On 2022-05-31, the DCOTSS profile exhibits significant structure between 16 km and 19 km, most likely due to sampling air impacted by overshooting convection. ICON-ART appears to capture some, but not all, of the complexity, whereas IFS has a smoother profile.

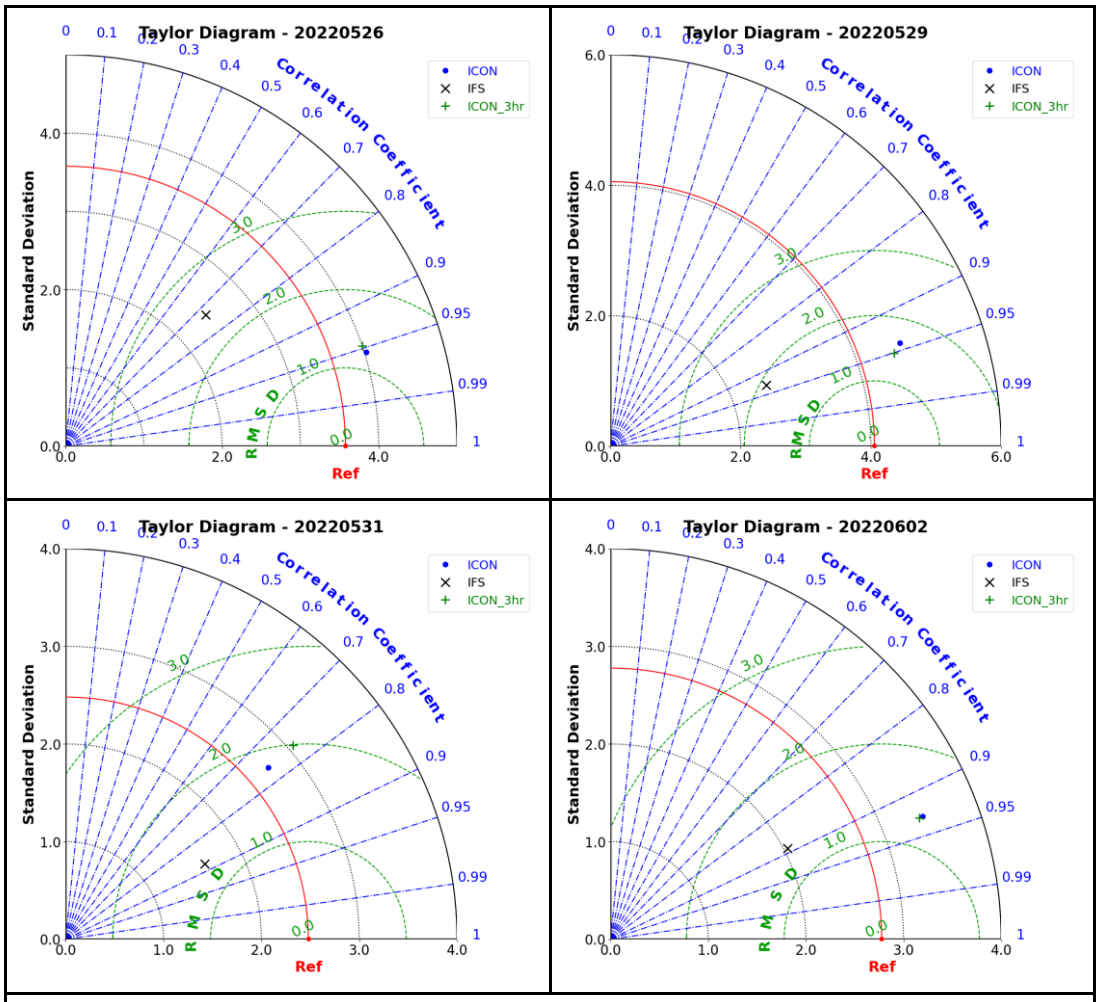


Figure 13: Taylor Diagrams comparing CO2 model outputs against DCOTSS data for selected dates

Vertical gradients provide direct indications of the strength and position of vertical mixing and transport barriers (Figure 14). The vertical gradient plots for CO_2, CH_4, and CO indicate that DCOTSS observations have a wide range of gradients at different elevations, indicating strong vertical structure and mixing activity. Both ICON-ART and IFS tend to underestimate the range of observed gradients, with model points clustering closer to zero than DCOTSS points. This

implies that the models result in smoother vertical profiles than observed, thereby underestimating the severity of abrupt boundaries or localized mixing events. On 2022-05-31, the CO_2 vertical gradient plot for DCOTSS shows significant positive gradients (CO2 increasing with height over a layer) in the altitude range of 16-19 km, which could be associated with sampling air masses with different CO_2 signatures due to convective transport (e.g., higher CO2 tropospheric air above lower CO_2 stratospheric air or mixed layers). Neither ICON-ART nor IFS properly reproduce the amplitude and distribution of the observed positive gradients on this particular day, which is crucial for assessing their representation of convective vertical transport into the UTLS.

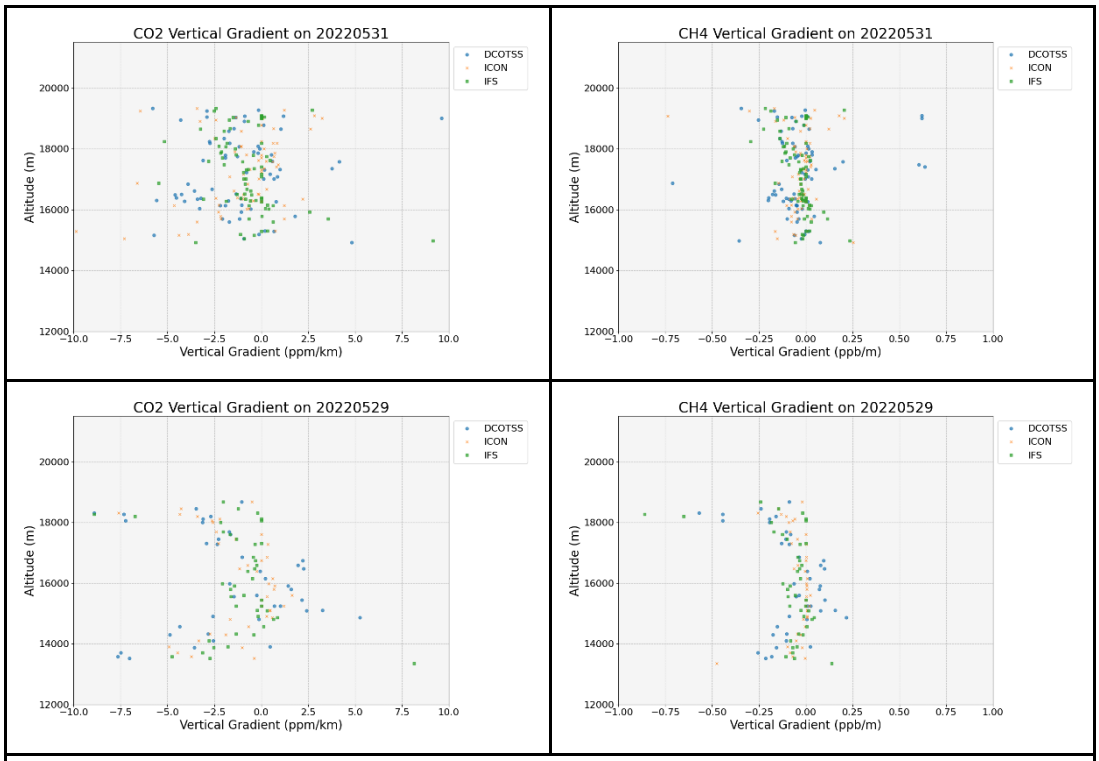


Figure 14: Vertical gradients plots of the CO2 and CH4 in the days 2022-05-29 and 2022-05-31

Normalized CO_2 time series plots (not shown here) show the temporal variability observed and simulated during the flights. DCOTSS observations vary considerably over time, indicating interactions with distinct air masses through regions with significant composition changes. While ICON-ART and IFS capture some of the broader temporal trends and overall shifts in normalized CO_2, they frequently fail to replicate the entire magnitude or precise timing of the quick fluctuations and sudden transitions observed in the DCOTSS data. ICON-ART (12-minute output) appears to capture more short-term fluctuation than IFS, which is smoother. The DCOTSS normalized time series for 2022-05-31 exhibits complicated changes, most likely representing the aircraft's path through the convective system and interactions with overshooting air parcels. Models capture some variability over this time period, but they may miss the fine-scale or rapid shifts associated with moving through heterogeneous air masses caused by deep convection.

In summary, the DCOTSS observations, particularly for the 2022-05-31 overshooting event, show compelling evidence of complex vertical structure and temporal variability in CO_2 mixing ratios in the UTLS, most likely caused by convective transport. The occurrence of non-zero

vertical gradients, particularly positive CO_2 gradients at high elevations, lends credence to the notion that the event involved extensive vertical mixing or layering. While both ICON-ART and IFS excel in reproducing the broad vertical structure and statistical characteristics (as seen in Taylor diagrams), they appear to struggle to reflect the complete complexity of vertical transport, particularly during occurrences such as overshooting convection. This is seen by the smoother vertical profiles, the underestimating of the range of vertical gradients, and the lower temporal variability as compared to DCOTSS. The models may fail to adequately resolve the turbulent mixing and fine-scale structures formed by deep convection, as well as accurately replicate air mass injection and mixing across the tropopause. Another cause can be the model setup and will be investigated by changing the way nudging is done and also by changing vertical resolution in ICON. For IFS we will also try a higher output frequency for this specific event.

2.4.3 IAGOS

The In-service Aircraft for a Global Observing System (IAGOS) project delivers a comprehensive, time and spatially resolved multi-component dataset focusing on Essential Climate Variables (ECVs) and Air Pollutants. This data provides crucial information regarding the distribution and long-term changes within the troposphere and lower stratosphere, as well as regular vertical profiles collected over major cities.

IAGOS operates in two main modes: IAGOS-CORE and IAGOS-CARIBIC. IAGOS-CORE involves continuous measurements of trace gases, aerosols, and cloud particles from a fleet of long-haul passenger aircraft, with approximately 500 flights per aircraft per year. Each IAGOS-CORE aircraft is equipped with a dedicated rack for automated instruments (Package 1) that in the scope of the CATRINE project measures CO. A second instrument package (Package 2) can be installed, with options for greenhouse gases (CO_2 and CH_4). Only one Package 2 option can be installed on a given aircraft at a time.

IAGOS-CARIBIC operates on a single aircraft, conducting 40-50 flights per year. The data from IAGOS-CORE and IAGOS-CARIBIC, along with data from precursor projects MOZAIC and CARIBIC, are stored in the IAGOS Data Base. Near real-time data are also provided to operational users, such as the COPERNICUS Atmosphere Monitoring Service (CAMS), via the WMO Information System (WIS).

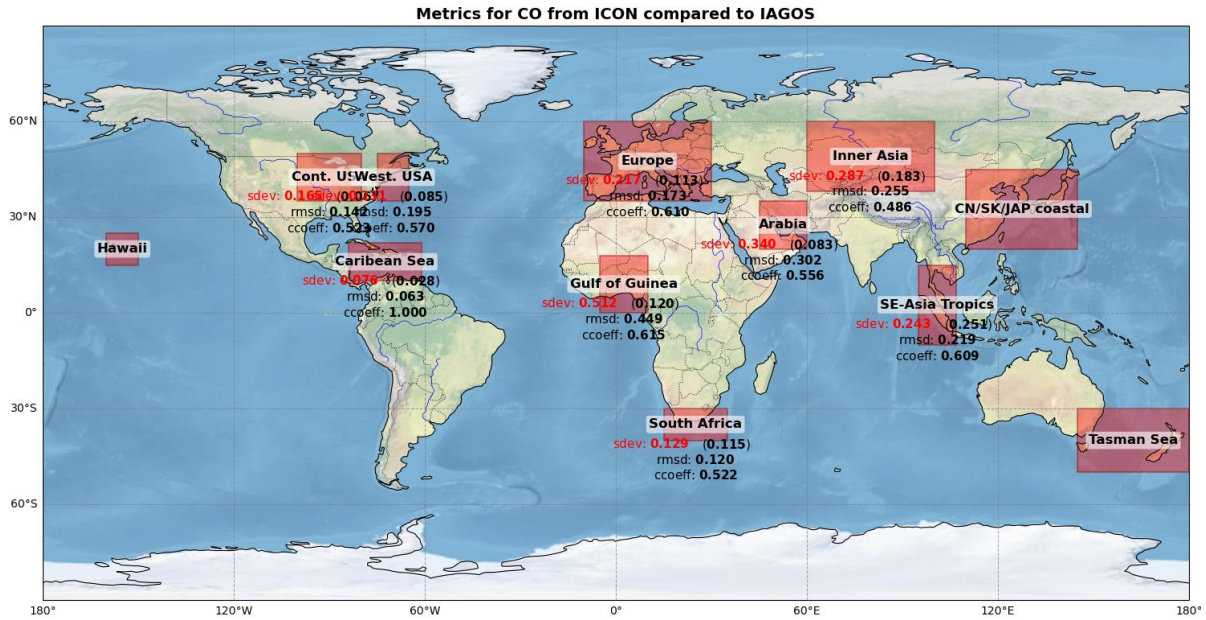


Figure 15: Regions selected for IAGOS with standard deviation from the observations and ICON (top row), root mean square deviation (middle row), and correlation coefficient (bottom

row) for each region for the tracer CO. Metrics are for all ascents and descents that fall into each region in the time frame 2022/05/18 to 2022/06/05.

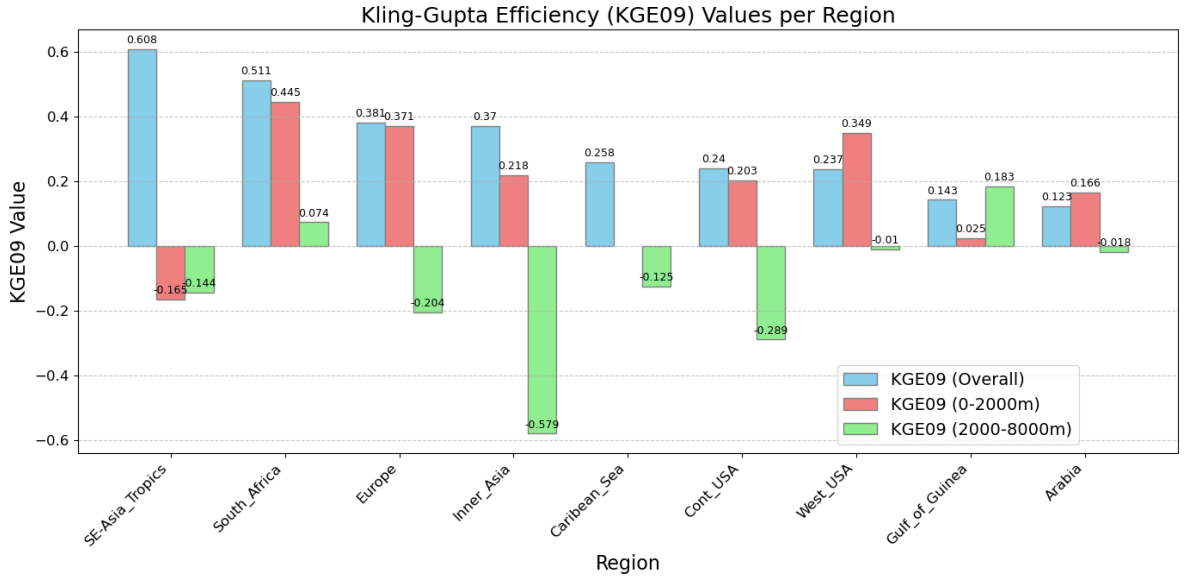


Figure 16: Kling-Gupta Efficiency per region for CO for the ICON model.

The IAGOS testbed provided an opportunity to assess model performance across diverse global regions. For this analysis, we utilized the same 12-minute model output as for the DCOTSS campaign. During the selected time period, only CO observations were available from IAGOS among the gases CO_2, CH_4, and CO.

We defined several distinct geographical regions (see Figure 15) for statistical analysis, focusing exclusively on ascent and descent profiles provided by IAGOS. This allowed us to examine the complete vertical structure of the atmospheric profiles. At this stage, the comparison was performed solely with the ICON model. It's important to note that not all defined regions had IAGOS flights within the chosen time period.

Instead of directly comparing CO mixing ratios, we used normalized CO (divided by the mean of the CO) profile. This normalization was necessary because both IFS and ICON models exhibited a significant positive bias in their absolute CO values. This bias is scheduled for improvement in the next phase of the CATRINE project. We averaged all available profiles within each defined region.

For this analysis, we compared quantities typically displayed in Taylor diagrams: standard deviation, root mean square deviation (RMSD), and correlation coefficient. Additionally, we calculated the Kling-Gupta Efficiency (KGE) for the complete vertical profile, and separately for the altitude ranges of 0 to 2000m and 2000m to 8000m for each region (see Figure 16). For definitions of these metrics, see section 3.2.

When assessing the variability (standard deviation, 'sdev' in Figure 15), we observed a consistent underestimation in the ICON-modelled variability compared to observations for the Northern Hemisphere regions. In contrast, for the South Africa and SE-Asia Tropics regions, the model and observational variability showed good agreement.

The RMSD of the normalized profiles was quite high across all regions, and notably exceptionally high for the Gulf of Guinea region. Given the current state of the model and the known bias in absolute values, this high RMSD should not be overemphasized.

The correlation coefficient, serving as a good indicator for the vertical gradient, showed mediocre agreement across all regions. The lowest agreement was found in the two entirely continental regions: Continental USA and Inner Asia.

Interestingly, the Kling-Gupta Efficiency (KGE) for the complete altitude range often appears to be reasonable. For instance, the SE-Asia Tropics region showed the highest KGE values among all considered regions for the complete profile. However, when looking at the KGE values for the segmented altitude ranges (0-2000m and 2000-8000m), the picture changes considerably: both KGE values are negative for the SE-Asia Tropics region. This pattern holds true for almost all regions, where the KGE for the altitude ranges consistently looks worse than for the complete profiles. This behavior will be further investigated in future work, and the IFS model will be included in the analysis.

2.4.4 STRATOCLIM

The Stratospheric and upper tropospheric processes for better climate predictions (STRATOCLIM) campaign focuses on detailed observations of atmospheric transport and physical-chemical processes that govern the input of air and aerosols into the (sub-)tropical stratosphere. The primary target of this mission is the Asian monsoon anticyclonic circulation, which, with its defined boundaries and outflow patterns, along with significantly polluted inflow, provides an ideal atmospheric laboratory for the M55 Geophysica aircraft payload, considering its operational range and capabilities, and the traceability of tropospheric pollutant signals through the Upper Troposphere (UT) and Tropical Tropopause Layer (TTL) into the stratosphere. The aircraft campaign was divided into two parts: four flights from Kalamata (Greece) in August/September 2016, and the main part in South Asia in 2017.

Several instruments were part of the STRATOCLIM campaign. We use HAGAR (High Altitude Gas AnalyzeR) which measures CH4, CO2, and SF6, and AMICA (Airborne Mid-Infrared CAvity) enhanced spectrometer which measures CO and CO2.

Chosen date for analysis are flights on July 29, 2017 (Flight 2) and August 2, 2017 (Flight 4). Examination of the vertical profiles of CO_2 mixing ratio versus altitude for both flights reveals fundamental characteristics of the observed and modelled distributions. STRATOCLIM data consistently show aa increase in CO_2 with increasing altitude throughout the troposphere (due to biogenic sink during NH summer), with values approaching typical lower stratospheric levels at higher altitudes. Comparing this to the models, the ICON-ART simulation generally exhibits a positive bias, simulating higher CO_2 mixing ratios than observed across most altitude ranges for both flights. IFS model also shows a positive bias, but simulating lower CO_2 mixing ratios compared to ICON, particularly in the troposphere. These biases suggest differences in the models' representation of large-scale transport, boundary conditions, or the global/regional CO_2 budget. Further detailing the vertical structure, plots of the CO_2 vertical gradient (dCO_2/dz) indicate that both ICON and IFS models tend to simulate weaker negative gradients in the troposphere compared to the steeper gradients observed by Stratoclim. This weaker gradient suggests that the models may not fully capture the strength of atmospheric stratification or that they are overly diffusive vertically, potentially leading to a smoothing of sharp vertical transitions present in the real atmosphere.

Taylor diagrams provide a statistical summary of model performance relative to Stratoclim data for specific flight phases (Figure 17).

These diagrams highlight differences in correlation, centered RMSE, and standard deviation for the ascent, stable, and descent periods. During the stable, high-altitude phase (primarily in the lower stratosphere), both models demonstrate strong statistical agreement with observations. This suggests that both models represent the structure of the CO_2 field at these altitudes well. But also, both are showing weaknesses in the RMSD and standard deviation. This suggests that both models are showing wrong amplitudes for the structures. However, during the dynamic phases of ascent and descent, which involve profiling through different

atmospheric layers and challenging the models' representation of vertical structure and mixing, performance metrics diverge more notably. Correlation coefficients are generally lower during ascent compared to stable phases, indicating greater difficulty in precisely matching the vertical structure.

While both models perform well statistically in the stable high-altitude phase, their differences become more apparent during the ascent and descent phases, where the accurate representation of vertical structure and transport becomes critical. IFS's overall better statistical performance during the dynamic phases, particularly its closer match to observed variability and lower RMSE, suggests a potentially more realistic representation of the combined effects of vertical mixing, stratification, and potentially horizontal advection encountered during profiling compared to ICON.

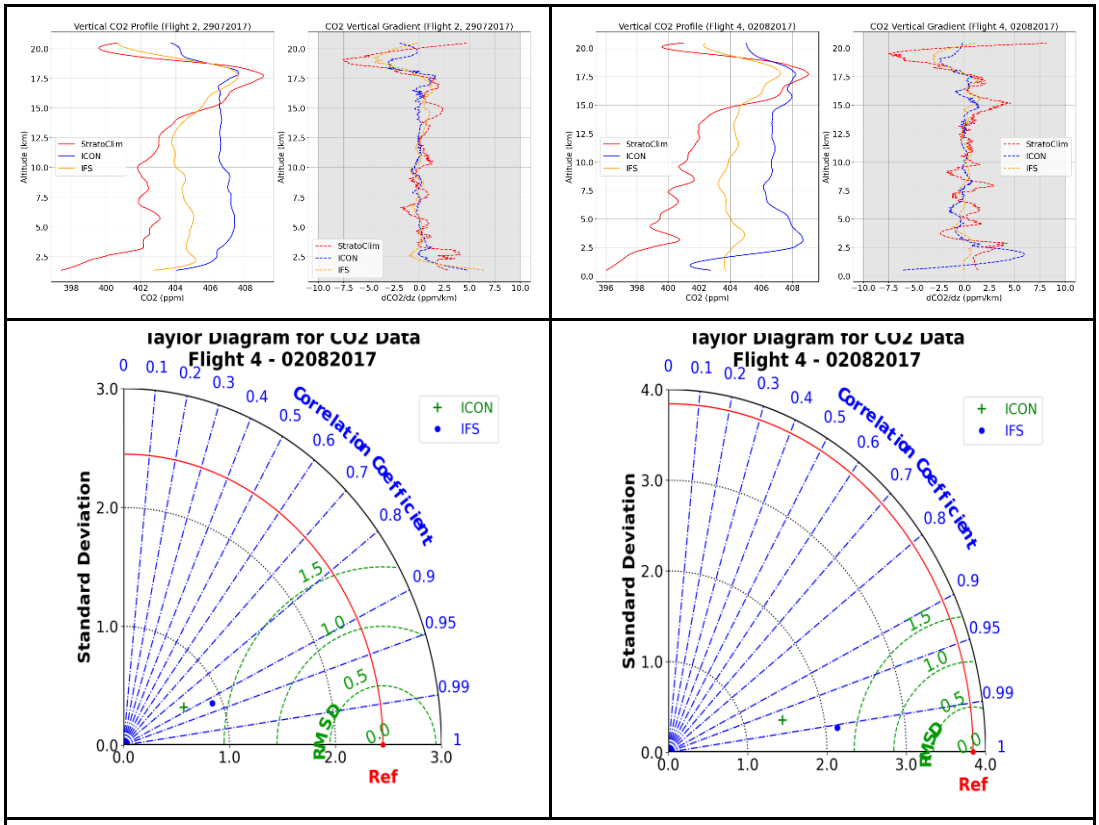


Figure 17: Vertical profile gradients plots of the CO2 in the days 2017-07-29 (Flight 2) and 2017-08-02 (Flight 4) (top), and Taylor diagrams for Flight 4 for Ascent (left) and “constant” flight level in 2017-08-02 (bottom)

2.4.5 WISE

The WISE (Wave-driven Isentropic Exchange) campaign aims to investigate the complex interrelations between atmospheric composition and dynamical structures within the Upper Troposphere and Lower Stratosphere (UTLS). The campaign seeks to quantify the physical and chemical processes, including air mass exchange and cirrus formation, that govern UTLS composition. It specifically addresses how mixing processes at the tropopause, spanning various scales, contribute to uncertainties in radiative forcing estimates. WISE focuses on three main research topics: the interrelation of the tropopause inversion layer (TIL) and trace gas distribution, the role of planetary wave breaking in water vapor transport into the extratropical lower stratosphere, the role of halogenated substances in ozone and radiative

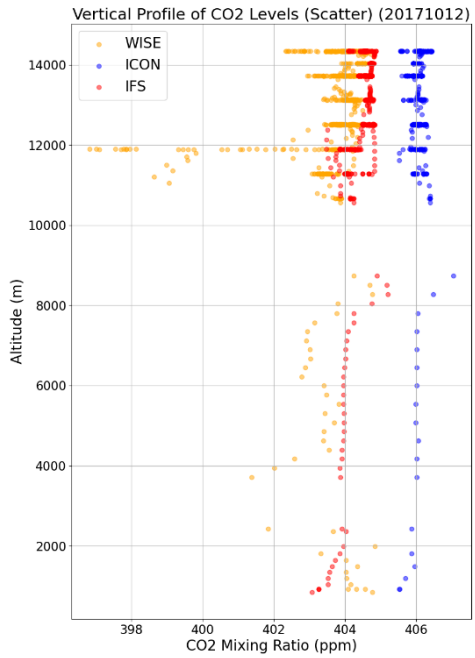
forcing in the UTLS, and the occurrence and effects of sub-visual cirrus (SVC) in the lowermost stratosphere.

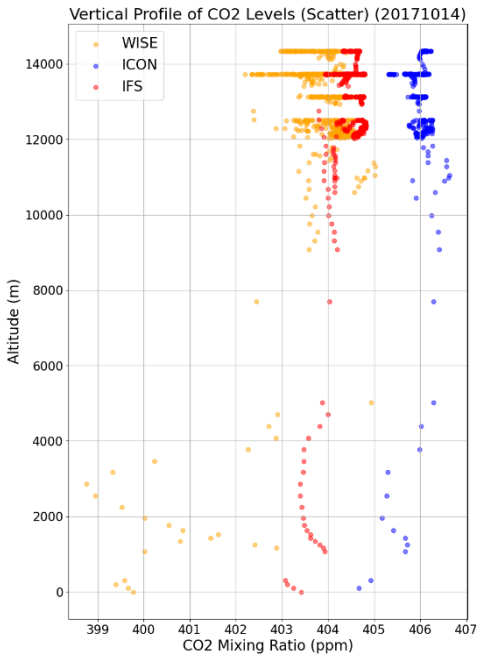
The campaign utilizes the German research aircraft HALO due to its unique capabilities, including carrying a substantial payload up to altitudes of 15.5 km, which is above the lowermost stratosphere (LMS) at mid-latitudes. This altitude range is ideal for profiling the LMS, particularly the "overworld" region significantly affected by the Asian summer monsoon. HALO's combination of in-situ and remote sensing instruments, providing high-resolution 3D measurements of temperature, static stability, various trace gases, and cirrus clouds, offers unprecedented detail and coverage for investigating vertical temperature and trace gas structures and their interactions.

The WISE campaign's location and season are strategically chosen to observe the evolution of baroclinic life cycles and Rossby wave breaking events and their role in cross-tropopause exchange, particularly over the Atlantic and North Sea.

The WISE-Aircraft-Data product features measurements collected by various instruments onboard the HALO research aircraft. We used the HAGAR-V instrument, which measured the trace gases CO_2, CH_4, and SF_6. Up to now we have only looked at CO_2.

Date chosen for analysis are 2017-10-12 and 2017-10-14, and 2017-10-15. Overall, both ICON and IFS models exhibit varying skill in reproducing the CO_2 observations from the WISE campaign. A consistent finding across dates and flight phases, as quantified by the Taylor diagrams (not shown), is the significant underestimation of observed CO_2 variability by both models. The standard deviation of model simulations is substantially lower than that of the WISE reference data, indicating a deficiency in capturing the amplitude of temporal or vertical fluctuations. E.g. IFS and ICON do not capture the decreasing mixing ratio of CO_2 between 2km and 4km very well. Furthermore, both models generally show a positive bias, meaning they tend to overestimate CO_2 mixing ratios compared to the WISE measurements. The consistent underestimation of variability across all phases points to a potential smoothing effect in the model simulations, likely related to the representation of turbulent mixing or other sub-grid scale processes.





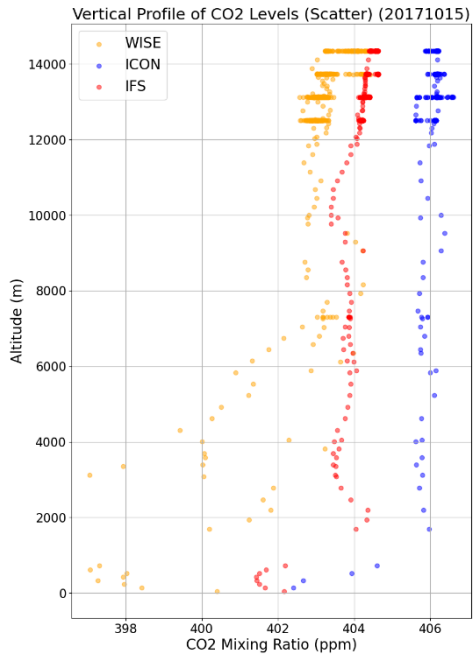


Figure 18: Vertical profile gradients plots of the CO2 in the days 2017-10-12, 2017-10-14, and 2017-10-15

The vertical profiles of CO_2 (mixing ratio vs. altitude Figure 18) clearly illustrate the models' performance in capturing the vertical structure. WISE observations reveal significant vertical

variability and structure, often showing decreasing CO_2 with altitude (negative gradients) or distinct layering. In stark contrast, the ICON model consistently produces a remarkably flat vertical profile across most altitudes for all analyzed dates (2017-10-12, 10-14, 10-15), hovering around 406 ppm and capturing virtually none of the observed vertical variability or structure. This is further emphasized by the vertical gradient plots, where ICON gradients are almost exclusively near zero, completely failing to reproduce the magnitude and distribution of the observed positive and negative gradients. The IFS model shows better performance in the vertical domain compared to ICON. While also exhibiting a positive bias relative to WISE, especially at higher altitudes, IFS does capture some aspects of the vertical structure, showing a tendency for decreasing CO_2 with height at lower to mid-altitudes (up to ~8000m) similar to WISE, particularly on 2017-10-12, 10-14, and 10-15. However, IFS generally underestimates the magnitude of the vertical gradients observed by WISE and tends to flatten out at higher altitudes, missing some of the fine-scale layering or variability. The vertical gradient plots for IFS reflect this, showing gradients that are closer to zero than the more extreme positive and negative values seen in WISE, especially at higher altitudes.

The WISE campaign took place during the Asian Summer Monsoon. Whether the models in general have problems to simulate CO_2 during such conditions will be investigated by using data from the PHILEAS campaign, for which the data unfortunately did not arrive in time for this report.

2.5 Deviations and counter measures

With respect to the two pilot BL-testbed cases—targeting the coupling between the atmospheric boundary layer and the free troposphere in contrasting ecosystems—the main scientific and technical goals have been successfully achieved. Specifically, we have demonstrated the feasibility of integrating intensive field campaigns, targeted numerical experiments, and model evaluation workflows within both rainforest and temperate forest environments. These efforts have yielded valuable insights into the vertical and horizontal transport of greenhouse gases, and the role of surface heterogeneity in modulating these processes. Due to the large amount of variables and processes, this is work in progress and we will give the full output after the workshop 2nd-3rd July 2025. While the comprehensive analysis of long-term datasets—each spanning over a year—collected at the two supersites is still in progress, preliminary results are promising. These datasets are rich and multidimensional, enabling the diagnosis of seasonal variability, diurnal cycles, and episodic events such as deep convection or stable boundary layers. At an upcoming workshop (July 2025) these results will form the basis for metrics to evaluate model output.

Our initial proposal for the UTLS tests envisioned utilizing testbeds from MAGIC, OSTRICH, and PHILEAS. Unfortunately, we did not manage to get data from MAGIC and OSTRICH. For PHILEAS, access was granted only recently, and as such, we do not yet have results to show.

To counteract these delays and expand our observational dataset, we incorporated the WISE campaign into our project. The IAGOS testbed, initially intended as an accompanying data source, has also been significantly extended to provide more comprehensive in-situ measurements. Furthermore, the inclusion of CO data from the Cafe Brazil campaign, centred around Manaus, will serve as an additional crucial countermeasure to bolster our observational data. The data was only provided very recently, and further analysis is ongoing.

Significant challenges were encountered during the execution of ICON model simulations.

For ICON high-resolution simulations, we experienced memory errors that prevented successful completion. These errors are currently under investigation by the modeling team.

In the UTLS (Upper Troposphere Lower Stratosphere) test cases, the ICON model exhibited a concerning drift in CO2 concentrations, showing an increase over the simulation period. This drift compromises the accuracy of results towards the end of the simulations. A thorough review of our emissions data, a common source of such issues, did not reveal any errors. Investigations into the root cause of this CO2 drift are ongoing.

Finally, the IFS (Integrated Forecasting System) model setup used in our simulations showed a consistent high bias in CO concentrations. As the CO and CH4 results from these IFS simulations were subsequently used as initialization for our ICON runs, this high bias was propagated into the ICON model, impacting the accuracy of CO outputs. We are actively working to understand and mitigate this systemic bias in our modeling chain. To counteract these biases, for this report the CO mixing ratios were normalized. This is the reason for not looking too much at tracer-tracer correlations for the moment. For the Atom campaign, however, we looked at SF6/CO2 ratios. Otherwise, we were calculating and investigating single tracer metrics.

3 Methods

3.1 Background Test Bed ABL-Free Troposphere

Figure 19 summarizes the research strategy employed in the testbeds to investigate the coupling between the atmospheric boundary layer (ABL) and the free troposphere, with a particular focus on the transport of greenhouse gases. The testbed integrates three core components:

Comprehensive super-sites and field campaigns, which provide high-resolution, vertically resolved observations of key meteorological and trace gas variables;
Dedicated numerical experiments, including Large-Eddy Simulations (LES), to gain process-level understanding of turbulent exchange and transport processes;
Model evaluation and improvement, focusing on parameterisations of tracer transport in weather and climate models, such as IFS and ICON-ART.

These three elements are embedded in an iterative framework (Figure 17) in which observational constraints and simulation insights are used to systematically **assess tracer transport errors**, and in turn **evaluate and improve model parameterisations**. Special focus is on the diagnosing transport errors in models, allowing us to assess how well processes such as convection, turbulence, and advection are represented across scales.

The strategy is applied to two pilot cases—**rainforest** and **temperate forest** ecosystems—chosen for their contrasting surface and atmospheric characteristics. These cases serve to:

1. **Systematically evaluate** the transport processes in models using multi-scale observations and simulations;

2. **Diagnose and quantify** key sources of error in the representation of greenhouse gas transport, both vertically (mixing, entrainment) and horizontally (advection, surface-induced heterogeneity).

This integrated approach forms the foundation for improving predictive capability of greenhouse gas distributions in numerical weather prediction and Earth system models.

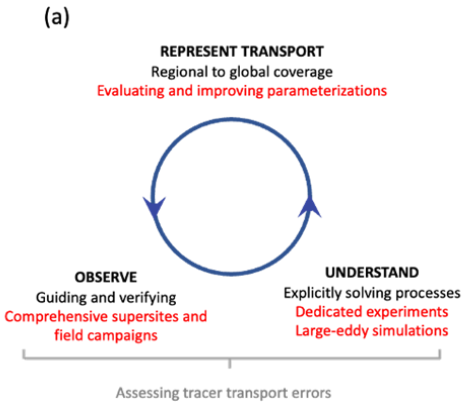


Figure 19: (a) Research strategy to integrate the observations collected during comprehensive campaigns, the understanding obtained through dedicated numerical experiments constrained by the observation, and the evaluation of the representation of the greenhouse transport as modelled by IFS and ICONART.

3.2 Metrics in UTLS testbeds

We used the following metrics within this report:

- **Spearman's Rank Correlation Coefficient (Spearman):** This non-parametric measure assesses the strength and direction of the monotonic relationship between two ranked variables. A value of 1 indicates a perfect monotonic increasing relationship, -1 a perfect monotonic decreasing relationship, and 0 no monotonic relationship. It is less sensitive to outliers than Pearson's correlation coefficient.
- **Kling-Gupta Efficiency (KGE):** KGE is a goodness-of-fit indicator widely used in the hydrologic sciences for comparing simulations to observations. It was created by hydrologic scientists Harald Kling and Hoshin Vijai Gupta with the intention to improve upon widely used metrics such as the coefficient of determination and the Nash–Sutcliffe model efficiency coefficient.
  It aims to provide a more comprehensive assessment of model performance by considering the correlation, bias, and variability between simulated and observed data. A value of 1 indicates a perfect fit, while values less than 1 suggest poorer performance.

  The formula for KGE is:

$$KGE = 1 - \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2}$$

  where: r is the Pearson correlation coefficient between the simulated and observed data, α is a term representing the variability of prediction errors, defined as:

$$\alpha = \frac{\sigma_s}{\sigma_o}$$

  where σ_s is the standard deviation of the simulated time series and σ_o is the standard deviation of the observed time series and β is a bias term, defined as:

$$\beta = \frac{\mu_s}{\mu_o}$$

  where μ_s is the mean of the simulated time series and μ_o is the mean of the observed time series.

- **Nash-Sutcliffe Efficiency (NSE):** The Nash–Sutcliffe model efficiency coefficient (NSE) is used to assess the predictive skill of hydrological models. It is defined as one minus the ratio of the error variance of the modeled time-series divided by the variance of the observed time-series. In the situation of a perfect model with an estimation error variance equal to zero, the resulting Nash–Sutcliffe Efficiency equals 1 (NSE=1). Conversely, a model that produces an estimation error variance equal to the variance of the observed time series results in a Nash–Sutcliffe efficiency of 0.0 (NSE=0). In reality, NSE=0 indicates that the model has the same predictive skill as the mean of the time-series in terms of the sum of the squared error. In the case of a modeled time series with an estimation error variance that is significantly larger than the variance of the observations, the NSE becomes negative. An efficiency less than zero (NSE<0)

occurs when the observed mean is a better predictor than the model. Values of the NSE nearer to 1 suggest a model with more predictive skill. For the application of NSE in regression procedures, the Nash–Sutcliffe efficiency is equivalent to the coefficient of determination (R2), thus ranging between 0 and 1.

The formula for NSE is

$$NSE = 1 - \frac{\sum_{t=1}^{T}(Q_{o,t} - Q_{m,t})^2}{\sum_{t=1}^{T}(Q_{o,t} - \bar{Q}_o)^2}$$

where: $Q_{o,t}$ is the observed value at time t, $Q_{m,t}$ is the modeled value at time t, Q_o is the mean

- **Normalized Root Mean Square Error** (NRMSE): NRMSE is a standardized measure of the root mean square error (RMSE), allowing for comparison between datasets with different scales. It represents the typical magnitude of the errors relative to the range or mean of the observed data. Lower NRMSE values indicate better model performance, with 0 representing a perfect fit.

- **Taylor Diagrams**: A Taylor diagram is a polar plot that visually summarizes how well a model or a set of models matches observations. It simultaneously displays three key statistical metrics: the correlation coefficient, the centred root-mean-square error (RMSE), and the standard deviation of the modeled and observed fields. The diagram plots the standard deviation of the model against that of the observations on the radial axis, and the correlation coefficient is represented by the azimuthal angle. The centred RMSE is proportional to the distance from the "reference" point (representing the observations) on the diagram. Models that perform well will be located close to the reference point, indicating high correlation, similar standard deviation, and low RMSE. Taylor diagrams provide a concise and intuitive way to compare the performance of multiple models against a reference dataset or to assess the skill of a single model across different variables or time periods.

We mainly focused on Taylor diagrams as they are good visual way to investigate different quantities at once. We utilized the Kling-Gupta Efficiency in a few testbeds. KGE is basically a Taylor diagram pressed into one value. Therefore, it is very well suited to automatically rank the quality of models or regions.

4 Outlook

With the rainforest, grassland, and temperate forest testbeds now in place, the work has been structured into three distinct phases:

1. General evaluation of more than 50 variables, grouped into key categories: radiation, surface energy balance, atmospheric boundary layer, and cloud properties.
2. Selection of integrative variables that are most relevant for diagnosing the transport of greenhouse gases, based on insights gained from the initial evaluation.

3. In-depth diagnosis of transport processes, guided by the behaviour of these selected variables and their interactions across different surface types and atmospheric conditions.

We anticipate that all these aspects will be discussed thoroughly during the upcoming workshop on 2–3 July 2025, to be held in Wageningen (see Annex for the full programme).

A crucial next step involves exploring the impact of increased vertical resolution within ICON, particularly for selected testbeds where fine-scale atmospheric processes are paramount. While technically straightforward to implement, a comprehensive tuning of the model for these higher resolutions falls outside the scope of the current project due to the significant computational and labour resources required. Nevertheless, preliminary tests at higher vertical resolution will provide valuable insights into the model's behaviour and its ability to resolve critical features, such as transport across atmospheric barriers like the tropopause. These initial experiments will serve to highlight areas where increased resolution offers substantial improvements and guide future, more extensive tuning efforts.

Furthermore, the current implementation of grid point nudging with strong relaxation to ERA5 data in ICON presents a potential limitation, as it may inadvertently suppress or distort inherently small-scale atmospheric phenomena. To address this, we plan to investigate alternative nudging strategies. Specifically, adopting a reinitialization approach every 24 hours, analogous to the methodology employed in the ECMWF Integrated Forecasting System (IFS), is envisioned. This approach is expected to allow for the development of more realistic small-scale features while still maintaining consistency with large-scale observational constraints. To further validate this revised nudging scheme, comparative tests utilizing observational data from the German Weather Service (DWD) as an alternative to IFS data will be conducted, providing an independent assessment of its performance. Additionally, to gain a more detailed understanding of transport processes, particularly within the context of overshooting events, we intend to utilize IFS data with a higher temporal output frequency. This will allow for a more precise analysis of the evolution and dynamics of transport in these highly energetic atmospheric phenomena.

To mitigate observed discrepancies in the representation of trace gases such as carbon monoxide (CO) and methane (CH4), future work will explore the utilization of initialization data from the Copernicus Atmosphere Monitoring Service (CAMS) in place of current IFS-derived simulations. CAMS data, being specifically tailored for atmospheric composition, is anticipated to provide a more accurate and comprehensive initial state for these crucial species, leading to improved model performance in simulating their sources, sinks, and transport. This transition is expected to significantly enhance the model's capabilities for atmospheric chemistry studies and contribute to a more robust understanding of air quality and climate-relevant processes.

We intend to integrate the LQM3DCONS limiter into the IFS model. This limiter is designed to improve the monotonicity and stability of tracer advection, particularly in regions with strong gradients, which is crucial for accurately representing the distribution of atmospheric constituents. Concurrently, the implementation of the IFS tracer mass fixer will address any numerical mass non-conservation issues that may arise during the advection process, ensuring that the total mass of each tracer is conserved over the simulation period. Furthermore, we plan to incorporate an improved version of the COMAD interpolation scheme. These specific advancements in tracer transport, the mass fixer, and the interpolation method are based on the detailed developments and insights documented in Deliverable 1.2 (D1.2) of the CATRINE project.

In collaboration with Work Packages 7 (WP7) and 8 (WP8), a dedicated effort is underway to investigate the persistently elevated carbon dioxide (CO2) and sulfur hexafluoride (SF6) ratios observed within ICON simulations. Comparisons conducted within WP7 have consistently

highlighted a disproportionately large growth rate for both these trace gases in ICON when contrasted with other participating models.

These investigations are currently ongoing, with the primary objective of identifying the root cause of this discrepancy within the ICON framework. Once the underlying issue is pinpointed and resolved, the relevant simulations will be repeated. Critically, these subsequent simulations will adhere to an updated WP7 protocol, which now includes the definition of specific altitudes. These designated altitudes will be leveraged for detailed flux diagnostics, providing a quantitative measure of transport across these crucial atmospheric interfaces. The resulting flux diagnostics will then be compared against other established metrics defined within the scope of this deliverable, offering a comprehensive assessment of the improvements achieved in the representation of CO2 and SF6 transport and evolution in ICON.

As part of ongoing international collaborative efforts, a dedicated workshop (together with WP8) will be convened in the beginning of July to address the critical need for robust metrics in evaluating atmospheric transport processes within global tracer transport models (Milestone M6). This event aims to bring together leading experts to delve into the complex interplay of atmospheric transport and its influence on greenhouse gas distributions across a spectrum of scales.

The workshop will center on three fundamental themes. Firstly, it will explore vertical transport processes, emphasizing the quantitative assessment of exchanges and gradients between the atmospheric boundary layer and the free troposphere. This includes a particular focus on cloud-mediated transport and the intricate interactions occurring with the lower stratosphere. Secondly, the workshop will examine greenhouse gas transport budgets, considering their calculation at local, regional, and global scales, while also addressing the significant influence of large point-emitting sources. Finally, discussions will encompass large-scale horizontal gradients, which will be evaluated over extended temporal integration periods to elucidate persistent patterns driven by transport.

A cornerstone of the workshop will be the utilization of the TestBeds that were investigated in this deliverable.

The anticipated outcomes of this workshop are multifold. It aims to deliver a well-defined protocol of metrics specifically designed for the evaluation of transport parameterizations. Furthermore, it seeks to provide a comprehensive analysis of transport representations through the application of budget diagnostics under a wide range of atmospheric conditions. Ultimately, the workshop will contribute to a quantitative assessment of the inherent uncertainties and systematic errors associated with current transport modeling approaches.

The next phase of our research, encompassed within Work Package 6 (WP6), will be dedicated to a comprehensive evaluation and assessment of the reliability and physical consistency of parametrizations for unresolved vertical transport (specifically turbulent mixing and convection) and unresolved circulations driven by surface heterogeneities in global transport models.

WP6 will systematically address systematic errors that typically arise at critical interfaces within the atmosphere: the canopy-atmosphere, atmospheric boundary layer-free troposphere, troposphere-stratosphere, and cloud-environment boundaries. These interfaces are characterized by strong gradients in thermodynamic and tracer variables, making their accurate representation paramount. We will leverage the extensive real and virtual data generated during Work Package 5 to achieve this. By combining insights from high-resolution large eddy simulations, comprehensive observational datasets from supersites and field campaigns (including both meteorological and atmospheric composition data), we aim to identify and attribute errors in trace gas transport driven by the combined effects of turbulence,

meso- and synoptic-scale weather. Both observational analyses and fine-scale simulations will provide essential guidance for evaluating physical schemes, particularly those related to turbulent mixing and convection. Based on these evaluation diagnostics and through detailed sensitivity analyses, we will propose targeted improvements wherever feasible. A key objective is to assemble specialized testbeds designed to assess physical parametrizations using detailed observations of specific processes. To systematically identify these errors, we plan to calculate specific metrics, including (a) mass-fluxes related to moist (cloud) convection and turbulent vertical transport, and (b) horizontal transport as further elaborated in relevant project documentation.

Task 6.1 will focus on the evaluation of operational turbulent mixing and shallow/deep convection parametrizations near the surface and tropopause. This task builds upon the work done in D5.2 to further develop appropriate tracer-transport scores for operational global transport models, drawing upon a wide array of observations from field campaigns and supersites. Statistical analysis of transport errors will be performed for different ecosystem conditions (tropical, semi-arid/irrigation, temperate, boreal) and across various tracers and thermodynamic variables. Once the sources of error are attributed, we will formulate and apply optimal strategies to mitigate these errors by analyzing both numerical aspects (model grid and effective resolutions) and physical aspects of the parametrizations, providing key recommendations to Task 6.2.

Task 6.2 will utilize testbeds to assess the development of parametrizations, their interactions with resolved processes, and uncertainty quantification within the IFS. The primary aim of this task is to understand the sensitivity of tracer transport to uncertainties in parametrized processes and model resolution, building on the testbeds developed in WP5 (D5.2) and the evaluation metrics from T6.1. We will explore both parameter uncertainty and model 'structural' uncertainty. Parameter uncertainty will be investigated through the Stochastically Perturbed Parametrizations (SPP) scheme, which is integral to the IFS Ensemble Prediction System (ENS) and represents model uncertainty originating from atmospheric physics parametrizations. By selectively activating individual SPP perturbations (e.g., in convection or turbulent diffusion schemes), we can pinpoint which parametrization components have the greatest impact on tracer transport and identify the atmospheric conditions, regions, or levels that are most sensitive. We will further investigate the transport properties of ensemble members—considering different SPP perturbations, model resolutions, and comparing them to initial conditions perturbations—with a focus on assessing the exchangeability of ensemble members. In contrast, structural uncertainties, arising from unconstrained or unknown functional forms of equations in parametrizations, will be explored by testing alternative functional forms (e.g., evaluating a new TKE scheme in the IFS), as their impact on tracer transport may not be evident within the phase space of parameter perturbations alone. The results from Work Packages 1 and 2 will be integrated into the error analysis to provide a holistic view of error attribution concerning different transport processes.

5 Conclusion

This deliverable describes progress in the ongoing development and evaluation of testbeds, to facilitate the evaluation of model transport. Two areas of interest are defined: one set of testbeds targets the transport between the boundary layer and the free troposphere, while another set of testbeds targets transport in the upper troposphere and lower stratosphere (UTLS).

The BL testbeds have been successfully developed as shown by initial model-observation comparisons.

For this deliverable, we conducted extensive evaluations of greenhouse gas (GHG) transport within the IFS and ICON global models using various observational datasets from missions like ATom, DCOTSS, IAGOS, STRATOCLIM, and WISE. These evaluations served as crucial testbeds to assess model accuracy in simulating GHG transport processes, particularly those influenced by turbulence and cloud dynamics, and their role in the vertical redistribution of CO_2 and other trace gases. This work will be completed with the diagnostic of the budget of CO_2 to identify which terms are governing the transport.

Key findings across the testbeds indicate that while both IFS and ICON models show considerable skill in simulating the global distribution of CO_2 and SF6, there are ongoing challenges. Differences in biases and weaker performance for tracer ratios suggest difficulties in perfectly representing the interplay of different transport processes, including vertical exchange between the boundary layer, free troposphere, and the Upper Troposphere Lower Stratosphere (UTLS) region. Specifically, models often struggle to fully capture the complex vertical structure and temporal variability observed in CO_2 mixing ratios, particularly during events like overshooting convection, leading to smoother vertical profiles and an underestimation of observed gradients and short-term fluctuations. Continued efforts are needed to refine model parameterizations and improve the accurate representation of fine-scale processes and turbulent mixing. Some of these deficiencies are inherently associated with the limited spatial and temporal resolution of the models and the complex and stochastic nature of convection. The very high resolved simulations done in the boundary layer testbeds within this workpackage and within WP3/WP4 will help understanding those limitations.

6 References

Boussetta, S., et al. (2013), Natural land carbon dioxide exchanges in the ECMWF integrated forecasting system: Implementation and offline validation, *J. Geophys. Res. Atmos.*, 118, 5923–5946, doi:10.1002/jgrd.50488.

Deng, F., Jones, D. B. A., Walker, T. W., Keller, M., Bowman, K. W., Henze, D. K., Nassar, R., Kort, E. A., Wofsy, S. C., Walker, K. A., Bourassa, A. E., and Degenstein, D. A.: Sensitivity analysis of the potential impact of discrepancies in stratosphere–troposphere exchange on inferred sources and sinks of CO2, Atmos. Chem. Phys., 15, 11773–11788, https://doi.org/10.5194/acp-15-11773-2015, 2015.

Gaubert, B., Stephens, B. B., Basu, S., Chevallier, F., Deng, F., Kort, E. A., Patra, P. K., Peters, W., Rödenbeck, C., Saeki, T., Schimel, D., Van der Laan-Luijkx, I., Wofsy, S., and Yin, Y.: Global atmospheric CO2 inverse models converging on neutral tropical land exchange, but disagreeing on fossil fuel and atmospheric growth rate, Biogeosciences, 16, 117–134, https://doi.org/10.5194/bg-16-117-2019, 2019.

Gerbig, C., Körner, S., and Lin, J. C.: Vertical mixing in atmospheric tracer transport models: error characterization and propagation, Atmos. Chem. Phys., 8, 591–602, https://doi.org/10.5194/acp-8-591-2008, 2008.

van Heerwaarden, C. C., van Stratum, B. J. H., Heus, T., Gibbs, J. A., Fedorovich, E., and Mellado, J. P.: MicroHH 1.0: a computational fluid dynamics code for direct numerical simulation and large-eddy simulation of atmospheric boundary layer flows, Geosci. Model Dev., 10, 3145–3165, https://doi.org/10.5194/gmd-10-3145-2017, 2017.

Heus, T., van Heerwaarden, C. C., Jonker, H. J. J., Pier Siebesma, A., Axelsen, S., van den Dries, K., Geoffroy, O., Moene, A. F., Pino, D., de Roode, S. R., and Vilà-Guerau de Arellano, J.: Formulation of the Dutch Atmospheric Large-Eddy Simulation (DALES) and overview of its applications, Geosci. Model Dev., 3, 415–444, https://doi.org/10.5194/gmd-3-415-2010, 2010

Peter A. Rochford (2016) SkillMetrics: A Python package for calculating the skill of model predictions against observations, http://github.com/PeterRochford/SkillMetrics

Schröter, J. et al. (2018). ICON-ART 2.1: a flexible tracer framework and its application for composition studies in numerical weather forecasting and climate simulations. Geosci. Model Dev., 11, 4043-4068, https://doi.org/10.5194/gmd-11-4043-2018

Schuh, A. E., Jacobson, A. R., Basu, S., Weir, B., Baker, D., Bowman, K., et al. (2019). Quantifying the impact of atmospheric transport uncertainty on CO2 surface flux estimates. *Global Biogeochemical Cycles*, *33*, 484–500. https://doi.org/10.1029/2018GB006086

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., and Denning, A. S.: Weak Northern and Strong Tropical Land Carbon Uptake from Vertical Profiles of Atmospheric CO2, Science, 316, 1732–1735, https://doi.org/10.1126/science.1137004, 2007

Vilà-Guerau de Arellano, J., and Coauthors, 2024: CloudRoots-Amazon22: Integrating Clouds with Photosynthesis by Crossing Scales. *Bull. Amer. Meteor. Soc.*, **105**, E1275–E1302, https://doi.org/10.1175/BAMS-D-23-0333.1.

Wofsy, S.C., and ATom Science Team. 2018. ATom: Aircraft Flight Track and Navigational Data. ORNL DAAC, Oak Ridge, Tennessee, USA. https://doi.org/10.3334/ORNLDAAC/1613

Zängl, G. et al. (2015). The ICON (ICOsahedral Non-hydrostatic) modelling framework of DWD and MPI-M: Description of the non-hydrostatic dynamical core. Q. J. R. Meteorol. Soc. 141, 563–579. https://doi.org/10.1002/qj.2378

7 Annex Workshop scope and agenda

Workshop

Metrics to evaluate the transport processes

in global tracer transport models

Wageningen University

2-3 July 2025

SCOPE

As part of the EU project *Carbon Atmospheric Tracer Research to Improve Numerical Schemes and Evaluation* (CATRINE), we are organizing a workshop at Wageningen University on **2–3 July 2025** to explore the role of atmospheric transport processes in shaping greenhouse gas distributions across multiple scales.

The workshop will focus on three core themes:

1. **Vertical transport processes**, with an emphasis on quantifying exchanges and gradients between the atmospheric boundary layer and the free troposphere, including cloud-mediated transport and interactions with the lower stratosphere.
2. **Greenhouse gas transport budgets**, examining their calculation at local, regional, and global scales, and addressing the influence of large point-emitting sources.
3. **Large-scale horizontal gradients**, evaluated over extended temporal integration periods to understand persistent transport-driven patterns.

A central element of the workshop will be the use of **TestBeds**, integrating a range of modelling strategies—from explicitly resolving turbulence and clouds to parameterizing these processes—combined with diverse observational datasets. These efforts will be supported by **ensemble-based analyses** to evaluate the sensitivity of transport representations.

The primary modelling tools will include **global-scale systems** (e.g., IFS, ICON-ART), underpinned by observations across multiple scales and high-resolution **large-eddy simulations**.

The workshop aims to deliver:

- A well-defined protocol of metrics to evaluate transport parameterizations;
- An analysis of transport representations using budget diagnostics under a wide range of condistions.
- A quantitative assessment of related uncertainties and systematic errors.

AGENDA

Wednesday 2nd July

10-12 Presentations EU CATRINE project (20-25 minutes)

Transport greenhouse in the ABL and FT: TestBeds

Transport greenhouse in the Upper Troposphere and Low Stratosphere

Local atmospheric transport models for monitoring emission hotspots

Metrics to evaluate global tracer transport

Alessandro Savazzi (ECMWF):

Diagnostics of Parameterised CO_2 Transport in the Boundary Layer

Vincent de Feiter (MAQ, WUR):

Parameterised and Resolved CO2 Exchange in the Lower Tropical Troposphere Across Clear-to-Cloudy Conditions”

Anja Raznjevic (MAQ, WUR):

Evaluating High-Resolution Simulations of Atmospheric Composition in Rotterdam Using Satellite and Ground-Based Observations

Anna Agusti-Panareda (ECMWF):

Towards a framework to explore diagnostics of the atmospheric CO_2 budget

Achraf Qor-el-Aine (KIT)

Modeling CO_2 in UTLS during extreme transport: Evaluating ICON-ART and IFS CO_2 Simulations Against Aircraft Observations'

12-13 Lunch

13:00-13:45 **Sarah-Jane Lock** (ECMWF):

Representations of model uncertainty and using the IFS ensemble to explore transport impacts

13:45-14:30 **Jochen Fönstner** (DWD Germany)

Investigation of transport uncertainties with ICON-ART and MH diagnostics in the German ITMS project

14:30-15:00 Tea break

15:00-15:45 **Harald Bönisch** (Karlsruhe Institute of Technology, Germany)

The in-situ problem - evaluate models with in-situ observations

16:00-17:00 Discussion

- Transport metrics
- Diagnostics transport
- Use of stochastic numerical experiments

19 Dinner

Thursday 3rd July

9-11 Presentations

11-12 **Yasmine Bennouna** (Laboratoire Aerologie, Université Paul Sabatier, France)
In-situ monitoring of carbon tracers by IAGOS for CAMS model evaluation

12-13 Lunch

13-15 Developing strategies to integrate local and global transport to assess the uncertainties in the transport processes: future plans

15 end workshop

8 Project partners:

Partners	
EUROPEAN CENTRE FOR MEDIUM-RANGE WEATHER FORECASTS	ECMWF
WAGENINGEN UNIVERSITY	WU
KARLSRUHER INSTITUT FUER TECHNOLOGIE	KIT

Document History

Version	Author(s)	Date	Changes
0.1 (Initial document created)	Stefan Versick, Jordi Vila, Anna Agusti-Panareda, Achraf Qor-El-Aine	06/06/2025	
1.0	Stefan Versick, Jordi Vila, Anna Agusti-Panareda, Achraf Qor-El-Aine	26/6/2025	Updated after review comments and issued

Internal Review History

Internal Reviewers	Date	Comments
Maarten Krol (WU) and Leena Jarvi (UH)	June 2025	Textual suggestions and requests for a clearer structure. Links to upcoming work and better alignment between BL and UTLS work needed.