CARBON ATMOSPHERIC TRACER RESEARCHTO IMPROVE NUMERICAL SCHEMES AND EVALUATION

CATRINE

Carbon Atmospheric Tracer Research to Improve Numerics and Evaluation

D6.1 Calculations of tracer transport metrics and variables based on selected test beds: score cards

Due date of deliverable	30th April 2026
Submission date	29th April 2026
File Name	CATRINE-D6-1.V5.0
Work Package /Task	Task 6.1
Organisation Responsible of Deliverable	Wageningen University Research (WUR)
Author name(s)	Jordi Vilà (WUR), Stefan Versick (KIT), Anna Agusti-Panareda (ECMWF), Vincent de Feiter (WUR), Alessandro Savazzi (ECMWF), Achraf Qor-el-aine (KIT), Sarah-Jane Lock (ECMWF), Annelize van Niekerk (ECMWF)
Revision number	5
Status	Issued
Dissemination Level / location	Public

Funded by the European Union

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

1 Executive Summary

This deliverable (D6.1) builds directly on the TestBed framework developed in Deliverable 5.1, which established an integrated observational and modelling platform to evaluate atmospheric transport in global models. D5.1 combined comprehensive observations, high-resolution large-eddy simulations (LES), and global modelling to identify key variables and diagnostic metrics controlling tracer transport. These TestBeds, initially demonstrated for the Amazonian CloudRoots campaign (rainforest ecosystem) and designed for extension to other ecosystems (grassland and temperate forest) such as Cabauw–Loobos, provide the observational and modelling basis for systematic model evaluation.

Building on this foundation, D6.1 develops a scorecard framework to evaluate the representation of atmospheric transport processes in global numerical models used for greenhouse gas monitoring. Accurate simulation of atmospheric transport is essential for interpreting atmospheric concentrations of trace gases such as CO_2 and CH_4 and for improving emission estimates derived from atmospheric inversion systems. However, many relevant processes—including turbulent mixing, cloud-driven transport, and circulations induced by surface heterogeneity and the exchange between the upper troposphere and lower stratosphere—occur at spatial scales that cannot be explicitly resolved in global models and must therefore be parameterized.

The scorecard methodology uses the TestBed datasets and diagnostics defined in D5.1 to systematically assess transport behaviour in numerical models. The framework combines several comparison pathways, including model–observation comparisons (OBS–IFS), comparisons between large-eddy simulations and global models (LES–IFS), and sensitivity experiments within the global model (IFSc–IFS). Model performance is evaluated using key diagnostic metrics such as atmospheric boundary-layer height, diurnal variability, vertical gradients, and flux ratios, which together quantify major transport pathways including surface exchange, ABL–free troposphere exchange, cloud transport, and advection.

Two prototypes scorecards are presented to explain the methodology, demonstrating how the framework can diagnose model transport behaviour and identify dominant processes contributing to tracer variability. Overall, D6.1 extends the TestBed concept developed in D5.1 by providing a structured, process-based evaluation framework for atmospheric transport in global greenhouse gas monitoring systems. In the conclusion section, preliminary recommendations are given on the use of scorecards.

Table of Contents

2 Introduction

2.1 Background

Atmospheric transport plays a fundamental role in determining the distribution of trace gases such as CO_2 and CH_4 and therefore directly affects the ability of atmospheric inversion systems to estimate surface fluxes. Global transport models used in monitoring systems typically operate at spatial resolutions ranging from tens to hundreds of kilometres, which are too coarse to explicitly resolve key processes governing vertical and horizontal transport. Representative examples of these transport contributions are the exchange between the atmospheric boundary layer and the free troposphere driven by turbulence and cloud transport, and horizontal transport driven by thermal contrasts such a sea-land differences. As a result, processes such as turbulent mixing, shallow and deep convection, and circulations induced by surface heterogeneity must be represented through parametrizations. These unresolved processes strongly influence the exchange of tracers across major interfaces in the atmosphere, including the surface–atmosphere interface, the top of the atmospheric boundary layer (ABL), and the upper troposphere–lower stratosphere (UTLS).

Errors in the representation of these processes can lead to systematic biases in tracer distributions and ultimately affect the accuracy of emission estimates derived from atmospheric observations. Improving the evaluation and understanding of these parametrizations is therefore a key requirement for advancing the performance of global atmospheric transport models and the Copernicus CO_2 Monitoring and Verification Support (CO_2MVS) capacity.

2.2 Scope of this deliverable

2.2.1 Objectives of this deliverable

The main objective of WP6 is to evaluate and assess the reliability and physical consistency of parametrizations of unresolved vertical and horizontal transport. To this end and considering the differences in the governing physics and spatiotemporal scales, the work is divided into two components: (i) the surface–atmospheric boundary layer–free troposphere interface (ABL–FT), led by WU, and (ii) the upper troposphere–lower stratosphere region (UTLS), led by KIT. For both the ABL–FT and UTLS components, the assessment of transport-related uncertainties is carried out using scorecards.

These scorecards are designed to monitor the performance of large-scale models, such as IFS and ICON-ART, against available observations and against high-resolution models that more explicitly resolve transport processes, such as large-eddy simulations (LES). For the ABL–FT component, the fully operational test beds established in Amazonia and the Netherlands (Deliverables 5.1 and 5.2) have proven essential in facilitating a systematic and harmonized analysis of the observational and modelling datasets.

2.2.2 Work performed in this deliverable

a. Scorecards Atmospheric Boundary Layer-Free Troposphere transport errors

We develop and design scorecards which map different aspects of the model performance. Figure 1 shows the roadmap of the possible scorecard configurations used in the analysis. The framework combines different **reference comparisons**, **metrics**, **transport processes**, **meteorological regimes**, **sensitivity experiments**, **spatiotemporal scales**, and **ecosystem types**. Model performance is evaluated through three main comparison pathways: observations (OBS) against the Integrated Forecast System (IFS) model (OBS–

IFS), high-resolution large-eddy simulations against IFS (LES–IFS) (LES, large-eddy simulations), and sensitivity experiments within IFS itself (IFSc–IFS). In all these intercomparison we assume that observations and large-eddy simulations are closer to the real values due to the in-situ measurements and the explicit representation of turbulence.

These comparisons allow us to assess both realism with respect to measurements and internal model consistency. The scorecards evaluate model behaviour using a set of **diagnostic metrics**, including the atmospheric boundary layer (ABL) height, diurnal amplitude range, vertical gradients, and flux ratios. These diagnostics are used to assess key **transport pathways** such as surface fluxes, ABL–free troposphere exchange, cloud mass flux, cloud–ABL coupling, advection, and residual transport. To capture the diversity of atmospheric regimes, the analysis is organized around **prototypical boundary-layer states**, including convective and stable conditions as well as transition periods and different cloud regimes (shallow cumulus, stratocumulus, and deep convection).

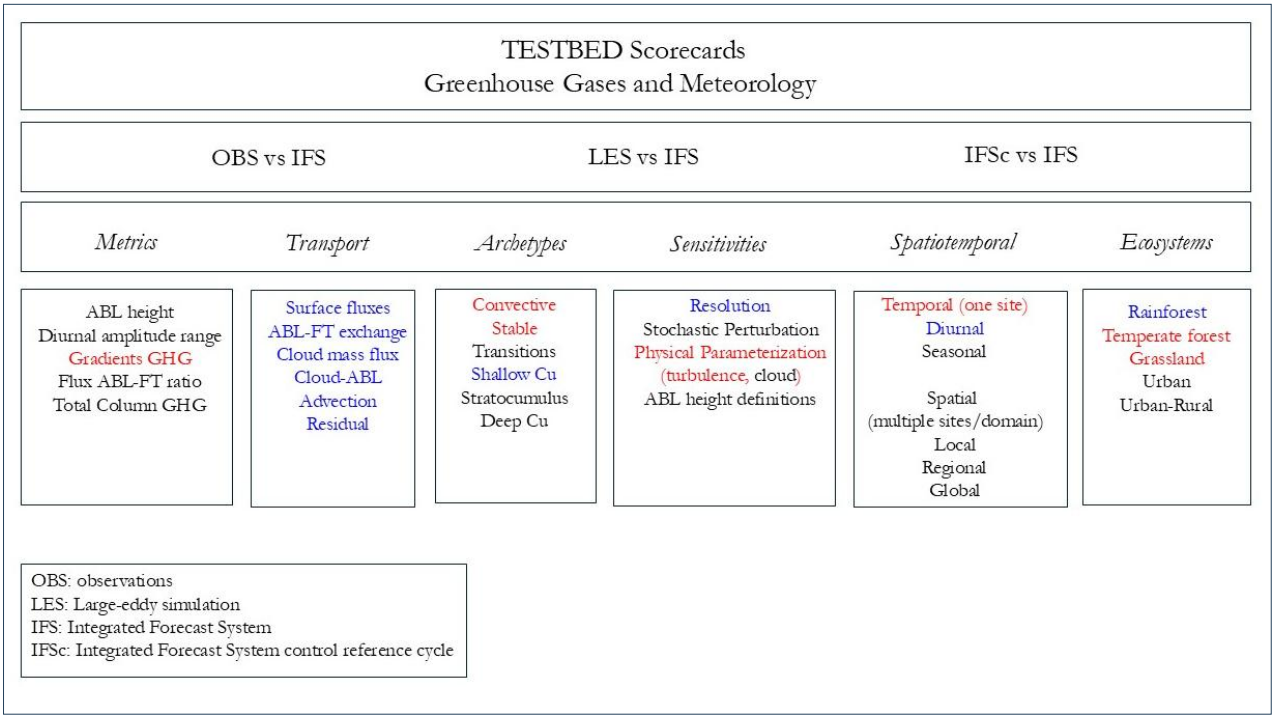

Figure 1. Roadmap framework of the Testbed scorecards used to assess model transport performance. The scorecards combine different comparison pathways (OBS–IFS, LES–IFS, IFSc–IFS) with diagnostic metrics, key transport processes across different atmospheric regimes, model sensitivities (e.g. resolution and parameterizations), spatiotemporal scales (diurnal, seasonal, local to global), and ecosystem types (rainforest, temperate forest, grassland, urban, and urban–rural). In red and blue, respectively, the prototype options used to intercompare carbon dioxide gradients of CO2 (red)and transport process (blue) in the Amazonia TestBed.

The scorecards are powerful tools to explore **model sensitivities** to key modelling choices, such as horizontal resolution, physical parameterizations (including stochastic perturbations), and alternative definitions of ABL height. The framework also considers **spatiotemporal variability**, evaluating performance at different temporal scales (diurnal and seasonal cycles) and spatial scales ranging from single-site analysis to local, regional, and global domains. Finally, the methodology is applied across different **ecosystem types**, including rainforest, temperate forest, grassland, and urban environments, as well as urban–rural contrasts. This roadmap design attempts to show all the possibilities of model evaluation and it is designed to harmonize the intercomparison of global models like IFS or ICON-ART

against all sort of observations and high-resolution simulations (LES) in a systematic manner and allowing the possibility to be reproduced.

This structured scorecard approach enables a systematic and harmonized model comparison through the assessment of transport processes in the models, and provides a comprehensive overview of strengths, weaknesses, and sensitivities across scales, regimes, and ecosystems. As representative examples on how we calculate these scorecards we will present two examples: (1) intercomparison IFS versus LES and OBS on the CO2 gradients at the surface and (b) assessing the transport contributions to specific processes

b. Scorecards Upper Troposphere-Lower Stratosphere (UTLS) transport errors

We develop and design scorecards which map different aspects of the model performance in the Upper Troposphere / Lower Stratosphere (UTLS). The framework combines different reference comparisons, metrics, transport processes, archetypes, sensitivities, spatiotemporal scales, and observational campaigns (see Figure 2). Model performance is evaluated through three main comparison pathways: observations (OBS) against the Integrated Forecast System (IFS) model (OBS-IFS), the IFS against the ICOsahedral Non-hydrostatic model with Aerosols and Reactive Trace gases (IFS vs ICON-ART), and comparisons of ICON-ART against the IFS at different resolutions. These comparisons allow us to assess both realism with respect to measurements and internal model consistency.

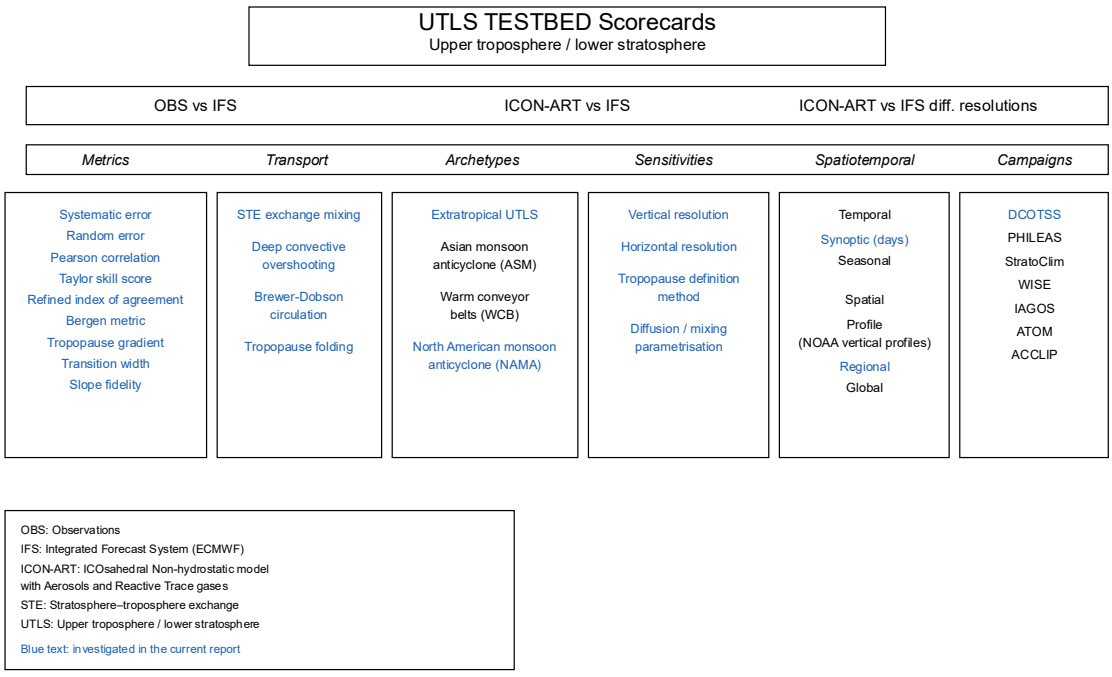

Figure 2. Roadmap framework of the UTLS Testbed scorecards used to assess model transport performance in the upper troposphere and lower stratosphere. The scorecards combine different comparison pathways (OBS vs IFS, IFS vs ICON-ART, and ICON-ART vs IFS at different resolutions) with diagnostic metrics, key transport processes across different atmospheric archetypes, model sensitivities such as resolution and tropopause definitions, spatiotemporal scales ranging from synoptic days to seasonal profiles, and specific observational flight campaigns. This framework enables a systematic and harmonized evaluation of stratosphere-troposphere exchange and other critical dynamics.

The scorecards evaluate model behaviour using a set of diagnostic metrics, including systematic error, random error, Pearson correlation, Taylor skill score, refined index of agreement, Bergen metric, tropopause gradient, transition width, and slope fidelity. These diagnostics are used to assess key transport pathways such as stratosphere-troposphere exchange (STE) mixing, deep convective overshooting, Brewer-Dobson circulation, and tropopause folding. To capture the diversity of atmospheric regimes, the analysis is organized around prototypical states, including the extratropical UTLS, the Asian monsoon anticyclone (ASM), warm conveyor belts (WCB) (publication Qor-El-Aine et al. (2026) in preparation), and the North American Monsoon Anticyclone (NAMA).

The scorecards are powerful tools to explore model sensitivities to key modelling choices, such as vertical resolution, horizontal resolution, the tropopause definition method, and diffusion or mixing parametrisations. The framework also considers spatiotemporal variability, evaluating performance at different temporal scales (synoptic days and seasonal cycles) and spatial scales ranging from profile analysis (such as NOAA vertical profiles) to regional and global domains. Finally, the methodology is applied across different flight campaigns, including DCOTSS (Bowman et al., 2026), PHILEAS (Riese et al., 2025), StratoClim (StratoClim Consortium, 2017), WISE (Riese and Hoor, n.d.), IAGOS (Petzold et al., 2015), ATom (Wofsy et al., 2018), and ACCLIP (Pan et al., 2025). This structured scorecard approach enables a systematic and harmonized model comparison through the assessment of transport processes in the models, and provides a comprehensive overview of strengths, weaknesses, and sensitivities across scales and regimes.

2.2.3 Deviations and counter measures

No deviations to report.

2.3 Project partners:

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

3 Results: selected and representative scorecards

3.1 Prototype scorecard ABL-FT transport errors

This structured scorecard approach enables a systematic and harmonized assessment of transport processes in the models and provides a comprehensive overview of strengths, weaknesses, and sensitivities across scales, regimes, and ecosystems. We will present two examples: (1) intercomparison IFS versus LES and OBS on the CO_2 gradients above the surface using the Testbed Cabauw-Loobos and (b) assessing the transport contributions to specific processes using the Testbed Amazonia. Here, as an example of the applicability we show two representative showcases.

a. Gradients of carbon dioxide above the surface: grassland ecosystem

A key and challenging aspect in the exchange of the processes occurs in the regions defined by strong gradients: (1) surface-ABK, ABL-FT, UT-LS. Therefore, one useful approach to evaluate vertical mixing in models is to analyse the mean vertical gradients of meteorological and greenhouse gases variables. This metric is particularly valuable for diagnosing stable thermodynamic stratified conditions, which are particularly critical for turbulent mixing. The vertical CO_2 gradient provides a way to largely separate errors in vertical turbulent transport from biases in the surface fluxes. Many observational towers measure CO_2 at two heights, allowing this evaluation to be applied broadly across sites. When multiple levels are available, such as at the Cabauw TestBed, where CO_2 is measured at four heights, the evaluation becomes more robust, allowing for a more detailed comparison in a region that is very sensitive to the sources and sinks of CO_2.

The mean gradient alone is not sufficient for a comprehensive evaluation and carries some limitations, which is why the scorecard combines it with several complementary metrics. Care is required under well-mixed conditions, where biases in the vertical gradients can become very small. In such regimes, even small absolute model–observation differences appear as disproportionately large percentage changes. This can misleadingly suggest substantial performance differences even when both simulations capture the observed gradient reasonably well.

These kinds of artefacts are also known from scorecards used in operational NWP model verification. While they are difficult to eliminate entirely, they should be minimised as far as possible. Therefore, we recommend complementing gradient-based metrics with additional diagnostics that remain robust across both stable and well-mixed regimes, ensuring a more balanced evaluation of vertical mixing performance.

A possible way to mitigate these issues is to stratify the evaluation by stability regime, for example by separating daytime (typically characterized by well-mixed conditions of the meteorological and greenhouse variables) and nighttime (typically stably stratified) conditions. Examining the daytime gradients in Figure 2 shows the challenge: even if the experiment j1kz (red line) reduces the CO_2 bias consistently at all measured levels, the gradient metric may still appear worse than the control iyw7 (green line), simply because gradients are very small during well-mixed conditions. This finding shows the difficulty of combining multiple metrics into a single, simple score. Any aggregated evaluation tool must therefore balance the sensitivity of gradient-based diagnostics with the absolute CO_2 biases, ensuring that improvements in one area are not obscured by artefacts in another.

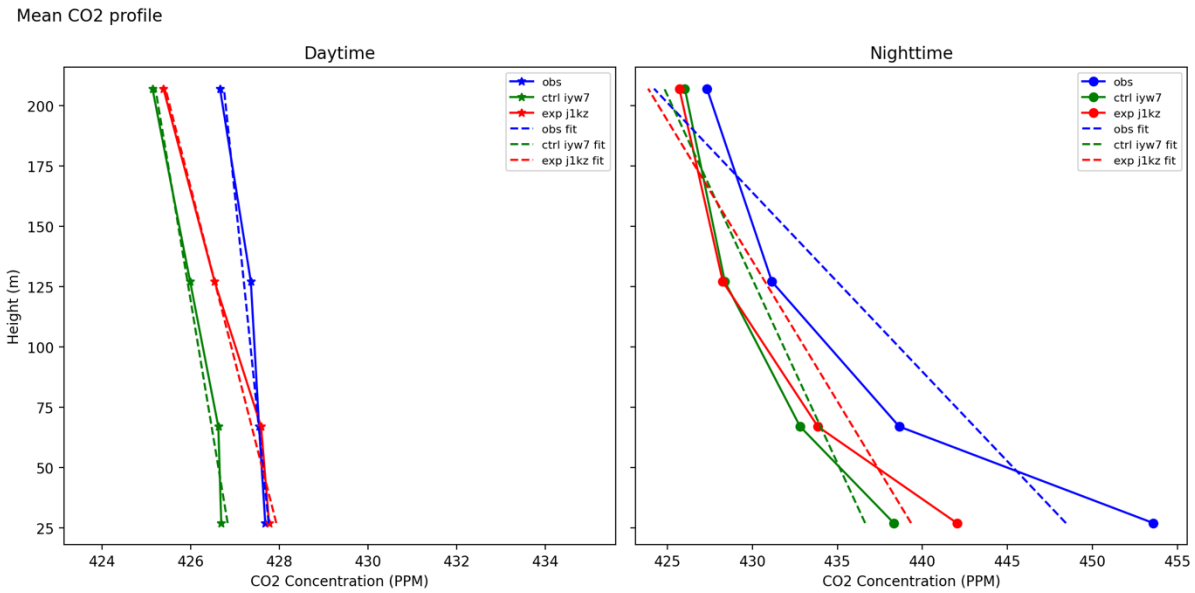

Figure 2. Comparison of the carbon dioxide profiles during daytime (unstable) and night (stable) conditions from the Testbed Cabauw-Loobos. The sensitivity analysis is done around the eddy diffusivity parameter K. Experiment iyw7 uses the standard value and experiment jikz has decreased the value by 2.

Building on these considerations, we have designed a scorecard that summarizes how the model performs. The example shown in **Figure 3** can be considered as part of an integrated and flexible framework which goes beyond the evaluation of variables relevant for the transport of carbon dioxide. Although in this case we focus on the comparison with measurement data, the tool is adaptable to a range of reference datasets (see Figure 1 for alternative configurations).

Figure 3 presents the scorecard results for an IFS experiment (j1kz) in which the eddy diffusivity coefficient K for tracers has been reduced by half, compared to a control IFS simulation (iyw7) based on the operational physics suite. These two simulations are used purely for illustrative purposes; as a more detailed discussion on the IFS sensitivity to different parametrizations and parameter uncertainty will follow in Deliverable 6.2

At this stage, we simply note that, as expected, modifying the mixing efficiency of tracers does not influence the meteorological fields, which is reflected by the 0 % change in both the bias and RMSE of metrics involving the boundary-layer height (BLH). In contrast, metrics that depend directly on tracer transport (rows 4 to 10), such as the diurnal CO_2 range proposed by Faassen et al. (2025) or the vertical CO_2 gradient, do respond to the reduction in eddy diffusivity.

Metric	Ctrl bias	Δ bias %	Δ RMSE %
blh_day [m]	-228.07	+0.00%	+0.00%
blh_night [m]	-10.47	+0.00%	+0.00%
blh_amplitude [m]	-153.88	+0.00%	+0.00%
co2_DR_67m [PPM]	-11.88	-14.77%	-7.85%
co2_27m [PPM]	-0.99	-91.47%	+8.02%
co2_207m [PPM]	-1.52	-16.21%	-0.12%
co2_gradient_day [PPM/m]	-0.00	+140.45%	+25.14%
co2_gradient_night [PPM/m]	+0.07	-29.45%	-13.83%
co2_flux_5m [PPM m/s]	+0.13	-1.05%	-0.26%
co2_flux_180m [PPM m/s]	+0.07	-7.63%	-2.46%

Figure 3. Scorecard of key variables of the land-atmosphere interaction: atmospheric boundary layer (blh), CO_2 concentration and CO_2 gradient. The sensitivity analysis is done around the eddy diffusivity parameter K. Experiment iyw7 uses the standard value and experiment jikz has decreased the value by 2.

b. Transport process contribution: Rainforest ecosystem

The main aim of this scorecard (column 2 in Figure 1) is to reconstruct the boundary-layer carbon dioxide concentration from its main physical components (see equation below). The formulation follows a mixed-layer budget approach derived by Pino et al. (2012) and updated by de Feiter et al. (2026).

In this framework, the CO_2 mole fraction is expressed as the sum of the key process that drive the diurnal variability of CO_2. The expression reads

$$\langle C\rangle = \underbrace{C_0^{FT} + (\langle C\rangle_0 - C_0^{FT})\frac{h_0}{h} + \frac{\gamma_c}{2h}(h-h_0)^2}_{\text{instantaneous contributions}} + \underbrace{\frac{1}{h}\int_{t_0}^{t}\left(\overline{w'c'}_s + M\Delta C - \overline{w'c'}_M + \int_{z_0}^{h}\left.\frac{\partial \overline{C}}{\partial t}\right|_{dyn} dz\right)dt}_{\text{time-integrated contributions}}$$

(Equation 1)

The scorecard compares the simulated CO_2 mole fraction from the IFS, ⟨C⟩ (left hand side of Equation 1), with the contributions (right-hand side) from the following terms:

I) Free tropospheric mole fraction.
II) Entrainment/detrainment during the night-to-day transition.
III) Surface fluxes, including plant assimilation and soil respiration.
IV) Entrainment/detrainment determined by the CO_2 lapse rate.
V) Boundary-layer dilution associated with clouds.
VI) Cloud mass flux.
VII) Horizontal advection.
R) Residual term representing the remaining imbalance due to uncertainties in the previous terms and numerical errors.

This decomposition allows a process-based evaluation of the model transport and surface–atmosphere-free troposphere, namely the atmosphere boundary layer-free troposphere (ABL-FT) exchange, represented and solved in the IFS under different resolutions (27, 9 and 4 km^2). Under a perfect situation the left-hand-side of the equation, ⟨C⟩ will be very close to the addition of the individual components of the right-and-side.

Figure 4 shows a representative example based on the Amazonian testbed. In this case, the CO_2 mole fraction tendency simulated by the IFS at 25 km resolution (left-hand side of the budget equation) is compared with the sum of the individual process-based contributions shown in Figure 1. Panel (a) presents the relative contribution of the different terms controlling the boundary-layer CO_2 budget throughout the diurnal cycle under shallow cumulus conditions. Panel (b) shows the initial mismatch between the reconstructed and simulated CO_2 mole fraction.

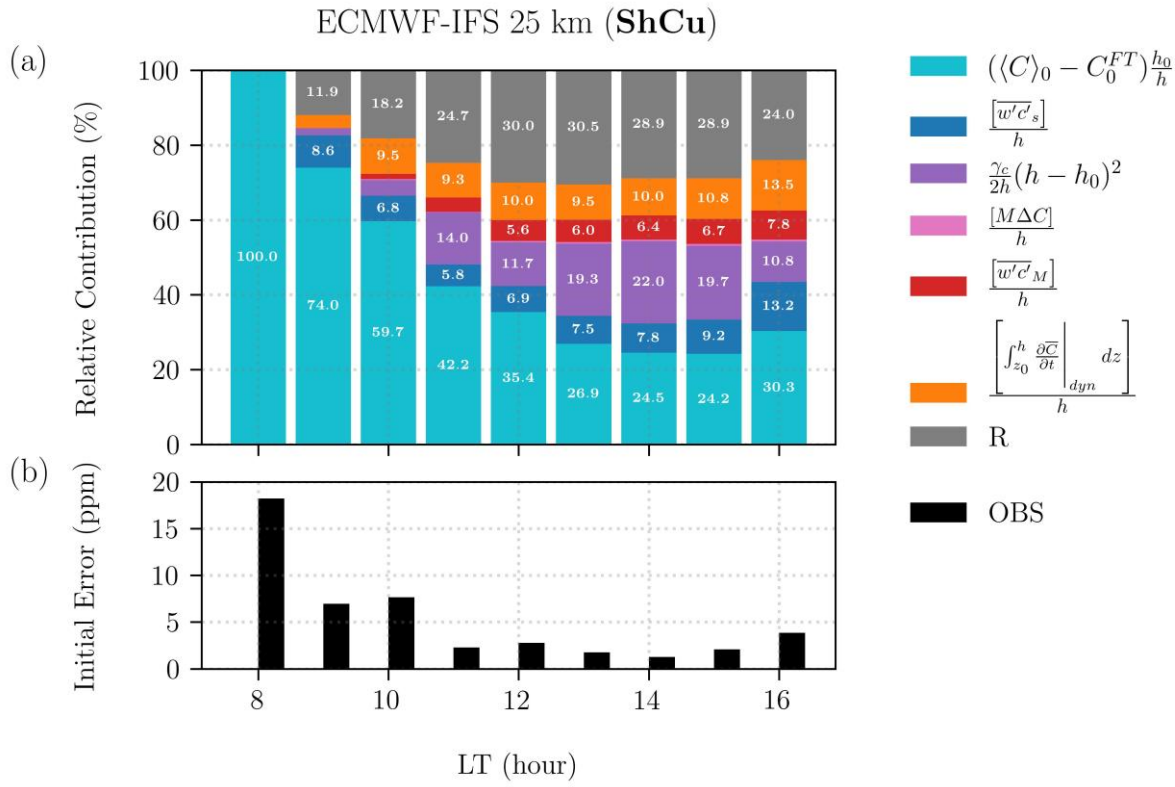

Figure 4. Example of the CO_2 budget scorecard for the Amazonian testbed under shallow cumulus (ShCu) conditions using the ECMWF-IFS at 25 km resolution. **(a)** Relative contribution of the different process terms to the reconstructed boundary-layer CO_2 mole fraction throughout the diurnal cycle. The coloured bars represent the individual contributions (free-tropospheric concentration, surface fluxes, entrainment processes, cloud-related transport, advection), while the grey segment indicates the residual. **(b)** Initial error between the reconstructed CO_2 mole fraction and the value simulated by IFS. This diagnostic illustrates how the scorecard decomposes the simulated CO_2 concentration into its underlying physical processes.

The figure shows how the dominant contributions evolve during the day, highlighting the roles of free-tropospheric concentrations, surface fluxes, entrainment processes, cloud-related transport, and advection. Similar diagnostics can be generated within the TESTBED framework using large-eddy simulations performed for the domain centred at the ATTO tower, as well as observational datasets. This allows a consistent comparison of process contributions across models and observations. For instance, and based on Figure 1, a similar figure could be done but now using the 4- or 9-kilometer resolution, or a new run with a stochastic perturbation. In that way, we can determine how the new change (resolution, physics)impacts each individual specific term and it yields a reduction of the residual term.

To complement the graphical representation of the scorecard, **Figure 5** provides a quantitative breakdown of the individual contributions to the boundary-layer CO_2 budget for the Amazonian testbed under shallow convective conditions. The table reports the relative contribution of each process term (II–VII and residual R) and the initial reconstruction error at

three key times of the diurnal cycle (12LT, in the Annex 2 there are the times 09 and 15 LT), comparing large-eddy simulations (DALES) with ECMWF-IFS at 25 km and 9 km resolutions. The need to have scorecards that depend on time show the relevance of the diurnal variability and the dynamics of each terms shown in Equation (1) if one wants to understand how the processes are acting on the XCO2 mole fraction.

Overall, the table highlights systematic differences between the high-resolution reference (DALES) and the IFS simulations. In the morning (09 LT), the CO_2 budget is dominated by entrainment during the night-to-day transition (Term II), which is underestimated by IFS, particularly at coarser resolution. As the day progresses (12–15 LT), surface fluxes (Term III) become increasingly important in DALES, while IFS strongly underestimates their contribution. Conversely, entrainment driven by the CO_2 lapse rate (Term IV) and cloud-related processes (Terms V–VI) tend to be overestimated in IFS, especially at higher resolution.

Horizontal advection (Term VII), not accounted in the DALES experiment, appears as a non-negligible contribution in IFS, indicating differences in large-scale forcing and transport representation. The residual term (R) remains substantially larger in IFS across all times, pointing to inconsistencies in the closure of the CO_2 budget and potential structural errors in the representation of transport processes. Although increasing resolution from 25 km to 9 km improves some terms (e.g. cloud mass flux), it does not systematically reduce the residual, suggesting that parametrization issues remain.

This quantitative decomposition demonstrates the added value of the scorecard framework in identifying which physical processes contribute most to model biases and how these errors evolve throughout the diurnal cycle. As Figure 1 shows, the scorecards presented in Figures 3 and 5 can be applied to all sort of numerical experiments settings: key metrics ABL-FT exchange, individual processes, sensitivity to processes, sensitivity to spatiotemporal scales and ecosystems

Scorecard — CloudRoots-Amazon22 — Sensitivity Framework: Shallow Convective Regime - 12 LT

Term	Physical Meaning	DALES	ECMWF 25 km	ECMWF 9 km	Δ (ECMWF 25 km − DALES)	Δ (ECMWF 9 km − DALES)
II	En/detrainment during the night-to-day transition	59.34	35.4	31.65	-23.94	-27.69
III	Surface fluxes (plant assimilation and soil respiration)	31.31	6.91	3.82	-24.4	-27.49
IV	En/detrainment determined by the CO_2 lapse rate	5.03	11.7	12.8	6.68	7.77
V	Boundary layer dilution associated with clouds	1.23	0.39	0.84	-0.83	-0.39
VI	Cloud mass flux	2.91	5.58	8.29	2.67	5.38
VII	Horizontal advection	-	9.98	13.65	9.98	13.65
R	Residual	0.19	30.04	28.96	29.85	28.77
Init. Error	Init. Error	1.82	2.78	4.42	0.96	2.59

Figure 5. Quantitative CO_2 budget scorecard for the Amazonian testbed under shallow convective conditions at 12 LT (in Appendix B you will find the time 09 and 15 LT). The table shows the relative contribution (%) of each process term to the reconstructed boundary-layer CO_2 mole fraction, comparing large-eddy simulations (DALES) with ECMWF-IFS at 25 km and 9 km resolutions. Differences (Δ) with respect to DALES are provided to highlight model biases. Green to red colours indicate the departure from ECMWF model results from being withing the range of DALES results to getting larger differences. The residual term (R) represents the remaining imbalance of the budget (differences between left hand side and right-hand side in equation (1), and “Init. Error” indicates the mismatch (ppm) between reconstructed and observed CO_2 mole fraction.

3.2 Prototype scorecard UTLS transport errors (KIT)

To comprehensively evaluate the performance of global models in the Upper Troposphere and Lower Stratosphere, we employ a diverse set of statistical and physically based metrics. Traditional statistical metrics such as Systematic Error and Random Error evaluate persistent overestimation or underestimation and unsystematic scatter not explained by the mean bias, respectively (Murphy & Epstein, 1989). To combine correlation and variance ratios, the Taylor Skill Score is utilized (Taylor, 2001). The Bergen Fractional Bias provides a variability normalized assessment of model performance (Hanna et al., 1991; Chang et al., 2004). The refined Index of Agreement (Willmott et al., 2012) is also incorporated as it is more sensitive to large errors than the original formulation. Beyond standard statistics, physically based diagnostics are critical for this highly stratified region. The tropopause gradient metric (Hoor et al., 2004) tests for numerical diffusion directly at the tropopause boundary. Transition width (Pan et al., 2004) is monitored because a width larger than observed indicates that the model transport schemes are excessively diffuse. The Age of Air bias metric evaluates transport timescales and large-scale circulation (Engel et al., 2009). Finally, slope fidelity tests whether the model reproduces the compact relationship between carbon dioxide and sulphur hexafluoride, which arises from shared transport histories along the Brewer Dobson circulation (Plumb & Ko, 1992; Konopka et al., 2025). A correct slope indicates proper relative ageing rates of these trace gases.

The observational baseline for these scorecards in this report is derived from the Dynamics and Chemistry of the Summer Stratosphere (DCOTSS) campaign using the data by NASA/LARC/SD/ASDC. Scorecards for further campaigns (see Appendix C) and combinations of them will be done in the next project phase. Every flight is classified into distinct categories (see figure 6) based on the meteorological context, tracer signatures, and temporal proximity to events. Background flights establish the unperturbed baseline against which perturbations are measured, providing a reference state for undisturbed stratospheric transport. Active Convection flights test the ability of the model to capture rapid convective injection, representing the most demanding test of transport and mixing schemes. Recent Convection tests the maintenance of these tracer signatures during the first days of mixing, while Aged Convection evaluates the slow decay of these signals into the background. Flights sampling the North American Monsoon Anticyclone test the representation of large-scale anticyclonic transport and its influence on atmospheric composition through confinement.

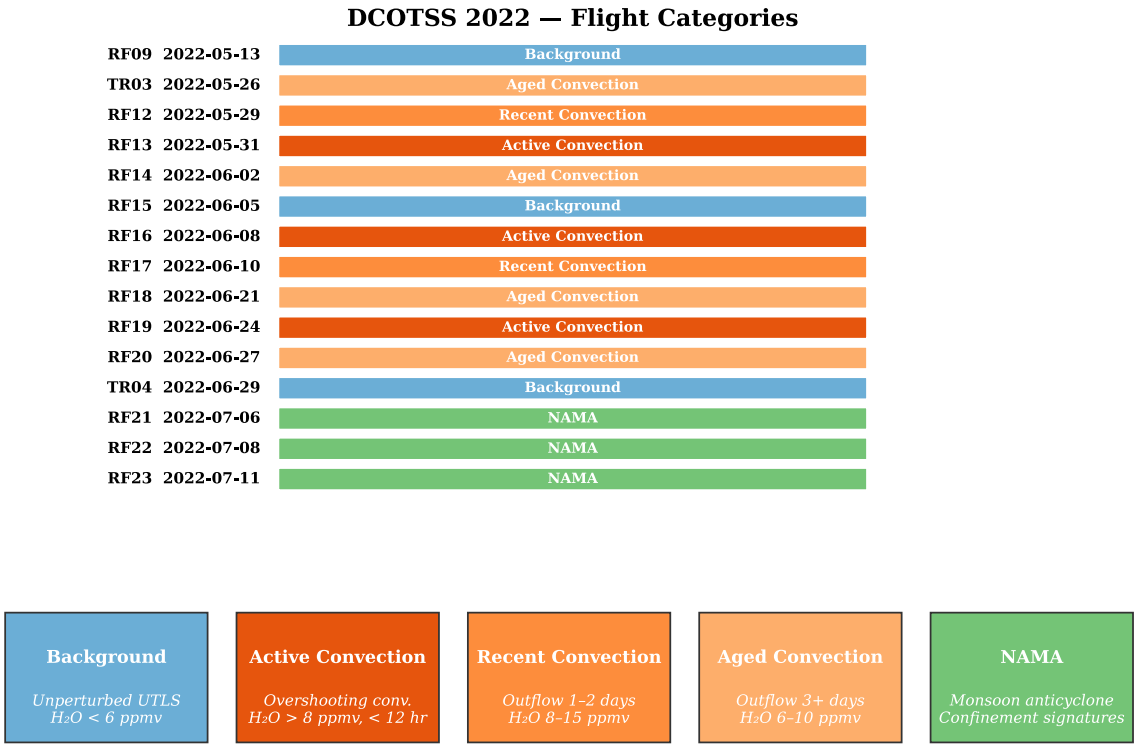

Figure 6. Classification of observation flights from the Dynamics and Chemistry of the Summer Stratosphere campaign into distinct meteorological categories. These categories are used to isolate specific transport phenomena ranging from unperturbed background states to active overshooting convection and monsoon anticyclone confinement. Classification are based on Homeyer et al., (2023) and Gordon et al., (2024) (for Active Convention), DCOTSS, (2022) (for Recent Convection and Aged Convection) and Sayres et al., (2024) (for NAMA)

Furthermore, to diagnose small scale mixing processes, each observation point is classified into a Stratosphere Troposphere Exchange air mass regime using a multiparameter approach (Pan et al., 2004, Pan et al., 2021, Hoor et al., 2004, Gettelman et al., 2011). This classification prioritizes specific thresholds for water vapour to identify special cases, followed by a joint ozone and carbon monoxide classification, and finally an altitude relative to the tropopause fallback. Regimes include the deep stratosphere containing photochemically aged air, the lower stratosphere with recently descended air, and the extra-tropical transition layer representing the key zone where irreversible mixing occurs across the tropopause. Convective injection is explicitly identified by the direct injection of tropospheric water vapour, while stratospheric intrusions are marked by dry stratospheric air moving below the tropopause fold.

A primary application of this scorecard framework is testing model sensitivity to horizontal resolution across these different regimes.

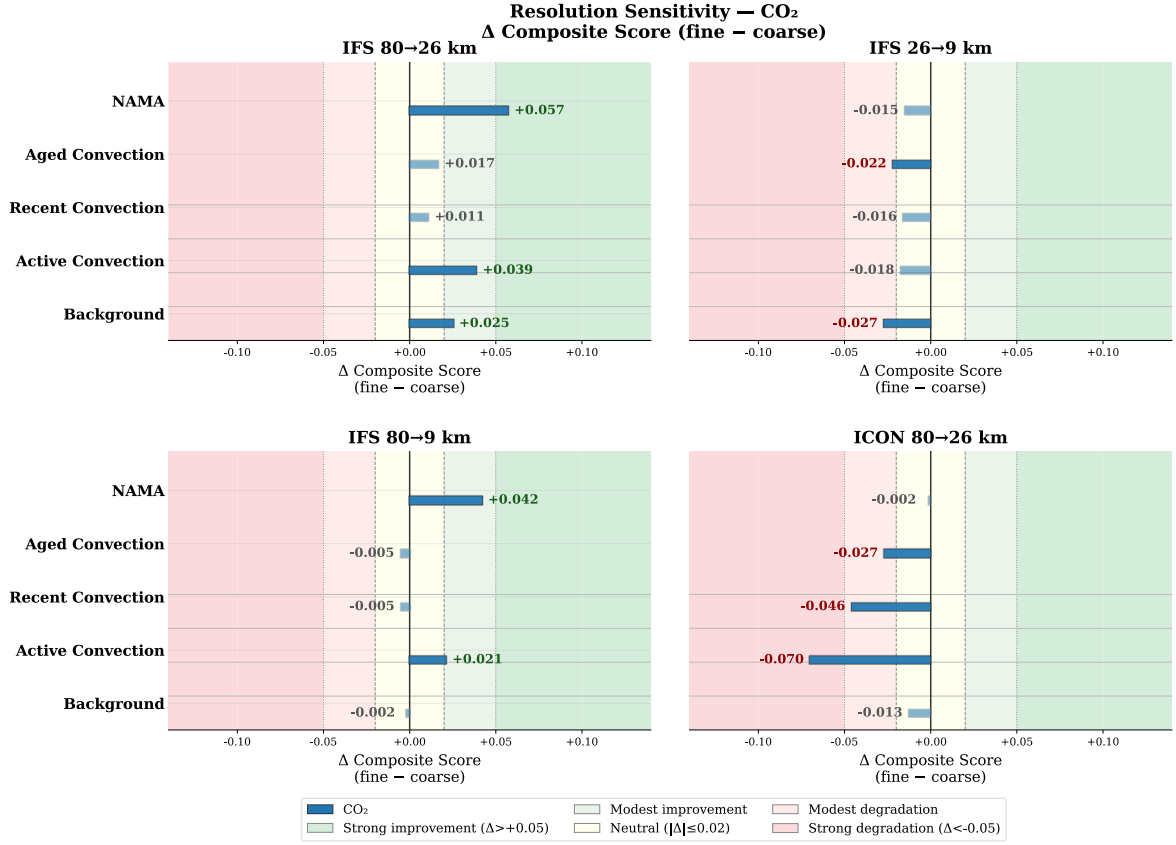

Figure 7. Resolution sensitivity scorecard for carbon dioxide within the extra tropical transition layer. The panels illustrate the change in composite score when increasing horizontal resolution for the ICON-ART model and the Integrated Forecast System. Positive values in green indicate an improvement in the simulation of transport processes at finer resolutions.

Analyzing the resolution sensitivity for carbon dioxide within the extra-tropical transition layer reveals distinct model behaviours (see Figure 7). Upgrading the Integrated Forecast System from eighty kilometers to twenty six kilometers, and further to nine kilometers, generally yields a strong improvement in the composite score across most flight categories. The highest resolution step demonstrates the most substantial performance gains, particularly for flights sampling the North American Monsoon Anticyclone and active convection events. Conversely, upgrading the ICON-ART model from eighty to twenty six kilometers results in a modest to strong degradation of the composite score in this specific atmospheric layer, highlighting that finer resolution does not automatically resolve mixing errors if the underlying parameterizations behave differently at smaller grid scales.

Another critical physical diagnostic utilized in the scorecard is the tracer-tracer correlation (see Figure 8), which isolates transport and mixing errors from pure emission biases.

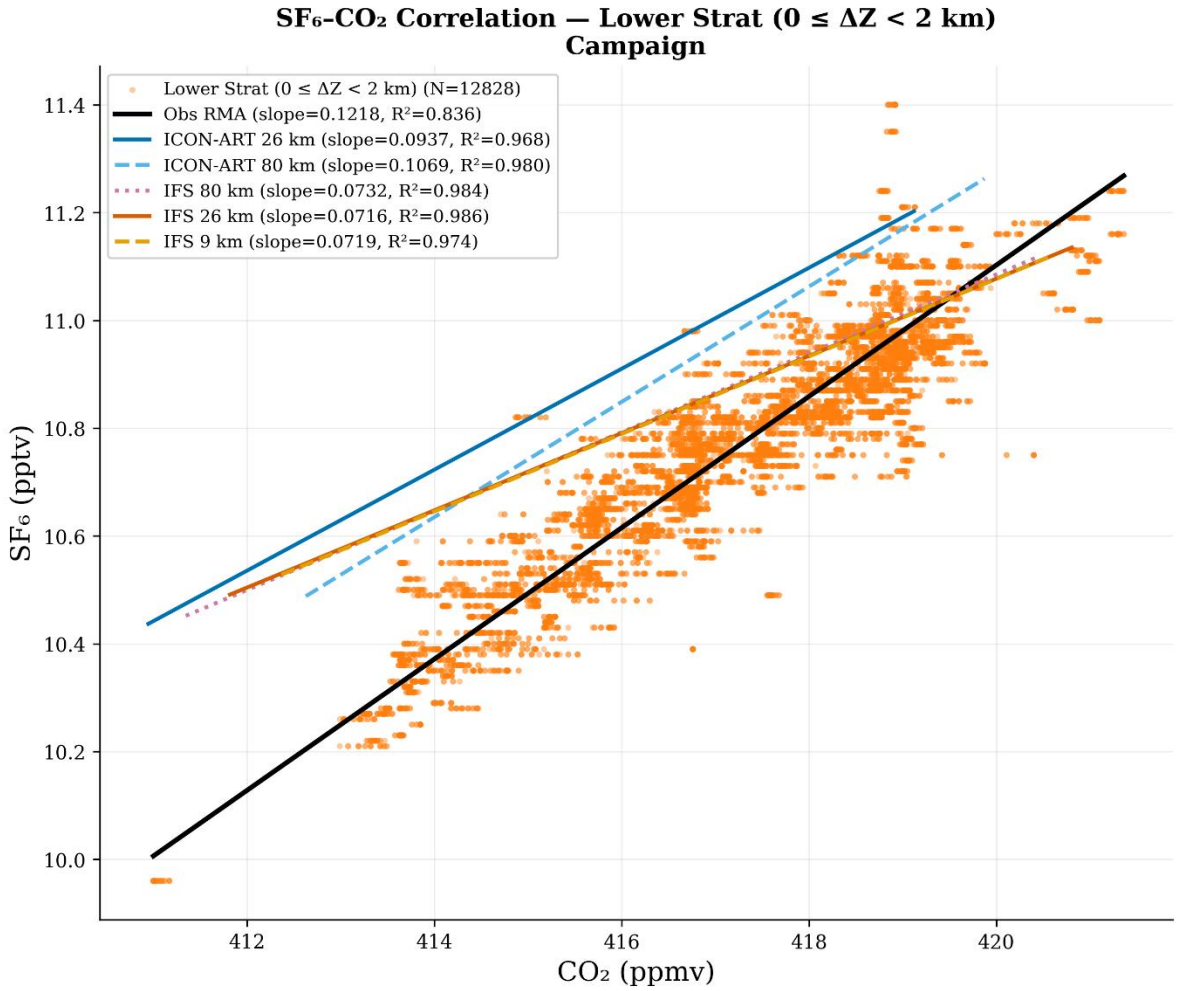

Figure 8. Tracer-tracer correlation between sulfur hexafluoride and carbon dioxide in the lower stratosphere. The solid black line represents the observed relationship, while the coloured lines denote the regression slopes produced by different configurations of the ICON-ART and Integrated Forecast System models.

Examining the relationship between sulphur hexafluoride and carbon dioxide in the lower stratosphere reveals how models represent relative ageing and isentropic mixing along the global circulation pathways. The observational data establish a specific regression slope that models attempt to replicate. The results indicate that all configurations of both the Integrated Forecast System and the ICON-ART model underpredict this slope compared to the observations. This discrepancy suggests systemic challenges across the models in accurately representing the differential transport histories and the precise timescales of these two trace gases as they cross the tropopause boundary.

The culmination of these diverse metrics, flight categories, and physical diagnostics is the complete multi tracer scorecard matrix (see Figure 9).

Campaign Metrics — All Tracers
Raw Values with ΔZ-Region Breakdown

	ICON 26	ICON 80	IFS 80	IFS 26	IFS 9	ICON 26	ICON 80	IFS 80	IFS 26	IFS 9	ICON 26	ICON 80	IFS 80	IFS 26	IFS 9
	CO_2					H_2O					SF_6				
Systematic Error	**0.709**	**0.210**	**1.102**	**0.497**	**0.221**	**3.572**	**3.080**	**2.475**	**1.960**	**1.579**	**0.325**	**0.304**	**0.198**	**0.255**	**0.317**
UpTrop	1.738	0.662	1.039	0.724	0.300	17.456	30.485	22.294	18.206	12.818	0.065	0.070	0.046	0.043	0.051
ExTL	1.628	1.352	0.584	0.220	0.084	3.533	3.218	2.318	1.836	1.810	0.074	0.072	0.008	0.008	0.022
LowStrat	1.995	1.323	0.886	0.500	0.017	1.426	0.077	0.939	0.928	0.940	0.082	0.082	0.055	0.079	0.111
MidUpStrat	0.202	0.856	1.220	0.521	0.312	0.356	0.122	0.081	0.083	0.011	0.455	0.422	0.287	0.363	0.447
Mean Bias	**-0.709**	**0.210**	**-1.102**	**-0.497**	**0.221**	**3.572**	**3.080**	**2.475**	**1.960**	**1.579**	**0.325**	**0.304**	**0.198**	**0.255**	**0.317**
UpTrop	-1.738	-0.662	-1.039	-0.724	-0.300	17.456	30.485	22.294	18.206	12.818	0.065	0.070	-0.046	-0.043	-0.051
ExTL	-1.628	-1.352	-0.584	-0.220	0.084	3.533	3.218	2.318	1.836	1.810	0.074	0.072	-0.008	0.008	0.022
LowStrat	-1.995	-1.323	-0.886	-0.500	-0.017	1.426	-0.077	0.939	0.928	0.940	0.082	0.082	0.055	0.079	0.111
MidUpStrat	-0.202	0.856	-1.220	-0.521	0.312	0.356	-0.122	0.081	0.083	0.011	0.455	0.422	0.287	0.363	0.447
Random Error	**1.594**	**1.634**	**0.808**	**0.788**	**0.855**	**34.324**	**30.433**	**30.252**	**27.077**	**31.269**	**0.316**	**0.285**	**0.209**	**0.239**	**0.277**
UpTrop	0.792	0.557	0.411	0.365	0.384	66.618	65.258	69.312	55.510	65.682	0.103	0.109	0.111	0.105	0.109
ExTL	0.804	0.660	0.741	0.636	0.692	8.966	10.486	8.297	7.049	8.276	0.147	0.129	0.123	0.126	0.134
LowStrat	1.315	1.166	0.900	0.860	0.858	4.844	5.568	4.659	4.531	5.109	0.158	0.150	0.132	0.134	0.141
MidUpStrat	1.450	1.376	0.746	0.771	0.852	1.417	1.186	1.111	1.106	1.231	0.304	0.272	0.186	0.209	0.242
Pearson R	**0.912**	**0.944**	**0.976**	**0.974**	**0.970**	**0.952**	**0.959**	**0.952**	**0.958**	**0.943**	**0.899**	**0.930**	**0.970**	**0.967**	**0.960**
UpTrop	-0.742	0.137	0.705	0.832	0.747	0.941	0.949	0.942	0.960	0.948	0.799	0.775	0.781	0.822	0.822
ExTL	0.724	0.802	0.772	0.819	0.770	0.900	0.879	0.909	0.931	0.913	0.421	0.566	0.612	0.598	0.513
LowStrat	0.690	0.745	0.861	0.871	0.871	0.650	0.505	0.675	0.685	0.663	0.665	0.708	0.814	0.819	0.792
MidUpStrat	0.849	0.916	0.962	0.955	0.944	-0.011	0.031	0.117	0.134	0.183	0.834	0.896	0.950	0.942	0.930
Refined IoA	**0.770**	**0.768**	**0.804**	**0.879**	**0.884**	**0.875**	**0.882**	**0.894**	**0.904**	**0.892**	**0.618**	**0.646**	**0.753**	**0.700**	**0.638**
UpTrop	-0.491	0.078	-0.148	0.169	0.510	0.855	0.838	0.839	0.866	0.852	0.603	0.572	0.603	0.622	0.602
ExTL	-0.098	0.068	0.491	0.664	0.655	0.774	0.766	0.813	0.828	0.819	0.454	0.489	0.621	0.605	0.562
LowStrat	0.303	0.506	0.658	0.740	0.786	0.562	0.607	0.617	0.623	0.616	0.573	0.597	0.674	0.648	0.595
MidUpStrat	0.739	0.682	0.707	0.830	0.831	0.142	0.385	0.440	0.442	0.483	0.352	0.403	0.590	0.491	0.380
AoA Bias (yr)	**-0.838**	**-0.785**	**-0.537**	**-0.686**	**-0.853**	**-0.838**	**-0.785**	**-0.537**	**-0.686**	**-0.853**	**-0.838**	**-0.785**	**-0.537**	**-0.686**	**-0.853**
UpTrop	-0.027	-0.028	-0.025	-0.027	-0.029	-0.027	-0.028	-0.025	-0.027	-0.029	-0.027	-0.028	-0.025	-0.027	-0.029
ExTL	-0.089	-0.090	-0.049	-0.078	-0.095	-0.089	-0.090	-0.049	-0.078	-0.095	-0.089	-0.090	-0.049	-0.078	-0.095
LowStrat	-0.223	-0.231	-0.167	-0.231	-0.311	-0.223	-0.231	-0.167	-0.231	-0.311	-0.223	-0.231	-0.167	-0.231	-0.311
MidUpStrat	-1.229	-1.141	-0.776	-0.982	-1.208	-1.229	-1.141	-0.776	-0.982	-1.208	-1.229	-1.141	-0.776	-0.982	-1.208
Taylor Skill Score (UTLS)	**0.859**	**0.801**	**0.975**	**0.976**	**0.967**	**0.928**	**0.907**	**0.940**	**0.951**	**0.946**	**0.704**	**0.749**	**0.819**	**0.746**	**0.662**
Bergen FB (UTLS)	**0.715**	**0.874**	**0.821**	**0.908**	**0.988**	**0.927**	**0.977**	**0.959**	**0.962**	**0.967**	**0.687**	**0.704**	**0.804**	**0.728**	**0.642**
Trop. Gradient	**0.58**	**0.64**	**0.70**	**0.78**	**0.74**	**0.62**	**0.58**	**0.71**	**0.76**	**0.78**	**0.65**	**0.72**	**0.64**	**0.64**	**0.59**
Trans. Width	**0.68**	**0.82**	**0.79**	**0.81**	**0.75**	**0.76**	**0.75**	**0.78**	**0.75**	**0.74**	**0.71**	**0.71**	**0.72**	**0.72**	**0.72**

Figure 9. Comprehensive scorecard presenting raw metric values broken down by specific altitude regions relative to the tropopause. The table aggregates systematic error, random error, correlation coefficients, and specialized transport metrics across carbon dioxide, water vapour, and sulfur hexafluoride for all model resolutions.

This comprehensive summary provides raw metric values broken down by specific altitude regions relative to the tropopause, explicitly isolating the upper troposphere, the extra-tropical transition layer, the lower stratosphere, and the middle to upper stratosphere. By aggregating standard statistical errors alongside physically meaningful transport metrics for carbon dioxide, water vapour, and sulphur hexafluoride, the scorecard enables a harmonized identification of structural model weaknesses. This structured approach mirrors the boundary layer evaluation framework, ensuring a consistent methodology for diagnosing tracer transport from the surface up to the stratosphere.

A publication based on these results is currently in preparation (Qor-El-Aine et al., 2026b).

4 Conclusions and Recommendations

This deliverable introduces a scorecard framework designed to systematically evaluate atmospheric transport processes in global numerical models used for greenhouse gas monitoring. The approach provides a structured methodology to assess how unresolved transport processes—such as turbulent mixing, cloud transport, and boundary-layer exchange—are represented in large-scale models.

The scorecard framework integrates multiple sources of information, including observations, large-eddy simulations, and sensitivity experiments within global models. By combining these comparison pathways with diagnostic metrics such as boundary-layer height, concentration gradients, diurnal variability, and flux ratios, the methodology allows a comprehensive evaluation of model transport performance across different atmospheric regimes and ecosystem types.

The prototype applications presented in this deliverable demonstrate the capability of the framework to diagnose specific transport processes. The evaluation of vertical CO_2 gradients highlights the sensitivity of tracer distributions to parameterized turbulent mixing, particularly under stable boundary-layer conditions. The CO_2 budget decomposition further shows how the scorecard methodology can attribute model behaviour to individual physical processes such as surface fluxes, entrainment, cloud transport, and horizontal advection in a systematic manner.

A key strength of the approach is its flexibility. The scorecards can be applied across different spatial and temporal scales, from local testbed analyses to regional and global simulations, and can be used to assess the impact of changes in model resolution, physical parameterizations, or stochastic perturbations. The framework therefore provides a consistent tool for diagnosing model sensitivities and identifying sources of transport uncertainty.

The methodology developed here forms the foundation for subsequent analyses within the CATRINE project. In particular, the scorecards will be further applied to evaluate transport processes in both the ABL–FT and UTLS regions and to assess the impact of model developments and parameterization changes. Ultimately, this work contributes to improving the representation of atmospheric transport in global models and supports the development of more reliable greenhouse gas monitoring and verification systems.

Preliminary recommendations are:

1. A systematic application of the scorecards across different locations and ecosystem types is essential to identify robust, process-level model biases and to assess the impact of model resolution. Given the wide range of configurations outlined in Figure 1, it is recommended to apply a harmonized set of scorecard configurations to ensure consistency and comparability across cases.
2. The selection of comprehensive observational sites (e.g. ICOS and WMO stations), providing co-located meteorological and greenhouse gas measurements, is crucial to enhance the robustness and interpretability of the scorecard analysis.

With respect to boundary-layer and free-troposphere exchange processes, the following preliminary findings emerge:

a. For applications in inverse modelling, transport-related variables should first be evaluated using the scorecard framework to assess their performance and sensitivity to definitions and ecosystem-specific conditions. A key example is the atmospheric boundary layer height, whose definition and representation can significantly influence transport diagnostics.
b. The representation of physical processes controlling greenhouse gas variability remains a critical source of uncertainty. The examples presented in this deliverable highlight systematic errors associated with the nocturnal stable boundary layer, particularly due to the model's limited ability to reproduce strong vertical gradients and the coarse representation of canopy processes.

With respect to upper troposphere-lower stratosphere (UTLS) exchange processes, the following preliminary findings emerge:

a. Horizontal resolution alone is not a universal fix for transport errors. While the IFS model showed strong improvement moving to finer grids (9 km), the ICON-ART model demonstrated that higher resolution can sometimes degrade performance if the underlying parameterizations for mixing and diffusion are not also refined for those scales.
b. The representation of differential transport histories remains a systemic challenge. Tracer-tracer correlations (such as the CO2-SF6 slope) reveal that models consistently struggle to accurately reproduce the relative ageing rates and precise timescales of trace gases as they cross the tropopause boundary into the lower stratosphere.
c. Utilizing targeted flight campaign data (e.g., DCOTSS) and categorizing observations by meteorological regime (such as active convection versus unperturbed background) is highly effective for isolating and diagnosing specific model transport weaknesses that are otherwise smoothed out in global averages.

Future work will aim to extend these prototypes to comprehensive ICOS station that cover the more important global ecosystems. Similar as presented in these report, the intercomparison of the ICOS observations with IFS or similar global models will map and assess systemic errors associated to transport processes at different spatiotemporal scales. For the UTLS the prototype shown will be extended to the campaigns given in Appendix C.

5 References

Bowman, Kenneth P., and Coauthors. *The Dynamics and Chemistry of the Summer Stratosphere (DCOTSS) Project*. Bulletin of the American Meteorological Society 107, no. 3 (2026). https://doi.org/10.1175/BAMS-D-24-0177.1

Chang, J. C., & Hanna, S. R. (2004). *Air quality model performance evaluation*. Meteorology and Atmospheric Physics, 87(1–3), 167–196. https://doi.org/10.1007/s00703-003-0070-7

de Feiter V. et al. Quantifying the contribution of individual processes in determining the CO2 diurnal variability/ Atmospheric Chemistry and Physics (in preparation), 2026.

Dynamics and Chemistry of the Summer Stratosphere Team. *DCOTSS Mission Overview, Data Access, and Data Analysis.* Presentation at the DCOTSS Open Data Workshop 2022, December 8, 2022. https://dcotss.org/workshops/2022/DCOTSS-ODW2022-Overview.pdf

Engel, A., Möbius, T., Bönisch, H., Schmidt, U., Heinz, R., Levin, I., Atlas, E., Aoki, S., Nakazawa, T., Sugawara, S., Moore, F., Hurst, D., Elkins, J., Schauffler, S., Andrews, A., & Boering, K. (2009). *Age of stratospheric air unchanged within uncertainties over the past 30 years*. Nature Geoscience, 2, 28–31. https://doi.org/10.1038/ngeo388

Faassen, K. A. P., González-Armas, R., Koren, G., Adnew, G. A., vanAsperen, H., deBoer, H., et al. (2025). Tracing diurnal variations of atmospheric CO2, O2, and δ13 ${\delta }^{13}$CO2 over a tropical and a temperate forest. Geophysical Research Letters, 52, e2025GL118016. **https://doi.org/10.1029/2025GL118016**

Gettelman, A., Hoor, P., Pan, L. L., Randel, W. J., Hegglin, M. I., & Birner, T. (2011). *The extratropical upper troposphere and lower stratosphere*. Reviews of Geophysics, 49(3), RG3003. https://doi.org/10.1029/2011RG000355

Gordon, A. E., Homeyer, C. R., Smith, J. B., Ueyama, R., Dean-Day, J. M., Atlas, E. L., Smith, K., Pittman, J. V., Sayres, D. S., Wilmouth, D. M., Pandey, A., St. Clair, J. M., Hanisco, T. F., Hare, J., Hannun, R. A., Wofsy, S., Daube, B. C., & Donnelly, S. (2024). *Airborne observations of upper troposphere and lower stratosphere composition change in active convection producing above-anvil cirrus plumes*. Atmospheric Chemistry and Physics, 24, 7591–7608. https://doi.org/10.5194/acp-24-7591-2024

Hanna, S. R., Strimaitis, D. G., & Chang, J. C. (1991). *Evaluation of fourteen hazardous gas models with ammonia and hydrogen fluoride field data*. Atmospheric Environment, 25A(4–5), 1229–1239. https://doi.org/10.1016/0960-1686(91)90253-U

Homeyer, C. R., Gordon, A. E., Smith, J. B., Ueyama, R., Wilmouth, D. M., Sayres, D. S., Hare, J., Pandey, A., Hanisco, T. F., Dean-Day, J. M., Hannun, R., & St. Clair, J. M. (2024). *Stratospheric hydration processes in tropopause-overshooting convection revealed by tracer-tracer correlations from the DCOTSS field campaign*. Journal of Geophysical Research: Atmospheres, 129, e2024JD041340. https://doi.org/10.1029/2024JD041340

Homeyer, C. R., Smith, J. B., Bedka, K. M., Bowman, K. P., Wilmouth, D. M., Ueyama, R., Dean-Day, J. M., St. Clair, J. M., Hannun, R., Hare, J., Pandey, A., Sayres, D. S., Hanisco, T. F., & Gordon, A. E. (2023). *Extreme altitudes of stratospheric hydration by midlatitude convection observed during the DCOTSS field campaign*. Geophysical Research Letters, 50, e2023GL104914. https://doi.org/10.1029/2023GL104914

Hoor, P., Gurk, C., Brunner, D., Hegglin, M. I., Wernli, H., & Fischer, H. (2004). *Seasonality and extent of extratropical TST derived from in-situ CO measurements during SPURT*. Journal of Geophysical Research: Atmospheres, 109, D15309. https://doi.org/10.5194/acp-4-1427-2004

Konopka, P., Ploeger, F., D'Amato, F., Campos, T., von Hobe, M., Honomichl, S. B., Hoor, P., Pan, L. L., Santee, M. L., Viciani, S., Walker, K. A., & Hegglin, M. I. (2025). *Isentropic mixing vs. convection in CLaMS-3.0/MESSy: Evaluation using satellite climatologies and in situ carbon monoxide observations*. Atmospheric Chemistry and Physics, 25, 17973–17996. https://doi.org/10.5194/acp-25-17973-2025

Murphy, A. H., & Epstein, E. S. (1989). *Skill scores and correlation coefficients in model verification*. Monthly Weather Review, 117(3), 572–581. https://doi.org/10.1175/1520-0493(1989)117\<0572:SSACCI>2.0.CO;2

NASA/LARC/SD/ASDC. *Dynamics and Chemistry of the Summer Stratosphere Airborne Data Products*. NASA Langley Atmospheric Science Data Center DAAC, May 19, 2022. https://doi.org/10.5067/ASDC/DCOTSS-Aircraft-Data_1

Pan, L. L., Randel, W. J., Gary, B. L., Mahoney, M. J., & Hintsa, E. J. (2004). *Definitions and sharpness of the extratropical tropopause: A trace gas perspective*. Journal of Geophysical Research: Atmospheres, 109, D23103. https://doi.org/10.1029/2004JD004982

Pan, L. L., Honomichl, S. B., Kinnison, D. E., Abalos, M., Randel, W. J., Paulik, J. W., & Thornberry, T. D. (2021). *Transport of chemical tracers from the boundary layer to stratosphere associated with the dynamics of the Asian summer monsoon*. Atmospheric Chemistry and Physics, 21, 595–611. https://doi.org/10.5194/acp-21-595-2021

Pan, L. L., Atlas, E. L., Newman, P. A., Thornberry, T., Jucks, K. W., Toon, O. B., et al. (2025). *The Asian summer monsoon chemical and climate impact project (ACCLIP): An*

overview. Journal of Geophysical Research: Atmospheres, 130, e2025JD044417. https://doi.org/10.1029/2025JD044417

Petzold, Andreas, Valerie Thouret, Christoph Gerbig, Andreas Zahn, Carl A. M. Brenninkmeijer, Martin Gallagher, Markus Hermann, et al.: *Global-Scale Atmosphere Monitoring by In-Service Aircraft - Current Achievements and Future Prospects of the European Research Infrastructure IAGOS*. Tellus B: Chemical and Physical Meteorology 67, no. 1 (2015). https://doi.org/10.3402/tellusb.v67.28452

Pino, D., et al.: *A conceptual framework to quantify the influence of convective boundary layer development on carbon dioxide mixing ratios*, Atmos. Chem. Phys., 12, 2969–2985, https://doi.org/10.5194/acp-12-2969-2012 , 2012

Plumb, R. A. (2007), *Tracer interrelationships in the stratosphere*, Rev. Geophys., 45, RG4005, https://doi.org/10.1029/2005RG000179

Qor-el-aine, A., Versick, S., Agusti-Panareda, A., and others.: *Assessing Vertical Transport of CO2 in Warm Conveyor Belts in ICON-ART and IFS models during PHILEAS Campaign*, in preparation, 2026.

Qor-el-aine, A., Versick, S., Agusti-Panareda, A., and others.: *Evaluation of stratospheric CO2 transport in global models using DCOTSS aircraft observations*: Implications for convective injection, in preparation, 2026b

Riese, M., & Hoor, P. (n.d.). *WISE: Dynamics & chemistry of mid-latitude upper troposphere / lower stratosphere*. HALO (High Altitude and Long Range Research Aircraft). Retrieved [Insert Date You Accessed the Site], from https://halo-research.de/science/previous-missions/wise/

Riese, M., Hoor, P., and Coauthors, 2025: *Long-range transport of polluted Asian summer monsoon air to high latitudes during the PHILEAS campaign in the boreal summer 2023*. Bull. Amer. Meteor. Soc., , BAMS-D-24-0232.1, https://doi.org/10.1175/BAMS-D-24-0232.1

Sayres, D. S., Smith, J. B., Wilmouth, D. M., Pandey, A., Homeyer, C. R., Bowman, K. P., & Anderson, J. G. (2024). *Using the NAMA as a natural integrator to quantify the convective contribution to lower stratospheric water vapor over North America*. Journal of Geophysical Research: Atmospheres, 129, e2024JD041641. https://doi.org/10.1029/2024JD041641

StratoClim Consortium. *StratoClim Factsheet 2: The StratoClim Aircraft Field Campaign*. Alfred Wegener Institute. Updated 2017. https://www.stratoclim.org/downloads/StratoClim_Factsheet_2-Aircraft_Campaign_updated_highres.pdf

Taylor, K. E. (2001). *Summarizing multiple aspects of model performance in a single diagram*. Journal of Geophysical Research: Atmospheres, 106(D7), 7183–7192. https://doi.org/10.1029/2000JD900719

Willmott, C. J., Robeson, S. M., & Matsuura, K. (2012). *A refined index of model performance*. International Journal of Climatology, 32(13), 2088–2094. https://doi.org/10.1002/joc.2419

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

6 Appendix A: Information on TestBed

The complete technical report of the TestBeds can be found in the following link

7 Appendix B: Scorecards transport processes. Diurnal variability.

Scorecard — CloudRoots-Amazon22 — Sensitivity Framework: Shallow Convective Regime - 9 LT

Term	Physical Meaning	DALES	ECMWF 25 km	ECMWF 9 km	Δ (ECMWF 25 km − DALES)	Δ (ECMWF 9 km − DALES)
II	En/detrainment during the night-to-day transition	78.84	74.01	64.73	-4.83	-14.11
III	Surface fluxes (plant assimilation and soil respiration)	15.74	8.61	7.05	-7.12	-8.69
IV	En/detrainment determined by the CO_2 lapse rate	1.5	1.98	1.95	0.48	0.45
V	Boundary layer dilution associated with clouds	0.0	0.0	0.37	0.0	0.37
VI	Cloud mass flux	0.0	0.0	1.56	-0.0	1.56
VII	Horizontal advection	-	3.48	9.5	3.48	9.5
R	Residual	3.93	11.92	14.84	7.99	10.91
Init. Error	Init. Error	8.23	6.96	1.82	-1.27	-6.42

Scorecard — CloudRoots-Amazon22 — Sensitivity Framework: Shallow Convective Regime - 12 LT

Term	Physical Meaning	DALES	ECMWF 25 km	ECMWF 9 km	Δ (ECMWF 25 km − DALES)	Δ (ECMWF 9 km − DALES)
II	En/detrainment during the night-to-day transition	59.34	35.4	31.65	-23.94	-27.69
III	Surface fluxes (plant assimilation and soil respiration)	31.31	6.91	3.82	-24.4	-27.49
IV	En/detrainment determined by the CO_2 lapse rate	5.03	11.7	12.8	6.68	7.77
V	Boundary layer dilution associated with clouds	1.23	0.39	0.84	-0.83	-0.39
VI	Cloud mass flux	2.91	5.58	8.29	2.67	5.38
VII	Horizontal advection	-	9.98	13.65	9.98	13.65
R	Residual	0.19	30.04	28.96	29.85	28.77
Init. Error	Init. Error	1.82	2.78	4.42	0.96	2.59

Figure B2.1 Quantitative CO_2 budget scorecard for the Amazonian testbed under shallow convective conditions at 09 LT (Top) and 15 LT (bottom)). The table shows the relative contribution (%) of each process term to the reconstructed boundary-layer CO_2 mole fraction, comparing large-eddy simulations (DALES) with ECMWF-IFS at 25 km and 9 km resolutions. Differences (Δ) with respect to DALES are provided to highlight model biases. The residual term (R) represents the remaining imbalance of the budget (differences between left hand side and right hand side in equation (1), and "Init. Error" indicates the mismatch (ppm) between reconstructed and observed CO_2 mole fraction.

8 Appendix C: Overview of campaigns in the UTLS

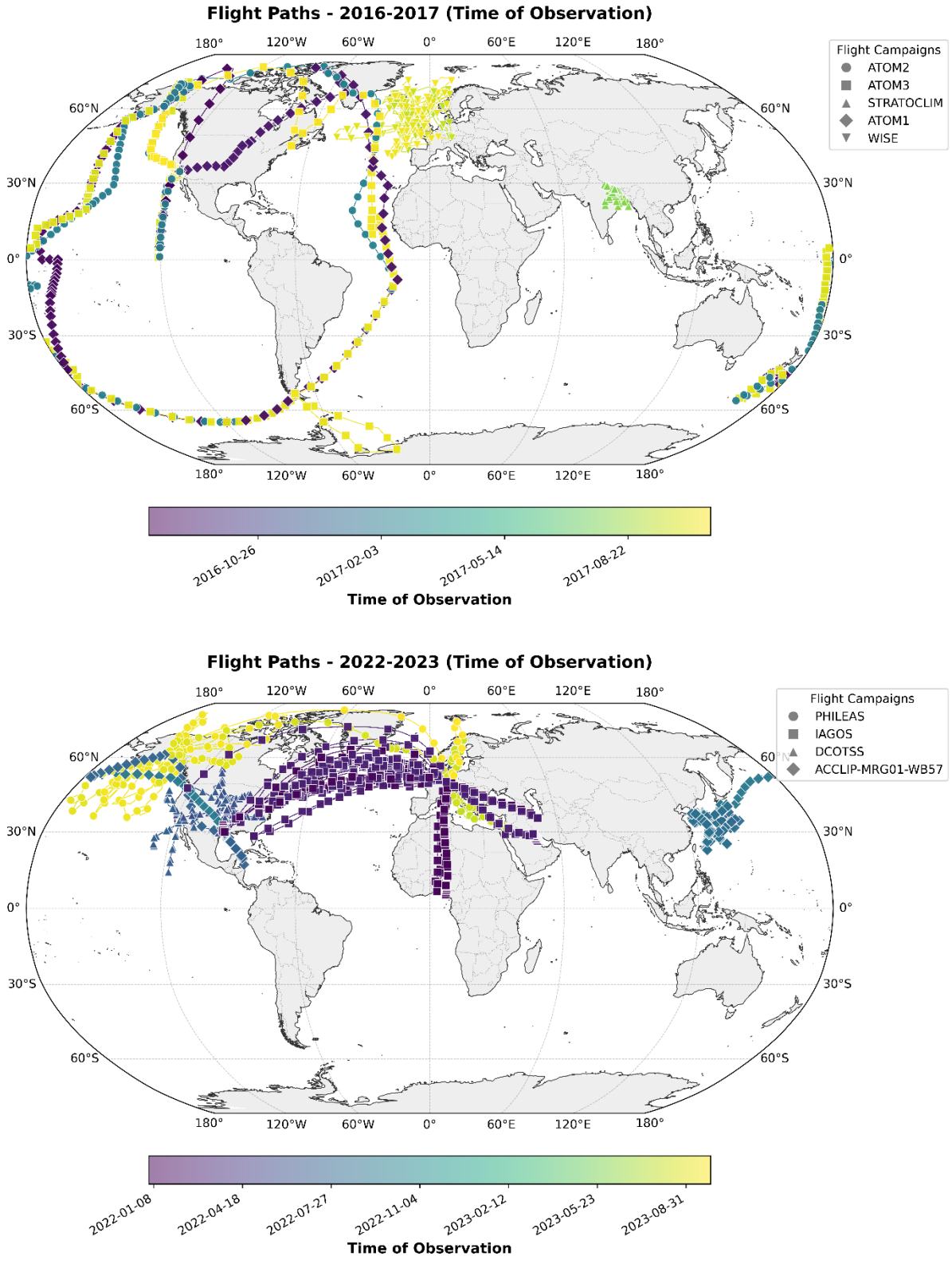

Figure C.1 Overview of flight campaigns for the UTLS during the chosen time frames (2016/17: top; 2022/23: bottom) that will be analyzed in detail during the final project phase. Symbol indicates the different campaigns, colours the date.

Campaign – archetype – spatiotemporal evaluation matrix

Which UTLS features each campaign can evaluate

Campaign	*Archetypes*				*Temporal*		*Spatial*	
	Extratropical	ASM	WCB	NAMA	Synoptic	Seasonal	Regional	Global
DCOTSS	●			●	●		●	
PHILEAS	●	●	●		●		●	
StratoClim		●			●		●	
WISE	●		●		●		●	
IAGOS	●		●		●	●	●	●
ATom	●			●	●	●		●
ACCLIP		●			●		●	

● = evaluated by campaign

DCOTSS — Deep convective overshooting over central US within the NAMA.

PHILEAS — HALO flights Aug–Oct 2023; ASM outflow into NH extratropics, WCB events over North Atlantic.

StratoClim — M-55 Geophysica from Kathmandu, Jul–Aug 2017; first in situ sampling of ASM anticyclone up to 20 km.

WISE — HALO from Shannon, Sep–Oct 2017; wave-driven isentropic exchange, Rossby wave breaking, WCB-related mixing over North Atlantic.

IAGOS — Continuous commercial aircraft measurements; long-term extratropical UTLS including trans-Atlantic corridors and WCB encounters.

ATom — NASA DC-8 global circuits (2016–2018); four seasonal deployments profiling 0.15–13 km over remote Pacific and Atlantic.

ACCLIP — NASA WB-57 and NCAR GV from Osan, South Korea, Jul–Aug 2022; ASM convective transport to western Pacific UTLS.

Figure C.2 Overview of flight campaigns for the UTLS and which of the analysis shown in Figure 2 can be done with them

Document History

Version	Author(s)	Date	Changes
0.1 (Initial document created)	Jordi Vilà	01-03-2026	Structure and coordination with Vincent de Feiter and Alessandro Savazzi. Meeting with Anna-Agusti Panareda and Stefan Versick
1.0	Jordi Vilà	24-03-2026	First complete draft of the project
2.0	Jordi Vilà	27-03-2026	Feedback is implemented after a general discussion hold 26-03-26. Main changes are figure 1 and writing recommendations
3.0	Stefan Versick	08-04-2026	Added UTLS
4.0	Jordi Vilà	10-04-2026	Final complete version to be revised
5.0	Stefan Versick	14-04-2026	Improved UTLS section
Final for revision	Jordi Vila	21-04-2026	Internal reviews will review it. Back 27th April 2026
Revised	Anna-Agusti Panareda	28-04-2026	Internal review
Final version	Jordi Vilà	29-04-2026	Processing comments. Submitting deliverable.

Internal Review History

Internal Reviewers	Date	Comments
Anna-Agusti Panareda	April 2026	Minor corrections