

Interpretable Data Quality Assessment for Energy and Emissions Monitoring: A Semiconductor Manufacturing Case Study in Singapore

Dingyi Zeng

Sussex Artificial Intelligence Institute, Zhejiang Gongshang University, Hangzhou, Zhejiang 310018, China

2337020114@pop.zjgsu.edu.cn

Abstract: Industrial energy and emissions analyses depend on transformations that connect activity records, conversion coefficients and aggregated indicators. An inconsistency in these transformations can resemble an operational anomaly while having a different physical and numerical meaning. This study develops and applies a rule-based framework that combines numerical consistency, semantic alignment and calculation-dependency analysis to explain anomalies in an industrial reporting dataset. A retrospective case study uses an eleven-sheet workbook supplied for a semiconductor facility in Singapore, containing 176.149 GWh of purchased electricity and a combined scope 1 and scope 2 inventory of 173,136.469 t CO₂e. Cross-view reconciliation identifies a 1,152.144 t CO₂e difference attributable to two source contributions excluded from one subtotal. A gas-factor mismatch produces a different response: replacing the displayed C₄F₈ coefficient changes a reverse-calculated quantity but leaves the emissions total unchanged. A fixed-mass calculation would instead suggest a hypothetical difference of 172.105 t CO₂e, illustrating the importance of dependency direction. Monthly auxiliary gas records exhibit a sixteen-fold change in an implied conversion multiplier, whereas the electricity factor cannot uniquely resolve competing reporting-year labels. The results distinguish aggregation, semantic, conversion and temporal inconsistencies within one dataset. The methodological contribution is an explanation procedure connecting each detected inconsistency to its computational mechanism and downstream effect. The case supports the use of dependency-aware data quality assessment before energy-performance modelling, while further datasets are needed to evaluate detection performance and transferability.

Keywords: Industrial energy monitoring; Data quality; Explainable analytics; Greenhouse gas inventory; Semiconductor manufacturing; Calculation dependencies.

1. Introduction

Energy-intensive manufacturing generates heterogeneous records: electricity purchases, fuel consumption, process-gas quantities, conversion factors and environmental performance indicators. These records are often combined at different temporal and organisational resolutions. Their usefulness for monitoring depends on whether the transformations between them preserve meaning. A numerically plausible total can conceal an omitted source, while an apparently incorrect coefficient may affect only a derived reporting field. Understanding such differences is necessary when energy and emissions data are reused for performance analysis.

Research on energy-consumption anomaly detection addresses unusual behaviour in observations and model residuals, including the role of artificial intelligence in identifying departures from expected consumption [6]. A complementary problem arises before an operational model is fitted: the recorded variables may already contain inconsistencies in units, identity, temporal coverage or aggregation. Spreadsheet research has established the importance of errors within calculation structures [1]. For industrial environmental datasets, this problem intersects with accounting conventions. Electricity-related emissions follow an activity-factor relationship, whereas semiconductor process-gas emissions can involve gas utilisation, abatement and by-product formation [4, 5]. A single displayed quantity can therefore have different meanings depending on its position in the calculation chain.

The GHG Protocol Corporate Standard and ISO 14064-1 define the accounting context for organisational inventories [2, 3]. These frameworks help establish source boundaries and reporting conventions, but the interpretation of a specific data inconsistency still requires an examination of the implementation. In particular, an emissions factor displayed beside a gas quantity is not sufficient evidence that the emissions result was calculated from that quantity. The reverse relationship is also possible. Treating every factor discrepancy as an emissions error would overlook this distinction.

This study asks how rule-based data quality assessment can distinguish inconsistencies by their computational mechanism and consequence in an industrial energy and emissions dataset. It examines three related questions: which relationships can be evaluated with the available records; how dependency tracing changes the explanation of a detected inconsistency; and which conclusions can be transferred from the reporting dataset to subsequent monitoring applications. The contribution is an integrated analytical procedure, demonstrated through a semiconductor case, rather than a new anomaly-detection algorithm. The procedure combines relationship tests with formula tracing and controlled perturbation, making the explanation of each finding explicit.

The term anomaly is used here for a departure from a specified relationship within the recorded data. Operational anomalies, such as abnormal equipment consumption, require time-resolved measurements and operating context. The present case concerns the data-processing layer that would supply such analyses. This distinction allows the study to

address energy monitoring without attributing the observed inconsistencies to unobserved sensor or equipment faults.

2. Materials and Methods

2.1. Case Context and Dataset

The dataset comprises a legacy Excel workbook with eleven worksheets and an accompanying training presentation. The provider describes the case as a semiconductor facility in Singapore, referred to as Facility S. The workbook combines annual energy and emissions calculations with monthly auxiliary process-gas records. Four data views are used in the analysis: activity inputs and source subtotals (S1), fluorinated-gas calculations (S2), the inventory summary (S3), and monthly carbon-monoxide records (S4). These labels preserve the analytical relationships while removing company-specific sheet names.

The selected summary records purchased electricity of 176,148,540 kWh and direct emissions of 101,179.791 t CO_{2e}. Purchased-electricity emissions add 71,956.679 t CO_{2e}. Fluorinated process gases account for 92,586.091 t CO_{2e} and other direct sources for 8,593.699 t CO_{2e}. These values characterise the supplied dataset; the study does not estimate sector-wide performance. The monthly CO records are examined as auxiliary process data. They are neither an electricity series nor a direct GHG category in this inventory.

No interval meter exports, production volumes, calibration records or independently measured emissions were available. The dataset therefore supports a retrospective study of data transformations and their consistency. The lack of production denominators prevents comparison of energy intensity across facilities, and competing year labels prevent an unambiguous assignment of the activity period. Scope 3 is outside the quantitative analysis.

2.2. Analytical Framework

The framework evaluates three linked properties. Numerical consistency concerns whether related values reproduce the specified arithmetic. Semantic alignment concerns whether the gas identity, unit, period and coefficient describe the same quantity. Dependency consistency concerns whether the interpretation of an apparent relationship agrees with the direction of the implemented calculation. Together, these properties produce an explanation consisting of a violated relationship, its computational mechanism and its effect on a selected output.

For an activity quantity A and factor f , the elementary calculation is $E = Af$. Numerical consistency is represented by the residual $r = E_{\text{recorded}} - Af$. For two views of a common inventory, the reconciliation residual is $\Delta = E_{\text{summary}} - E_{\text{activity}}$ after conversion to common units. A residual identifies a disagreement; tracing its contributing terms establishes whether it arises from a source-selection difference or another transformation. An absolute tolerance of 1 kg CO_{2e} was used for emissions reconciliation. Year labels and displayed coefficients were compared exactly.

Dependency analysis treats the relevant formula chain as a directed graph: a link from variable x to variable y means that y uses x in its calculation. Only the focal paths needed to explain the detected inconsistencies were interpreted. A controlled change in a selected input or reference was then propagated by recalculation. The resulting change in a downstream output was compared with the hypothesised activity–factor response. This test is especially useful when a

displayed mass may be reconstructed from an emissions result.

Table 1 summarises the analytical relationships. Sixteen checks were consolidated retrospectively from the exploratory analysis and rerun for the final assessment. The complete conditions, cell references and results are retained in the supplementary check register. The checks are dependent and purposively selected, so their outcomes do not estimate an error prevalence or classifier accuracy.

Table 1. Relationships evaluated in the case study

Dimension	Relationship	Interpretive purpose
Numerical	$E = Af$; component sums; cross-view totals	Separate arithmetic agreement from aggregation differences
Semantic	Gas–factor identity; units; reporting period	Determine whether related fields refer to compatible quantities
Dependency	Formula direction and recalculated response	Identify which quantities are affected by a changed reference
Temporal conversion	Monthly implied multiplier $k = wP/p$	Locate discontinuities in recorded conversion behaviour

2.3. Formula Extraction and Computational Checks

The original binary workbook was retained unchanged. Formula records, including shared formulas, were decoded with xlrd; 3,400 formula cells were identified. One formula outside the focal interpreted paths could not be fully decoded by the native parser. The relevant paths were cross-checked after LibreOffice conversion to .xlsx and inspected with openpyxl. A diagnostic copy was saved with automatic and forced-full-recalculation settings, then opened and saved in headless LibreOffice.

Recalculated values were compared with all 5,945 originally numeric cells using a tolerance of $\max(10^{-6}, 10^{-10} \times |\text{original value}|)$. A separate diagnostic copy changed only the reference supplying the displayed C_{4F}₈ coefficient. This intervention tested the local dependency response; it did not replace the source workbook or redesign its emissions methodology. The supplementary material provides the selected formula extracts, numerical results and an arithmetic reproduction script.

2.4. Sensitivity and Monthly Conversion Analysis

Two quantities were distinguished in the gas-factor analysis. The first was a hypothetical fixed-mass sensitivity, $\Delta E_{\text{fixed}} = m(f_1 - f_2)$, where the originally displayed mass m was held constant. The second was the observed change in the emissions output after changing the factor reference in a diagnostic copy. Comparing these quantities tests whether the fixed-mass assumption represents the implemented calculation.

For each monthly CO record, an implied multiplier was reconstructed as $k_m = w_m P_m / p_m$, where w_m is the stored mass, P_m the filling-pressure entry and p_m the consumed-pressure entry. All twelve records contained finite positive values. The resulting sequence describes the algebraic relationship among the stored fields. It is not a calibrated physical gas-law model, since temperature, residual pressure and the detailed aggregation convention were unavailable.

3. Results

3.1. Reproducible Totals and Unresolved Temporal Meaning

All 5,945 originally numeric cells agreed with the recalculated copy within the specified tolerance. The summary components also reconciled, and electricity

multiplication reproduced the stored value to 0.01 kg CO₂: $176,148,540 \text{ kWh} \times 0.4085 \text{ kg CO}_2/\text{kWh} = 71,956,678.59 \text{ kg CO}_2$. The calculation uses a grid-average factor and is interpreted as location-based [5]. The factor expresses CO₂; the workbook's CO₂e notation is retained for its combined inventory totals. Figure 1 shows the composition of the selected summary.

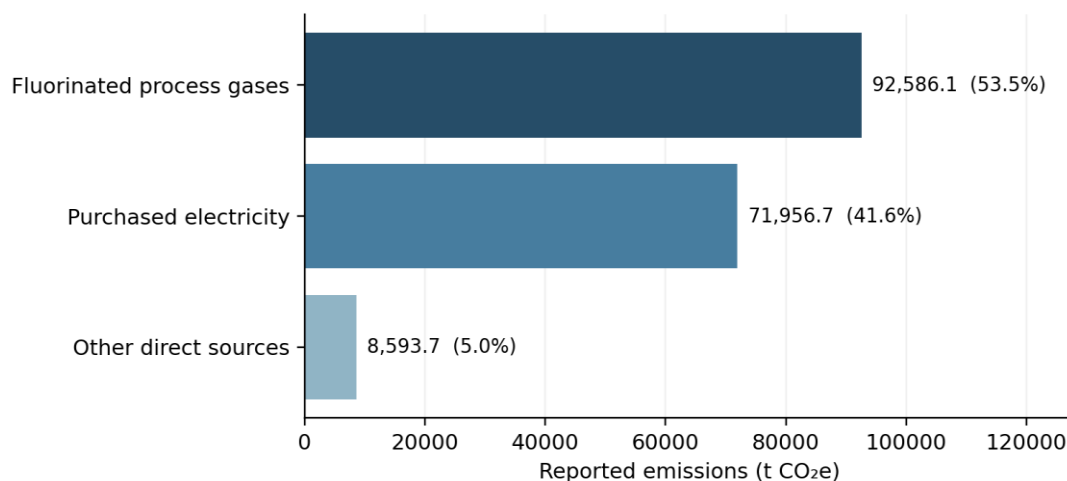


Figure 1. Composition of the recorded scope 1 and scope 2 summary. Percentages use the combined total of 173,136.469 t CO₂e; rounding causes the displayed percentages to sum to 100.1%.

Arithmetic reproducibility did not resolve the reporting period. Activity and detailed calculations carry 2020 labels, whereas the summary carries 2021. Published factor sources contain more than one match for 0.4085. An NEA notice using EMA data assigns it to the 2019 average operating margin [7], and NEA used that factor in its FY2020 reporting [8]. A later GreenGov.SG report lists 0.4085 for 2021 and 0.4074 for 2020 [9], while earlier NEA reporting cited 0.4080 for 2020 [10]. Consequently, the numerical match alone cannot identify the factor release or activity year used in this workbook. This result separates a reproducible multiplication from an unambiguous temporal interpretation.

3.2. Aggregation Differences with a Direct Numerical Consequence

The activity-view scope 1 subtotal was 100,027.647 t CO₂e, compared with 101,179.791 t CO₂e in the summary. Their difference of 1,152.144 t CO₂e equals two source contributions: 1,129.174 t from LPG and 22.970 t from the septic-system entry. The discrepancy is 1.139% of summary scope 1 and 0.665% of the combined scope 1 and scope 2 total.

Formula tracing showed that the activity subtotal explicitly added selected source terms and the fluorinated-gas result, excluding these two contributions. The summary reached the non-fluorinated total through column aggregations that included them. The two values therefore represent different source selections. Forced recalculation preserved the discrepancy, providing no support for a stale numeric cache as its explanation on the tested software path. Whether the source selection was intentional cannot be determined from the workbook, but the calculation mechanism is identifiable. An apparently general cross-view anomaly was thus reduced to a specific aggregation relationship.

3.3. Semantic Mismatch and the Direction of Calculation

The C₄F₈ summary field displayed a GWP100 of 12,200 through a reference to a detailed C₂F₆ calculation block. The

corresponding detailed C₄F₈ block used 10,300. A neighbouring inconsistent text label made gas identity ambiguous at first inspection, but the gas-type field and adjacent calculation rows identified the referenced block as C₂F₆. The published comparison confirms that both values occur in AR4 for different gases (Table 2), supporting a gas-factor mismatch rather than an explanation based only on assessment-report versions [11–14].

Table 2. Published GWP100 values for the two gas identities [11–14]

Gas	AR4	AR5	AR6
c-C ₄ F ₈	10,300	9,540	10,200
C ₂ F ₆	12,200	11,100	12,400

The displayed mass was 90.581402 kg. Holding that mass fixed gives an emissions difference of 172.105 t CO₂e between factors of 12,200 and 10,300, equivalent to 0.1701% of summary scope 1. However, the formula chain ran in the opposite direction from this assumed mass-to-emissions relationship. The emissions result was imported from detailed process calculations, and the displayed quantity was calculated by dividing that result by the summary GWP.

Changing only the reference to the value 10,300 increased the derived quantity to 107.290592 kg. The category emissions, scope 1 and combined total were unchanged (Table 3). The fixed-mass sensitivity therefore describes a hypothetical relationship rather than an emissions correction. In addition, the upstream process calculation contains parent-gas and by-product terms. Dividing the combined CO₂e by one parent-gas GWP does not establish the physical mass of that gas. The anomaly concerns both the reference identity and the interpretation of the derived quantity.

Table 3. Controlled factor perturbation and output response

Variable	Original	Perturbed
Displayed GWP100	12,200	10,300
Derived quantity (kg)	90.581402	107.290592
Category emissions (t CO ₂ e)	1,105.093100	1,105.093100
Scope 1 (t CO ₂ e)	101,179.790727	101,179.790727
Combined total (t CO ₂ e)	173,136.469327	173,136.469327

3.4. Discontinuity in Auxiliary Gas Conversion

The implied CO multiplier was 480 from January to September and 30 from October to December, a sixteen-fold change. The monthly sheet contained no native formula records, and the decoded formula search found no explicit links from this sheet into the emissions calculations. The discontinuity is therefore observable in the stored relationship, but its generating procedure and any effect on reported emissions remain unidentified.

An annotation of 8.0 kg does not provide a complete explanation. Dividing the two multipliers by eight gives 60 and 3.75. The first could represent an integer cylinder count; the second requires fractional consumption or a different aggregation convention if interpreted in that way. Pressure basis, cylinder movements and conversion records would be needed to discriminate among these explanations. The case illustrates how a temporal pattern can be detected without establishing a physical fault or a defensible corrected series.

4. Discussion

4.1. From Detecting Disagreement to Explaining Consequences

The main methodological result is that similar-looking inconsistencies have different implications once calculation direction is considered. The aggregation discrepancy changes the source population represented by a subtotal. The GWP mismatch changes a derived quantity while leaving the selected emissions outputs constant. The auxiliary gas discontinuity is present in the stored records but lacks an identifiable generating formula. These outcomes support an explanation procedure that links a relationship test to its computational mechanism before assigning environmental significance.

This distinction complements established concerns about spreadsheet errors [1]. In an environmental dataset, a formula can be numerically executable while the associated field has an unclear physical meaning. The C₄F₈ example demonstrates this particularly clearly: software recalculation preserves the emissions result, yet the reverse-calculated mass remains unsuitable for interpretation as an independent measurement. Numerical reproducibility and semantic validity are therefore separate properties. Neither one can be inferred solely from agreement between totals.

The controlled perturbation adds evidence beyond a static comparison of coefficients. It tests whether the hypothesised causal path exists in the calculation. Here, a fixed-mass calculation produced a non-zero sensitivity while the actual reported-emissions response was zero. The comparison does not demonstrate that all upstream emissions calculations are correct. It shows that this particular coefficient replacement cannot justify the inferred total adjustment. This is a useful methodological safeguard when domain knowledge about factors is applied to inherited analytical files.

4.2. Implications for Industrial Monitoring and Explainable Analytics

Energy monitoring combines physical measurement with data processing. A method that interprets a reporting inconsistency as equipment behaviour can misdirect subsequent analysis. In this case, the cross-view difference arose from aggregation, and the sixteen-fold auxiliary conversion change could not be assigned to a physical event. Preserving variable origin, units, time coverage and calculation links would allow monitoring systems to distinguish changes in the plant from changes in the data-processing chain.

The explanation generated by the proposed framework is inspectable because it identifies the relationship being tested and follows the relevant transformations. This form of interpretability is appropriate for deterministic data processing. It does not require a learned model or a post-hoc feature attribution. Its scope is narrower than operational anomaly detection, but it can provide a foundation for it: a model trained on variables with unresolved meaning may reproduce those inconsistencies rather than reveal industrial behaviour.

The case also shows why data quality should not be reduced to one compensatory score. Electricity arithmetic was reproducible while temporal applicability remained uncertain. A summary total reconciled internally while a parallel subtotal used a different source population. Recording these properties separately retains information that an average score could conceal. For downstream analyses, the relevant question becomes whether a specific variable is suitable for a specific use, such as annual inventory comparison, physical gas consumption or load forecasting.

The Singapore context makes the distinction between reporting boundaries and physical monitoring practically relevant. NEA's measurement and reporting framework defines facility-level requirements [15], whereas corporate inventory totals follow their stated accounting boundaries [2, 3]. The dataset alone does not establish equivalence between those populations. A monitoring architecture should retain the mapping between sources and intended outputs, rather than assuming that every aggregate describes the same system.

4.3. Transferability and Future Evaluation

The procedure is most directly transferable to datasets containing repeated quantities, explicit transformation rules and multiple aggregation views. Cross-view residuals, gas-factor identity checks and local perturbations can be implemented without labelled operational events. Their usefulness will depend on the availability and stability of those relationships. Values-only exports offer less explanatory depth because the generating transformations may have been lost.

A multi-facility study could evaluate the method by comparing its explanations with independently reconstructed calculation histories and domain-expert classifications. Such evaluation should distinguish detection coverage from correctness of the inferred mechanism, and distinguish both from the significance of any downstream output change. The present purposive case provides examples of those distinctions but no performance estimates.

For subsequent time-series work, interval electricity data, production load and maintenance records would support transparent baselines using operating variables and, where

appropriate, seasonal or regression residuals. Energy-anomaly research provides the broader modelling context [6]. Any later energy-saving claim would additionally require a baseline, measurement boundary and adjustment variables consistent with measurement-and-verification practice [16, 17]. The current findings motivate these data requirements without selecting or validating a predictive model.

4.4. Limitations

The study is exploratory and based on one dataset selected through availability. The sixteen checks were consolidated retrospectively; eight passed, five failed, two remained unresolved and one sensor-assessment check was inapplicable. These outcomes describe the analysis performed, not the prevalence of errors across the workbook or the industry. The computational test covered originally numeric cells and focal dependency paths; it did not establish equivalence for macros, all external dependencies, graphical objects or error-valued cells.

Physical accuracy could not be established without independent source measurements. The reporting period, source-boundary intent and original conversion procedure remain unresolved. Removing company identifiers also does not eliminate re-identification risk when exact values and cell coordinates are retained. The underlying workbook is not distributed; public release of extracted records requires data-owner permission. These limits constrain inference to the recorded relationships and their computational consequences.

5. Conclusions

This study presented a rule-based data quality framework for explaining inconsistencies in industrial energy and emissions records. Applied to a semiconductor dataset, it distinguished a source-aggregation difference of 1,152.144 t CO₂e, a gas-factor mismatch affecting a reverse-calculated quantity, a sixteen-fold auxiliary conversion discontinuity and an unresolved temporal interpretation of the electricity factor. Formula tracing and controlled perturbation showed that the consequence of an inconsistency depends on its position and direction in the calculation chain.

The central contribution is a reproducible path from relationship-based detection to an explanation of computational mechanism and output effect. The case demonstrates why numerical agreement, semantic alignment and physical validity should be evaluated separately before records are reused for energy-performance modelling. Further cases with independent source data are needed to determine how consistently the approach identifies and explains industrial data-quality problems.

6. Declarations

Data availability. Selected de-identified numerical records, formula extracts and the complete retrospective check register are provided in the accompanying supplementary material for author review. Public release is subject to the data owner's permission. The original workbook is not redistributed.

AI assistance. OpenAI ChatGPT assisted with manuscript drafting and revision, literature retrieval and analytical code preparation. The computational results were produced from the supplied workbook.

Acknowledgment

This research was supported by Enerstay Sustainability

Academy (Singapore) Grant Call (Call 7/2025) _SDG (Project ID CUMT-001), Singapore.

References

- [1] Panko, R. R. (1998). What we know about spreadsheet errors. *Journal of End User Computing*, 10(2), 15–21. <https://doi.org/10.4018/joeuc.1998040102>
- [2] World Resources Institute, & World Business Council for Sustainable Development. (2004). The greenhouse gas protocol: A corporate accounting and reporting standard (Revised edition). <https://ghgprotocol.org/corporate-standard>
- [3] International Organization for Standardization. (2018). Greenhouse gases—Part 1: Specification with guidance at the organization level for quantification and reporting of greenhouse gas emissions and removals (ISO 14064-1:2018). <https://www.iso.org/standard/66453.html>
- [4] Intergovernmental Panel on Climate Change. (2019). 2019 refinement to the 2006 IPCC guidelines for national greenhouse gas inventories: Volume 3, Chapter 6: Electronics industry emissions. https://www.ipcc-nggip.iges.or.jp/public/2019rf/pdf/3_Volume3/19R_V3_Ch06_Electronics.pdf
- [5] World Resources Institute. (2015). GHG protocol scope 2 guidance. <https://ghgprotocol.org/scope-2-guidance>
- [6] Himeur, Y., Ghanem, K., Alsalemi, A., Bensaali, F., & Amira, A. (2021). Artificial intelligence based anomaly detection of energy consumption in buildings: A review, current trends and new perspectives. *Applied Energy*, 287, 116601. <https://doi.org/10.1016/j.apenergy.2021.116601>
- [7] National Environment Agency. (2020, December 11). Information on emission factors (For CDM projects in Singapore). <https://www.nea.gov.sg/docs/default-source/envision/clean-development-mechanism/information-on-emission-factors-dec2020.pdf>
- [8] National Environment Agency. (2021). Integrated sustainability report 2020/2021: Walking the talk. <https://www.nea.gov.sg/committee-of-supply-2022/integrated-sustainability-report-2020-2021/review-of-fy2020/enabled-by-a-high-performance-and-future-ready-nea/walking-the-talk>
- [9] Ministry of Sustainability and the Environment, Singapore. (2022). GreenGov.SG report FY2022. Appendix, Table 8: Scope 2 emission factors, p. 59. <https://www.mse.gov.sg/files/resources/greengovsg-report-fy2022.pdf>
- [10] National Environment Agency. (2022). Integrated sustainability report 2021/2022: Walking the talk. <https://www.nea.gov.sg/integrated-sustainability-report-2021-2022/review-of-fy2021/enabled-by-a-high-performance-and-future-ready-nea/walking-the-talk>
- [11] GHG Protocol. (2024, August). Global warming potential values: Cross-report GWP100 compilation. [https://ghgprotocol.org/sites/default/files/2024-08/Global-Warming-Potential-Values%20\(August%202024\).pdf](https://ghgprotocol.org/sites/default/files/2024-08/Global-Warming-Potential-Values%20(August%202024).pdf)
- [12] Forster, P., et al. (2007). Changes in atmospheric constituents and in radiative forcing. In *Climate change 2007: The physical science basis (IPCC AR4 WGI, Chapter 2, Table 2.14)*. Cambridge University Press. <https://www.ipcc.ch/report/ar4/wg1/changes-in-atmospheric-constituents-and-in-radiative-forcing/>
- [13] Myhre, G., et al. (2013). Anthropogenic and natural radiative forcing. In *Climate change 2013: The physical science basis (IPCC AR5 WGI, Chapter 8, Appendix 8.A)*. Cambridge University Press. <https://www.ipcc.ch/report/ar5/wg1/>

- [14] Smith, C., et al. (2021). The earth's energy budget, climate feedbacks, and climate sensitivity: Supplementary material. In Climate change 2021: The physical science basis (IPCC AR6 WGI, Chapter 7, supplementary GWP tables). <https://www.ipcc.ch/report/ar6/wg1/>
- [15] National Environment Agency. (n.d.). Measurement and reporting requirements for greenhouse gas emissions: Includes guidance for RY2019–RY2023 and subsequent periods. <https://www.nea.gov.sg/our-services/climate-change-energy-efficiency/climate-change/carbon-tax/measurement-and-reporting-requirements-for-greenhouse-gas-emissions>
- [16] ASHRAE. (2023). Guideline 14-2023: Measurement of energy, demand, and water savings. <https://www.ashrae.org/technical-resources/standards-and-guidelines/standards-addenda>
- [17] Efficiency Valuation Organization. (2022). International performance measurement and verification protocol: Core concepts 2022. <https://evo-world.org/en/news-media/evo-news/1280-release-of-the-new-ipmvp-core-concepts-2022>