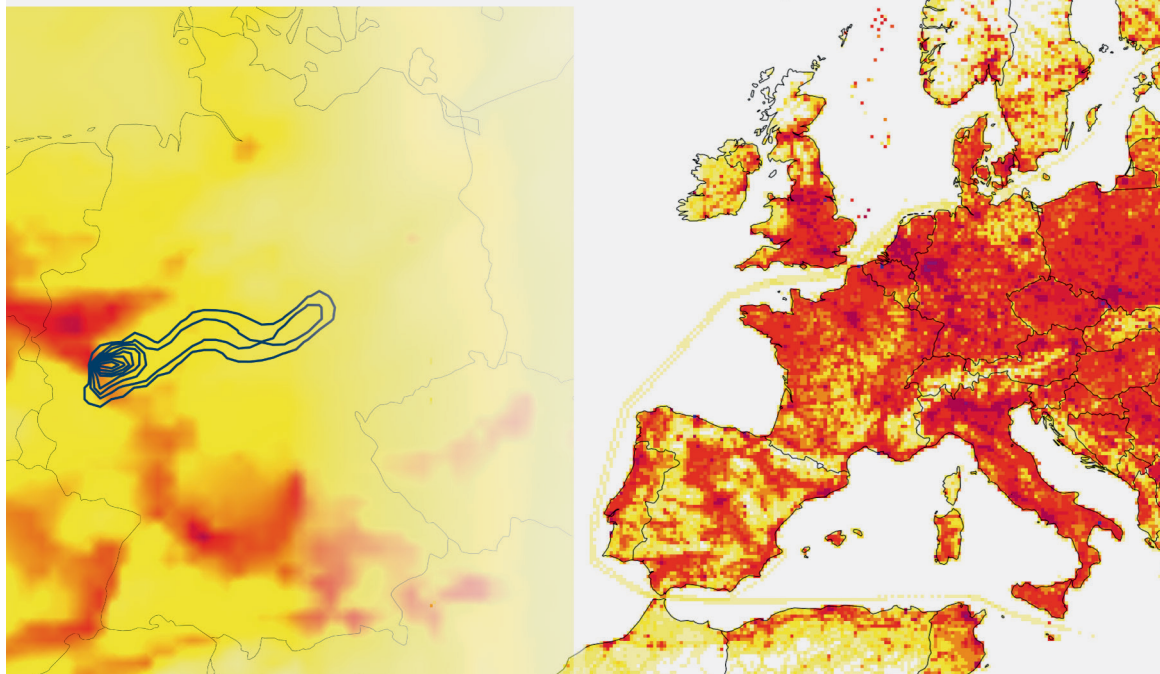


$$O(\chi) = \int \bar{S}(\chi, \xi) \times E^{inv}(\xi) \times BG_{RMSE}(\chi) d\xi$$



Model informed assessment of observation impact for emission rate optimization using four-dimensional variational data assimilation

Alexander Hermanns

Energie & Umwelt / Energy & Environment

Band / Volume 724

ISBN 978-3-95806-959-6

Forschungszentrum Jülich GmbH
Institute of Climate and Energy Systems (ICE)
Troposphäre (ICE-3)

Model informed assessment of observation impact for emission rate optimization using four-dimensional variational data assimilation

Alexander Hermanns

Schriften des Forschungszentrums Jülich
Reihe Energie & Umwelt / Energy & Environment

Band / Volume 724

ISSN 1866-1793

ISBN 978-3-95806-959-6

Bibliografische Information der Deutschen Nationalbibliothek.
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der
Deutschen Nationalbibliografie; detaillierte Bibliografische Daten
sind im Internet über <http://dnb.d-nb.de> abrufbar.

Herausgeber
und Vertrieb: Forschungszentrum Jülich GmbH
Zentralbibliothek, Verlag
52425 Jülich
Tel.: +49 2461 61-5368
Fax: +49 2461 61-6103
zb-publikation@fz-juelich.de
www.fz-juelich.de/zb

Umschlaggestaltung: Grafische Medien, Forschungszentrum Jülich GmbH

Druck: Grafische Medien, Forschungszentrum Jülich GmbH

Copyright: Forschungszentrum Jülich 2026

Schriften des Forschungszentrums Jülich
Reihe Energie & Umwelt / Energy & Environment, Band / Volume 724

D 38 (Diss. Köln, Univ., 2026)

ISSN 1866-1793
ISBN 978-3-95806-958-9 (Print)
ISBN 978-3-95806-959-6 (E-Book)

Vollständig frei verfügbar über das Publikationsportal des Forschungszentrums Jülich (JuSER)
unter www.fz-juelich.de/zb/openaccess.



This is an Open Access publication distributed under the terms of the [Creative Commons Attribution License 4.0](https://creativecommons.org/licenses/by/4.0/),
which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Air pollution poses a significant threat to human health and the environment. Although efforts are being made to develop emission inventories, and characterise and quantify emission rates, the absence of direct measurements introduces uncertainty into model-based estimations, leading to uncertain air pollution predictions. Chemistry transport models, such as the EUROpean Air pollution Dispersion-Inverse Model (EURAD-IM), can combine physical, chemical, and meteorological information to simulate the state of the atmosphere. A common methodology for incorporating observational data into chemistry transport models for further analysis is data assimilation. The EURAD-IM uses four-dimensional variational data assimilation, which allows for the correction of emission inventories. However, the accuracy of these corrections depend on the information provided by the observations. It is therefore a fundamental requirement of the observational data to be representative of the state of the atmosphere.

In order to improve the representativity of the model analysis, a methodology to generate a representative split of ground-based observation data is developed in this work. A clustering algorithm is used to determine observation stations with common statistical features e.g. mean and variance of the annually averaged diurnal cycle. By sampling each cluster individually, the resulting assimilation and validation data sets are representative of the available measurement data. The representativity is evaluated as the difference in the root mean square error of the assimilation data to the validation data after the model analysis. It is shown that the representativity of the new data sets is improved compared to an operational configuration.

To assess the influence of observation configurations on the model result, a series of observing system simulation experiments is performed. Hereby, artificial observations are extracted from a model simulation, the nature run, which serves as ground truth to gauge the accuracy of subsequent model simulations. The nature run is generated with a known emission perturbation for CO (carbon monoxide), NO_x (nitric oxide and nitrogen dioxide), NMVOC (non-methane volatile organic compounds), and SO_x (sulfur oxide) from EURAD-IM simulations. This allows to examine the influence of the observation network on the emission rate correction directly. A systematic benchmark study addressing the relation between spatial coverage of the observational network and the quality of the model analysis is performed. The ability of the model to reproduce the emission rate perturbation of the nature run is evaluated for different observation networks with varying coverage. It is shown that the best performance is achieved by an observation configuration covering approximately 11 % of the model grid. Increasing the observation density further reduces the accuracy of the emission rate corrections.

In addition, a surrogate model for the evaluation of the impact of individual observations is developed, focusing on the definition of an optimal observation network for emission rate

optimization. It combines an adjoint sensitivity field of the simulated day with the emission inventory data and the root-mean square error (RMSE) of the model concentrations of a reference simulation (without data assimilation) with respect to the observation. An algorithm for the generation of an observation configuration based on the observation impact is developed and the resulting observation configuration is tested in the framework established by the benchmark study. The generated observation network outperforms the best benchmarking configurations in terms of RMSE reduction and convergency rate with a lower number of observations. The algorithm is extended to supplement satellite-based column density observations to ground observation networks. It is shown that additional satellite observations to an optimal ground-based observation configuration are detrimental to the emission rate optimization. The emission rate RMSE reductions are decreased by 7 to 16 %. This is due to the column density observations affecting the entire model column in the assimilation, while mainly the ground layers are relevant for the emission rate optimization. Additional satellite observations to an operational configuration show an improvement in the emission rate RMSE reductions (17 to 29 %). This is attributed to the improved capturing of emission sources not adequately captured by the operational configuration. The studies in this thesis enhances the understanding of the impact of observation configurations for the correction of emission rates, which has not been presented in the literature to date. They lay the foundation for future investigations on the optimal combination of different measurement instruments to enable more accurate emission rate assessments and improved air quality predictions.

Kurzzusammenfassung

Luftverschmutzung stellt eine signifikante Bedrohung für die menschliche Gesundheit und die Umwelt dar. Während Anstrengungen unternommen werden, um Emissionsinventare zu entwickeln und Emissionsraten zu charakterisieren und zu quantifizieren, führt die Abwesenheit direkter Messungen zu Unsicherheiten in modellbasierten Emissionsabschätzungen, welche wiederum zu unsicheren Vorhersagen der Luftverschmutzung führen. Chemie-Transportmodelle, zum Beispiel das EUROpean Air pollution Dispersion-Inverse Model (EURAD-IM), sind in der Lage physikalische, chemische und meteorologische Informationen zu kombinieren, um den Zustand der Atmosphäre zu simulieren. Eine häufig genutzte Methode zur Einbindung von Beobachtungsdaten in Chemie-Transportmodelle für weitere Analysen ist die Datenassimilation. Das EURAD-IM nutzt vier-dimensionale variationelle Datenassimilation, welche unter anderem die Korrektur von Emissionsraten erlaubt. Die Genauigkeit dieser Korrekturen hängt jedoch von den Informationen ab, die durch die Beobachtungen bereitgestellt werden. Es ist daher ein grundsätzlicher Anspruch an die Beobachtungsdaten, dass sie repräsentativ für den Zustand der Atmosphäre sind.

Um die Repräsentativität der Modellanalyse zu verbessern, wird in dieser Arbeit eine Methodik zur Erzeugung einer repräsentativen Aufteilung von boden-gestützten in situ Beobachtungen entwickelt. Ein Clustering-Algorithmus wird genutzt um Beobachtungsstationen mit gemeinsamen statischen Merkmalen zu identifizieren, z.B. Mittelwert und Varianz des jährlich gemittelten Tagesgangs. Durch die individuelle Aufteilung jedes Clusters sind die resultierenden Assimilations- und Validierungsdatensätze repräsentativ für die verfügbaren Messdaten. Die Repräsentativität wird durch die Differenz der Wurzel des quadratischen Mittelwertfehlers der Assimilationsdaten zu den Validierungsdaten nach der Modellanalyse bewertet. Es zeigt sich, dass die Repräsentativität der neuen Datensätze im Vergleich zu einer operativen Konfiguration verbessert ist.

Um den Einfluss der Beobachtungskonfigurationen auf das Modellergebnis abzuschätzen, wird eine Reihe von Experimenten mit simulierten Beobachtungssystemen durchgeführt. Dabei werden künstliche Beobachtungen aus dem sogenannten 'nature run' extrahiert, dieser dient als Vergleichsgrundlage, um die Genauigkeit weiterer Modellsimulationen zu bewerten. Der nature run wird mit einer bekannten Störung der CO (Kohlenstoffmonoxid), NO_x (Stickoxide), NMVOC (flüchtige organische Verbindungen außer Methan), und SO_x (Schwefeloxide) Emissionen aus EURAD-IM Simulationen generiert. Dadurch lässt sich der Einfluss des Beobachtungsnetzes auf die Korrektur der Emissionsraten durch das Assimilationssystem direkt evaluieren. Es wird eine systematische Benchmark-Studie durchgeführt, die sich mit dem Zusammenhang zwischen der räumlichen Abdeckung des Beobachtungsnetzes und der Qualität der Modellanalyse befasst. Die Fähigkeit des Modells, die Störungen der Emissionsraten des nature runs zu reproduzieren, wird für verschiedene Beobachtungsnetze mit variierenden Abdeckungen evaluiert. Es zeigt sich, dass die bes-

te Leistung mit einer Beobachtungskonfiguration erreicht wird, die näherungsweise 11 % des Modellgitters abdeckt. Eine weitere Erhöhung der Beobachtungsdichte reduziert die Genauigkeit der Emissionsratenkorrektur.

Darüber hinaus wird ein Surrogatmodell zur Bewertung der Auswirkung einzelner Beobachtungen entwickelt, wobei der Schwerpunkt auf der Definition eines optimalen Beobachtungsnetzes zur Optimierung der Emissionsraten liegt. Es kombiniert ein adjungiertes Sensitivitätsfeld des simulierten Tages mit den Emissionsinventardaten und der Wurzel des quadratischen Mittelwertfehlers (RMSE) der Modellkonzentrationen aus einem Referenzlauf (ohne Datenassimilation) in Bezug auf die Beobachtung. Es wird ein Algorithmus zur Generierung einer Beobachtungskonfiguration auf der Grundlage der Auswirkung einzelner Beobachtungen entwickelt, und die resultierende Beobachtungskonfiguration wird im Rahmen der Benchmark-Studie getestet. Die generierte Beobachtungskonfiguration übertrifft die Güte der besten Benchmark-Konfiguration hinsichtlich der RMSE Reduktion und der Konvergenzrate mit einer geringeren Anzahl an Beobachtungen. Der Algorithmus wird erweitert, um satellitengestützte Beobachtungen von Säulendichten zu Bodenbeobachtungsnetzen zu ergänzen. Es zeigt sich, dass zusätzliche Satellitenbeobachtungen zu einer optimalen bodengestützten Beobachtungskonfiguration sich nachteilig auf die Optimierung der Emissionsraten auswirken. Die Reduktion des RMSE der Emissionsraten wird um 7 bis 16 % vermindert. Das liegt daran, dass die Beobachtungen der Säulendichten die gesamte Modellsäule in der Assimilation beeinflussen, während für die Optimierung der Emissionsraten hauptsächlich die Bodenschichten relevant sind. Zusätzliche Satellitenbeobachtungen zu einer operativ genutzten Konfiguration zeigen eine Verbesserung der Reduktion des RMSE der Emissionsraten (17 bis 29 %). Dies ist auf die verbesserte Erfassung von Emissionsquellen zurückzuführen, welche von der operativ genutzten Konfiguration nicht hinreichend erfasst werden. Die Untersuchungen in dieser Arbeit verbessern das Verständnis des Einflusses von Beobachtungskonfigurationen auf die Korrektur von Emissionsraten, was bisher in der Literatur noch nicht behandelt wurde. Sie legen den Grundstein für zukünftige Studien zur optimalen Kombination verschiedener Messinstrumente, um genauere Emissionsratenbewertungen und verbesserte Luftqualitätsvorhersagen zu ermöglichen.

Contents

Abstract	i
Kurzzusammenfassung	iii
List of Figures	vii
List of Tables	xiii
Acronyms	xv
1. Introduction	1
2. Methodologies	7
2.1. Data assimilation / 4D-var	7
2.2. The atmospheric chemistry transport model EURAD-IM	12
2.2.1. Forward model	12
2.2.2. Data assimilation scheme	14
2.3. Theory of observing system simulation experiments	16
3. Data clustering to optimise the representativity of observational data in air quality data assimilation: a case study with EURAD-IM (version 5.9.1 DA)	20
4. Benchmark experiment	38
4.1. OSSE setup	38
4.2. Assessment of the nature run	46
4.3. Analysis of the benchmark OSSEs	49
4.3.1. Emission rate correction	49
4.3.2. Trace gas concentration	53
4.4. Conclusion of the benchmark experiment	57
5. Improving observational impact	59
5.1. Surrogate model	59
5.2. Algorithm for the generation of targeted observation configurations	62

5.3. Analysis of a targeted ground observation network	64
5.3.1. Analysis of the OSSE result	65
5.4. Application to satellite data	70
5.4.1. Setup of the satellite OSSE	70
5.4.2. Analysis of the satellite OSSE	72
6. Summary and conclusions	78
General acknowledgments	82
A. Appendix: Supplement material to the benchmark experiments	83
A.1. Influence of Gaussian observation error on the assimilation run	83
A.2. Comparison of the emission rate RMSE for grid parameter 10	85
A.3. Regional distribution of the emission rate correction errors	86
B. Appendix: Supplement material to the observation impact experiment	91
B.1. Comparison of the concentration RMSE evolution	92
B.2. Configurations of the observation networks in Sect. 5.4	93
B.3. Initial value gradient and concentration difference for 17 March 2016	97
Bibliography	99
Personal acknowledgments	110

List of Figures

2.1.	Simplified schematic of the EURAD-IM forward model. The inputs are shown in green, the outputs in red and the forward model itself as a blue box containing various processes.	13
2.2.	Flowchart of the 4D-var assimilation process implemented in the EURAD-IM. The input data is shown in green, the assimilation cycle itself in blue, and the output in red. $\mathbf{x}_0^b$ denotes the background model state at initial time, $\mathbf{x}_0$ the initial state for each iteration, and $\mathbf{u}$ the transformed emission factor.	15
2.3.	Schematic of the OSSE framework, adapted from Timmermans et al. (2015). ‘DA’ is short for data assimilation.	17
4.1.	Plot of the emission inventory used in the benchmark OSSE. The emissions for CO are displayed in the top left, NO _x top right, SO _x bottom left and NMVOC bottom right. The values are shown as average daily emission over the year 2016 in t day ⁻¹ . The pixels correspond to the used model grid with a 15 km × 15 km horizontal resolution. The inventory data are taken from CAMS-REG-AP (version 3.1) inventory (Kuenen et al., 2022) and supplemented with the reported emission values from Umweltbundesamt Deutschland (2019) for Germany.	40
4.2.	Configuration of the six regular grid observation networks. The graphics for grid parameters 10, 7, and 5 show the whole model domain. For grid parameters 3, 2, and 1, only central Europe is shown to improve readability. Each dot represents a ground-based observation site of hourly concentration measures for CO, NO ₂ , O ₃ and SO ₂	42
4.3.	Configuration of the irregular grid observation networks for the four observed species CO, NO ₂ , O ₃ and SO ₂ . The grids for the correlation based networks, NC15 and NC50, are the grids for the first day of the assimilation experiment, 10 March 2016. Each dot represents an observation location, where hourly observations are extracted from the NR analogous to operational available observations.	45

4.4. Comparison of the spatially averaged time series of hourly concentrations of the nature run (blue), control run (green), and of real observations (orange) at ground measurement sites for the experiment horizon from 10 to 24 March 2016. Each plot shows the time series of an observed species CO (top left), NO ₂ (top right), O ₃ (bottom left), and SO ₂ (bottom right). The x-axis is marked at 12 UTC of the indicated day.	46
4.5. Evolution of the relative emission rate RMSE reductions in [%] of the maximum RMSE for the regular observation grid configurations from 9 to 24 March 2016. The values for the emission rates of CO (top left), NO _x (top right), SO _x (bottom left), and NMVOC (bottom right) are shown. The y-axis is limited to values less than 30 % for NMVOC and 80 % for the other species.	50
4.6. Emission rate RMSE reductions in [%] of the maximum RMSE for the benchmarking configurations at 17 March 2016 depending on the observation density for CO (blue), NO _x (orange), SO _x (green), and NMVOC (pink) according to Tab. 4.2 and Tab. 4.3. The observation densities of NO ₂ and SO ₂ are shown for NO _x and SO _x , respectively. Since no observations for NMVOC are used, the average observation density is shown. The RMSE of the regular grid configurations are indicated by dots, those for the correlation based configurations NC15 and NC50 by "x" and "+"-markers, respectively, and those for the KSC configuration by triangles. The observation density for the grid parameter 10 configuration is restricted to observations over land.	52
4.7. Evolution of the concentration RMSE reductions of CO (top left), NO ₂ (top right), SO ₂ (center left), O ₃ (center right), and NMVOC (bottom left) relative to the maximum RMSE for the regular grid benchmark configurations from 9 to 24 March 2016. The y-axes are limited to values less than 80 %, except for NMVOC, where the upper limit is decreased to 25 %. For O ₃ , the lower limit is extended to -20 %.	55
4.8. Reductions of the concentration RMSEs relative to the maximum RMSEs for the benchmarking configurations at 17 March 2016 depending on the observation density for CO (blue), NO ₂ (orange), SO ₂ (green), NMVOC (pink), and O ₃ (grey). The RMSE values of the regular grid configurations are indicated by dots, those for the correlation based configurations NC15 and NC50 by "x" and "+"-markers, respectively, and those for the KSC configuration by triangles. The observation density for the grid parameter 10 configuration is restricted to observations over land. For NO _x and SO _x the observation densities of NO ₂ and SO ₂ respectively are shown. Since no observations for NMVOC are used, the average observation density is shown.	57

5.1. Example of a normalized sensitivity field for an observation location in Germany for 10 March 2016. The wind barbs show the mean wind direction and speed at the ground level, where each half flag depicts 5 kn and each full flag 10 kn. A circle indicates winds below 5 kn.	61
5.2. Observation configurations for CO, NO ₂ , O ₃ , and SO ₂ generated using the observation impact algorithm on 10 March 2016. Each dot represents a time series of hourly ground-based concentration values.	65
5.3. Evolution of the emission rate RMSE reductions in [%] of the maximum RMSE for the observation configuration derived using the observation impact algorithm (Eq. 5.6) from 9 to 17 March 2016. The values for the emission rates of CO (top left), NO _x (top right), SO _x (bottom left), and NMVOC (bottom right) are shown. The y-axis is limited to values less than 30 % for NMVOC, and 80 % for other species. The evolution of the regular grid configurations is shown as dashed lines for comparison. The observation impact configuration results are depicted as solid black line.	66
5.4. Emission rate RMSE reductions in [%] relative to the maximum RMSE at 17 March 2016 for CO (blue), NO _x (orange), SO _x (green), and NMVOC (pink). For NO _x and SO _x the observation densities of NO ₂ and SO ₂ are shown, respectively. Since no observations for NMVOC are used, the average observation density is indicated. The RMSE is shown for the regular grid configurations by dots, by "x"-markers for the correlation based configuration NC15, by "+"-markers for configuration NC50, by triangles for the KSC configuration, and by diamonds for the observation impact configuration. The observation density for the grid parameter 10 configuration is restricted to observations over land. This figure is identical to Fig. 4.6, except for the addition of the results for the observation impact configuration.	68

5.5. Reductions of the concentration RMSE in [%] relative to the maximum RMSE for the observation impact assimilation run, compared to the regular and irregular grid benchmark configurations at 17 March 2016. The RMSEs are shown for the species CO (blue), NO ₂ (orange), SO ₂ (green), O ₃ (grey), and NMVOC (pink) in dependence to the observation densities. For NO _x and SO _x the observation densities of NO ₂ and SO ₂ are shown, respectively. Since no observations for NMVOC are used, the average observation density is shown. The RMSEs of the regular grid configurations are indicated by dots, by "x"-markers for the correlation based configuration NC15, by "+"-markers for configuration NC50, by triangles for the KSC configuration, and the RMSEs of the observation impact configuration by diamonds. The observation density for the grid parameter 10 configuration is restricted to observations over land. Note that the values for CO and SO ₂ of the observation impact assimilation run overlap.	69
5.6. Configuration of the CO observation networks considered in this section for 10 March 2016. Each dot represents an observation site with the ground observations in blue and satellite observations in red.	72
5.7. Concentration (bars) and emission (circles) RMSE reductions in [%] of the maximum RMSE for the supplemented and base configurations on 17 March 2016 for CO, NO ₂ (NO _x emissions), SO ₂ (SO _x emissions), NMVOC, and O ₃ . For O ₃ , no emission rates are available as it is a secondary pollutant. The y-axis is limited to values less than 80 %.	73
5.8. Normalized initial value gradient averaged along model y-direction for the Sat (left) and the OI+grnd assimilation run (center) and the normalized and along y-direction averaged concentrations difference between the NR and CR (right). The values are shown for 10 March 2016 and for the species CO, NO ₂ , SO ₂ , and O ₃ in descending order. The y-axis shows the height above the ground in [km] and the x-axis the model x-coordinate.	77
A1. Evolution of the emission rate RMSE reductions in [%] of the maximum RMSE for the regular observation grid configuration with grid parameter 10 during the assimilation window from 10 to 17 March 2016. The solid lines shows the result using unperturbed observations extracted from the nature run. The dashed lines show the result for the analysis using observations with additional Gaussian perturbation to reflect observation errors. Results are shown for CO (blue), NO _x (orange), SO _x (green), and NMVOC (pink). The RMSE fraction for 09 March 2016 is shown for completeness and the y-axis is limited to values less than 80 %.	83

A2.	Logarithmic plot of the CO emission rate correction errors of the grid parameter 10 assimilation run for 19 and 24 March 2016. Blue signifies an underestimation of the nature run emission rates while red indicates an overestimation.	85
A3.	Logarithmic plot of the emission rate correction errors of CO for 17 March 2016. Shown are the results for the KSC (top left), grid parameter 10 (top right), grid parameter 3 (bottom left), and grid parameter 1 (bottom right) configuration.	87
A4.	Logarithmic plot of the emission rate correction errors of NO _x for 17 March 2016. Shown are the results for the KSC (top left), grid parameter 10 (top right), grid parameter 3 (bottom left), and grid parameter 1 (bottom right) configuration.	88
A5.	Logarithmic plot of the emission rate correction errors of SO _x for 17 March 2016. Shown are the results for the KSC (top left), grid parameter 10 (top right), grid parameter 3 (bottom left), and grid parameter 1 (bottom right) configuration.	89
A6.	Logarithmic plot of the emission rate correction errors of NMVOC for 17 March 2016. Shown are the results for the KSC (top left), grid parameter 10 (top right), grid parameter 3 (bottom left), and grid parameter 1 (bottom right) configuration.	90
B1.	Evolution of the concentration RMSE in [%] relative to the maximum from 9 to 17 March 2016 for the observation impact (solid black line) and the regular grid benchmark assimilation runs (dashed lines). The figure shows the values for CO (top left), NO ₂ (top right), SO ₂ (center left), O ₃ (center right), and NMVOC (bottom left). The y-axes are limited to values between 0 % and 80 %, except for O ₃ , where the lower limit is extended to -20 % and NMVOC, where the upper limit is set to 25 %.	92
B2.	Configuration of the NO ₂ observation networks considered in Sect. 5.4 for 10 March 2016. Each dot represents an observation site with the ground observations in blue and satellite observations in red.	93
B3.	Configuration of the SO ₂ observation networks considered in Sect. 5.4 for 10 March 2016. Each dot represents an observation site with the ground observations in blue and satellite observations in red.	94
B4.	Configuration of the O ₃ observation networks considered in Sect. 5.4 for 10 March 2016. Each dot represents an observation site with the ground observations in blue and satellite observations in red.	95

B5. Normalized initial value gradient averaged along model y-direction for the Sat (left) and the OI+grnd assimilation run (center) and the normalized and along y-direction averaged concentrations difference between the NR and CR (right). The values are shown for 17 March 2016 and for the species CO, NO₂, SO₂, and O₃ in descending order. The y-axis shows the height above the ground in [km] and the x-axis the model x-coordinate. 97

List of Tables

4.1. Summary of the OSSE components and their realization in the benchmark OSSE. The model resolution is 15 km × 15 km horizontally for all simulations.	39
4.2. Number of observations and the observation density over the land-surface depending on the grid parameter. Each observation constitutes 24 h time series for CO, NO ₂ , O ₃ and SO ₂ . For grid parameter 10, the number of observations in parentheses is the total number of observations and the plain number corresponds to the observations over land.	43
4.3. Averaged number of observation locations and the density of observations over land surface for the irregular grid observation configurations, distinguished by trace gas. The averages are computed over the entire analysis episode from 10 to 24 March 2016. When applicable, the minimum to maximum number of observation are shown in brackets.	44
4.4. Correlation coefficient, bias, and RMSE of the hourly averaged mean concentrations at ground measurement sites of the control run compared with the ground observations and the nature run. Bias and RMSE are shown in µg m ⁻³ and in %. The percentage is calculated with respect to the mean of the ground observation and NR time series, respectively.	47
4.5. Values for the maximum of the emission rate RMSE for the investigated emission species in the OSSE setup.	50
5.1. Description of the algorithm to generate a targeted observation configuration based on the observation impact of individual observations.	63
5.2. Summary of the observation configurations analysed in this section. The KSC configuration has been modified to include the grid based observations over the ocean, which are also present in the OI configuration.	71

Acronyms

3D-var	Three Dimensional-Variational data assimilation
4D-var	Four Dimensional-Variational data assimilation
AR	Assimilation Run
AV-difference	Assimilation Validation-difference
CAMS	Copernicus Atmosphere Monitoring Service
CAMS-REG-AP	Copernicus Atmosphere Monitoring Service-Regional Anthropogenic emission inventory
CD	Clustering Dirunal
C-IFS	Composition-Integrated Forecasting System
CR	Control Run
CTM	Chemistry Transport Model
DA	Data Assimilation
DOY	Day Of Year
ECMWF	European Center for Medium range Weather Forecasting
EDGAR	Emissions Database for Global Atmospheric Research
EEA	European Environmental Agency
EnKF	Ensemble Kalman Filter
ESM	Earth System Modelling
EURAD-IM	EUROpean Air pollution Dispersion-Inverse Model
GEOS-chem	Goddard Earth Observing System atmospheric chemistry model
GMD	Geoscientific Model Development

GRAPES-CUACE

Global/Regional Assimilation and PrEdiction System-China Meteorological Administration Unified Atmospheric Chemistry Environmental Forecasting System

HDS-LEE Helmholtz School for Data Science in Life, Earth and Energy

IFS Integrated Forecasting System

JSC Jülich Supercomputing Centre

JUWELS Jülich Wizard for European Leadership Science

KSC K-mean Soft Constraint

L-BFGS Limited-memory-Broyden-Fletcher-Goldfarb-Shanno algorithm

NMVOC Non-Methane Volatile Organic Compound

NR Nature Run

OSE Observing System Experiment

OSSE Observing System Simulation Experiment

PAM Partitioning Around Medoids

PDF Probability Density Function

PM₁₀ Particulate Matter with an aerodynamic diameter $\leq 10 \mu\text{m}$

PM_{2.5} Particulate Matter with an aerodynamic diameter $\leq 2.5 \mu\text{m}$

RACM-MIM Regional Atmospheric Chemistry Mechanism - Mainz Isoprene Mechanism

RMSE Root Mean Square Error

STEM Sulfur Transport Eulerian Model

WHO World Health Organization

WPS Weather research and forecasting model Preprocessor System

WRF Weather Research and Forecasting model

1. Introduction

Air pollution and the current amount of pollutants in the air have detrimental effects on human health. In particular, the trace gases CO (carbon monoxide), NO₂ (nitrogen dioxide), O₃ (ozone), and SO₂ (sulfur dioxide), and particulate matter PM₁₀ and PM_{2.5} (particulate matter with an aerodynamic diameter $\leq 10\text{ }\mu\text{m}$ and $\leq 2.5\text{ }\mu\text{m}$, respectively) are known to cause health problems with $\sim 99\%$ of the world's population breathing air that exceeds the WHO guideline limits (World Health Organization, 2021). Air pollution causes cardio-vascular diseases, strokes, and is generally deemed to be carcinogenic (World Health Organization, 2013; Lee et al., 2020; Orellano et al., 2021). It is estimated that in 2022, 357,000 premature deaths in the EU-27 member states were caused by air pollution exceeding the WHO guideline limits (European Environment Agency, 2024).

A main driver of air pollution are anthropogenic emissions, especially in densely populated areas such as Central Europe. Thus, understanding them is a main task in the assessment and development of environmental policies (Amann et al., 2013; Franke et al., 2024). To this end, emission inventories are developed, which capture various emission sources and estimate their contributions based on reports and estimations. They provide spatially and temporally resolved data on the emissions and are generated with a bottom-up or top-down methodology. In the former, emissions for a specific sector (e.g. traffic) are calculated by combining information about the sector (e.g. amount of emissions per 100 km) with estimations about associated activities (e.g. traffic counts) and spatial distributions (e.g. registered cars of a certain type). In top-down estimations, inverse modeling is used to derive emissions from observations of trace gas concentrations. A comparison of the methodologies is shown in Elguindi et al. (2020). Emission inventories are compiled by environmental agencies and research institutes, for example by the Umweltbundesamt Deutschland (2019) for Germany, the CAMS-REG-ANT inventory from the CAMS (Copernicus Atmosphere Monitoring Service; Kuenen et al., 2022) for European anthropogenic emissions, and the global emission inventory EDGAR (Emissions Database for Global Atmospheric Research; Crippa et al., 2020) from the European commission, Joint Research centre.

However, emission inventory data is subject to large uncertainties. The composition, quantity, and spatial and temporal resolution of the emissions themselves are not precisely known. Solazzo et al. (2021), for example, estimate the uncertainties of NO₂ emissions of the power sector at $\sim 112\%$ in the EDGAR inventory. These discrepancies arise due to differences in the methodologies used for the estimation of the EDGAR inventory and

the methods used by individual countries for their own reported emission values. Further, not all countries report uncertainty estimates for their emission inventories (Schindlbacher et al., 2021). Hence, methods that combine numerical models with trace gas observations in the atmosphere to reduce uncertainties, analyse, and correct discrepancies in the emission inventory data have been developed.

Chemistry transport models (CTMs) are an important tool for the forecast and the assessment of air quality. These models are used to describe and analyse the transport and chemical interactions of trace gases and aerosols in the atmosphere. By combining meteorological fields with emission inventory data, chemical interactions and aerosol formation, the continuity equations for the trace gases are solved, enabling air quality predictions and sensitivity analysis. For Europe, the Copernicus Atmosphere Monitoring Service (CAMS, <https://atmosphere.copernicus.eu/regional-air-quality-production-systems>; last access: 29 December 2025) provides a regional ensemble of CTMs (CAMS regional), which generates daily forecast for air quality. It consists of 11 numerical models, which are detailed in Colette et al. (2025) and Marécal et al. (2015).

In order to improve the accuracy of the calculated concentration fields, information about the state of the atmosphere, provided by observations, is incorporated into the models. Common methods for this task are data assimilation techniques. These aim to decrease the discrepancies between the calculated and measured concentration fields by the adjustment of the model state or parameters of the model simulation. One prominent example is four-dimensional variational data assimilation (4D-var). It is capable of jointly optimizing the initial concentrations of the trace gases and the emission rates, given by emission inventories. This allows for reanalyses and especially corrections of emission inventories to be performed. Elbern et al. (2007) developed this method and implemented it into the European Air pollution Dispersion-Inverse Model (EURAD-IM) demonstrating that the joint optimization significantly improved the performance for SO₂ compared to only optimizing the initial concentration values. Furthermore, the O₃ forecast was improved by adjustments of the NO_x (nitric oxide (NO) and NO₂) emission rates. Wang et al. (2021) developed a 4D-var system based on the Global/Regional Assimilation and PrEdiction System – China Meteorological Administration Unified Atmospheric Chemistry Environmental Forecasting System (GRAPES–CUACE). They used an emission inventory from 2007 and observation data from 2016. By employing emission optimization only, the model is able to replicate black carbon emission rates in agreement with values from a 2016 emission inventory. Wang et al. (2022) updated the aforementioned model and were able to analyse and quantify regional emission contributions to an O₃ pollution event in Beijing. By assimilating satellite observations of NO₂ column densities, Kurokawa et al. (2009) were able to adjust NO_x emission rates. They showed a discrepancy in the increase of the a priori (19 %) and a posteriori emission rates (49 %) between 1996 to 2002.

The quality and availability of the observations (especially a sufficient spatial coverage)

strongly impact the accuracy of the data assimilation analysis. Furthermore, observations serve as an important tool to validate model analyses. It is common practice to withhold a subset of the available observation data, the validation data set, from the data assimilation. This validation data is then compared to the model output and enables the quantification of the model accuracy on ‘unseen’ data. Thus, the validation data set has to be representative of the modeled atmosphere and capture regional differences (Joly and Peuch, 2012). This validation procedure however, is only able to validate the model results for the concentration fields. For emission rates, no direct observations exist and the concentration validation has to serve as proxy for the accuracy of the emission rate optimization.

A wide range of sensing methods is available for atmospheric observations. Ground-based in situ measurements can provide good temporal resolution and can be placed close to areas of interest, such as emission sources and densely populated areas. However, they are costly, require maintenance, and only provide information about surface concentrations. In recent years, the availability of remote sensing observations using satellites has increased. These are able to provide measurements in areas, which are not covered by ground-based monitoring networks and provide large scale spatial coverage (ESA, 2025a,b). However, they usually provide column integrated measurements, which do not contain information about the vertical concentration distribution. Zhang et al. (2019) assimilated CO, NO₂, and O₃ observations from multiple satellite instruments to estimate CO and NO_x emissions. They showed that the simultaneous assimilation of multiple species generated regional differences in the model results compared to single species assimilation. For example, the adjustment of NO_x emissions by NO₂ assimilation in addition to O₃ assimilation improved the agreement of O₃ model concentrations with the observations. Alvarado et al. (2019) evaluated the effect of adding additional satellite observations to existing ground based observation networks for PM_{2.5} concentration estimation. They showed that the estimations of daily average PM_{2.5} concentrations with assimilation of satellite observations shows large uncertainties. Furthermore, the data coverage is limited due to cloud interference. However, they complement existing observation networks in inland cities of low altitude.

Due to the global coverage of satellite measurements, they are also used for the validation of CTMs. Zhang et al. (2010) used combined O₃ retrievals from two satellites and show that the GEOS-chem (Goddard Earth Observing System atmospheric chemistry model) underestimates tropospheric O₃ in the tropics and overestimates the concentrations in the northern subtropics and southern extra-tropics.

Newer methods of in situ measurements exist to improve model analyses. Erraji et al. (2024) showed that drone observations of vertical profiles have great potential to improve modeled vertical profiles of O₃ and NO. This is due to improved information about the species concentrations in the planetary boundary layer, which is commonly not resolved by other observation types.

With the amount of the available observation data increasing, the consideration of dis-

advantages of assimilating large amounts of data becomes more important. Especially for high resolution satellite data, data storage becomes an operational difficulty. Furthermore, assimilating large amounts of data increases the computational demand of already computationally expensive CTMs. Liu and Rabier (2002) have shown in a theoretical framework that an observation density exists after which additional observations are detrimental to the model result due to spatial correlated errors. Thus, thinning of high density observational data is necessary. Ochotta et al. (2005) developed two methods to reduce the number of observations to a subset which contains the most representative observations. Both lead to high observation densities in areas with higher variance in the original observations.

On the other hand, sufficient information about the atmosphere needs to be assimilated to ensure adequate accuracy of the model analysis. Several studies have examined techniques for the assessment of the impact of observations on model results. Fisher (2003) developed an algorithm for the quantification of the entropy reduction and the degrees of freedom for signal by the assimilation of observations. These measures are proxies for the provided information. The algorithm requires multiple calculations of the gradient of the cost function under incorporation of the observations. It is especially suitable for the intercomparison of the impact of the data provided by different observation instruments. Cardinali (2009) examined the forecast sensitivity to observations in a forecast from the ECMWF (European Centre for Medium-Range Weather Forecasts). This is done by performing an observing system experiments (OSE). Herein, the differences between simulations with and without the investigated observation type is quantified. They were able to quantify the impact of a variety of observation instruments on the forecast with seasonal changes. However, this requires a large amount of independent model simulations to fully assess an observational network. Langland and Baker (2004) employed a tangent linear adjoint model to analyse the impact of observations on forecast errors. The method is restricted to a specific cost function with respect to a chosen error norm. Furthermore, it requires choosing a specific set of observations for the evaluation prior to the analysis. Goris and Elbern (2015) used singular vector analysis to identify the optimal placement of observations based on the sensitivity of model states. They were able to make conclusions about the relative influence of emitted species and determined general characteristics of well placed observations.

Other methods to analyse the impact of observations are OSEs or data-denial experiments. They are most commonly used to gauge the impact of specific observation data sets by comparing model simulation with and without them. Bouttier and Kelly (2001) examined the influence of different observation systems on the data assimilation system of the ECMWF. They were able to show regional and qualitative difference due to the incorporation of various measurement instruments on the forecast, and how different types of observations complement the forecast when used in conjunction.

A special case of these experiments are observation system simulation experiments (OSSEs) where instead of real data, artificial observations are used. This is done by generating a na-

ture run, which serves as the true state of the atmosphere. From it, the artificial observations are extracted, which are subjected to data assimilation with an (ideally) different model. The nature run serves then as comparison for the evaluation of the impact of the chosen observations. OSSEs are commonly employed to estimate the effectiveness of proposed instruments on model forecasts and to aid in instrument design. Liu et al. (2017) estimated the impact of artificial NO_2 column densities and meteorological observations on the estimation of NO_x emissions. They determined the contribution of meteorological uncertainties to NO_2 concentrations at 35 % and showed that the emission estimation degrades when the hourly wind vector RMSE exceeds 1 m s^{-1} . Zhang et al. (2018) investigated the capability of satellite methane measurements to monitor global hydroxyl radical (OH) concentrations using an OSSE framework. The best performance is achieved by jointly assimilating short-wave infrared and thermal infrared measurements. The general setup and application of OSSEs will be described in more detail in the proceedings of this thesis.

While these approaches provide insight into the impact of observations, they require substantial computations or are not suitable to evaluate the impact of individual observations but rather whole observation instruments.

The objectives of the thesis are thus formulated as follows:

- Evaluation of the impact of observation networks on emission rate optimization
- Establishment of a method to generate a representative subset for the validation of CTM model results
- Quantifying the effect of different observation networks on the emission rate optimization
- Establishing a method to gauge the impact of an individual observation on the emission rate optimization prior to model simulations
- Examining the generation of an observation network for emission rate optimization and extension of an existing network

The structure of the thesis is as follows: Chapter 2 describes the methods and framework of the presented investigation. The theory of data assimilation with a focus on 4D-var is discussed first. Then the current version of the EURAD-IM and its components are introduced and an overview on OSSEs is given. Chapter 3 encompasses an article published in the peer-reviewed Geoscientific Model Development journal, Hermanns et al. (2025) as part of this work. It describes a method for sub-sampling ground observations to increase the representativity of the model analysis and the validation. Chapter 4 describes the setup, execution, and the analysis of OSSEs performed to benchmark the EURAD-IM by using different ground observation configurations and analysing their influence on the emission rate corrections and trace gas concentrations. In Chapter 5, a measure to quantify the impact of

a specific observation on the emission rate optimization is proposed. It is used to generate a ground observation network in the OSSE framework from Chapter 4, which is compared to the benchmark OSSEs and then used to supplement observation configurations with satellite observations. Chapter 6 concludes the thesis by discussing the potential and limitations of the presented approaches. An outlook to further studies and improvements is given.

2. Methodologies

This chapter describes the methodologies and the CTM used in this work. First, data assimilation, with a focus on 4D-variational data assimilation, is introduced and compared to the Ensemble Kalman Filter. Then, the EUROpean Air pollution Dispersion-Inverse Model is described and an overview of the forward model and the data assimilation scheme is given. The framework of observing system simulation experiments is introduced and its application reported in the literature and in the present work are discussed.

2.1. Data assimilation / 4D-var

The analysis of the state of the atmosphere is an important element of atmospheric science. A primary objective in air quality predictions is the modelling of atmospheric composition and the comprehension of physical and chemical processes within the atmosphere. In atmospheric models, an initial state of the atmosphere is propagated in time using physical and chemical equations. This propagated state is the *a priori* or *background* of the model prediction. However, this *a priori* diverges from the true atmospheric state due to the presence of uncertainties in the initial state and emissions, the approximations made in the modelling of atmospheric processes, and possible unknown or unrepresented physical or chemical processes. The divergence can be reduced through data assimilation by incorporating information from observational data into the model simulation to improve the model state. The corrected model state is then called the *a posteriori* or *analysis*.

Data assimilation enables the correction of the model state during the simulation and the propagation of information gathered from observations from well observed to less observed areas. Furthermore, it can be used for the analysis of unobserved variables via physical and chemical coupling to observed variables. Several methods of data assimilation exist, which are based on different mathematical and computational approaches. An overview on data assimilation algorithms used in atmospheric science is shown in Bouttier and Courtier (2002) and Lahoz and Schneider (2014). The algorithms with the highest complexities are the Ensemble Kalman Filter (EnKF) and the 4D-variational data assimilation (4D-var). In EnKF, the initial state of the model is varied to generate an ensemble, where each ensemble member is then subjected to the model integration (Houtekamer and Mitchell, 2003). These ensemble members are then compared to observations and the optimal model state is determined. Detailed information can be found in, e.g., Lahoz et al. (2010) and Evensen

(2009). As the evaluations presented in this thesis are undertaken with 4D-var, a more detailed description of this method is provided below.

The state vector $\mathbf{x}_j \in \mathbb{R}^n$ of the model contains the concentration of all variables at all simulated grid points for the time t_j . In general, the evolution of this state vector is described by the relation

$$\mathbf{x}_j = \mathcal{M}_j \mathbf{x}_{j-1}. \quad (2.1)$$

$\mathcal{M}_j$ is the nonlinear model operator performing the integration of the model state from time t_{j-1} to t_j . It is postulated that the model is without errors and the evolution of the model state is deterministic. Due to this constraint, the data assimilation algorithm is also called strong-constraint 4D-var (Evensen et al., 2022). When model errors are included, the algorithm is called weak-constraint 4D-var (Trémolet and Fisher, 2010). This thesis focuses on the strong-constraint 4D-var.

Equation (2.1) is defined on the time interval $t \in [t_0, \dots, t_N]$, which is called the assimilation window. With respect to the initial state of the model $\mathbf{x}_0$ at t_0 , Eq. (2.1) becomes

$$\mathbf{x}_j = \mathcal{M}_j \mathcal{M}_{j-1} \dots \mathcal{M}_1 \mathbf{x}_0. \quad (2.2)$$

By combining all observations, associated with time t_j , into a single observation vector $\mathbf{y}_j \in \mathbb{R}^m$, the relation between the model state and the observation vector can be expressed as

$$\mathbf{y}_j = \mathcal{H}_j \mathbf{x}_j + \epsilon_j. \quad (2.3)$$

Here, $\mathcal{H}_j \in \mathbb{R}^{m \times n}$ is the forward model operator mapping the model state into the phase space of the observations at time t_j . $\epsilon_j \in \mathbb{R}^m$ are the observation errors combining measurement and representation errors. The discrepancies or departures of the model state from the observation can then be expressed as the innovation vector $\mathbf{d}_j = [\mathbf{y}_j - \mathcal{H}_j \mathbf{x}_j]$, which enables to determine and to quantify necessary changes in the initial conditions.

The problem of determining the optimal initial conditions given the model state and the observations can be described in terms of Bayes' Theorem on conditional probabilities

$$P(\mathbf{x}|\mathbf{y}) = \frac{P(\mathbf{y}|\mathbf{x})P(\mathbf{x})}{P(\mathbf{y})}. \quad (2.4)$$

$P(\mathbf{x}|\mathbf{y})$ is the posterior probability density function (PDF) of model state $\mathbf{x}$ conditioned on the observations $\mathbf{y}$. Similarly, $P(\mathbf{y}|\mathbf{x})$ is the conditional PDF of observing $\mathbf{y}$ while $\mathbf{x}$ is given. $P(\mathbf{x})$ and $P(\mathbf{y})$ are the PDFs for the occurrence of model state $\mathbf{x}$ and observations $\mathbf{y}$, respectively. The objective of data assimilation is to maximize the posterior probability. By assuming that all PDFs are Gaussian distributed PDFs and applying the negative natural logarithm,

Eq. (2.4) leads to the cost function $J(\mathbf{x}_0)$ (Lorenc, 1986):

$$\begin{aligned} J(\mathbf{x}_0) &= J_{iv} + J_{obs} \\ &= \frac{1}{2} [\mathbf{x}_0 - \mathbf{x}^b]^T \mathbf{B}^{-1} [\mathbf{x}_0 - \mathbf{x}^b] \\ &\quad + \frac{1}{2} \sum_{j=0}^N [\mathbf{y}_j - \mathcal{G}_j(\mathbf{x}_0)]^T \mathbf{R}^{-1} [\mathbf{y}_j - \mathcal{G}_j(\mathbf{x}_0)]. \end{aligned} \quad (2.5)$$

Here, $\mathbf{x}^b$ denotes the background state of the model, which can be obtained by a model forecast or climatology. $\mathbf{B} \in \mathbb{R}^{n \times n}$ is the background error covariance matrix, and $\mathbf{R} \in \mathbb{R}^{m \times m}$ describes the covariance matrix of the observation errors. $\mathcal{G}_j$ is defined as

$$\mathcal{G}_j := \{\mathcal{H}_j \mathcal{M}_j \mathcal{M}_{j-1} \dots \mathcal{M}_1\}. \quad (2.6)$$

The cost function, Eq. (2.5), consists of two terms: J_{iv} , referring to the costs due to the mismatch of the updated initial value to $\mathbf{x}^b$, and J_{obs} , which describes the costs due to the discrepancy between the model state and the observations, accounting for all observations within the assimilation window. The objective of maximizing the posterior probability distribution, Eq. (2.4), is equivalent to minimizing the cost function, Eq. (2.5).

The minimization of the cost function of 4D-var is computationally demanding in terms of the necessary number of iterations, due to the non-linearity of $\mathcal{M}$ and $\mathcal{H}$. Since, for the minimization, small perturbations of the model state around the initial model state are considered, an incremental formulation can be introduced. Here, the tangent-linear operators of $\mathcal{M}$ and $\mathcal{H}$ are M and H , respectively, and they are defined by

$$\begin{aligned} \mathcal{M}_j(\mathbf{x}_0 + \delta \mathbf{x}_0) &\approx \mathcal{M}_j \mathbf{x}_0 + M_j \delta \mathbf{x}_0, \\ \mathcal{H}_j(\mathbf{x}_j + \delta \mathbf{x}_j) &\approx \mathcal{H}_j \mathbf{x}_j + H_j \delta \mathbf{x}_j. \end{aligned} \quad (2.7)$$

The cost function with respect to the increment $\delta \mathbf{x}_0 = \mathbf{x}_0 - \mathbf{x}^b$ then becomes

$$\begin{aligned} J(\delta \mathbf{x}_0) &= \frac{1}{2} [\delta \mathbf{x}_0]^T \mathbf{B}^{-1} [\delta \mathbf{x}_0] \\ &\quad + \frac{1}{2} \sum_{j=0}^N [\mathbf{d}_j - G_j(\delta \mathbf{x}_0)]^T \mathbf{R}^{-1} [\mathbf{d}_j - G_j(\delta \mathbf{x}_0)]. \end{aligned} \quad (2.8)$$

Here, the innovation vector is modified to be $\mathbf{d}_j = \mathbf{y}_j - \mathcal{G}_j(\mathbf{x}^b)$, describing the deviation of the observation vector to the propagated background state at time t_j . G_j is defined analogous to Eq. (2.6) applying the tangent-linear operators H_j and M_j . This incremental formulation is used to reduce the computational cost of the minimization, and is applicable since small perturbations of the model state are considered. For an in-depth discussion of the incremental formulation, see Courtier et al. (1994).

Elbern et al. (2007) proposed a joint optimization of the initial values and the emission factors, which can be achieved by considering the influence of the emission on the evolution of the model state. To reduce the degrees of freedom, it is assumed that the diurnal profiles of the emission rates are correct, such that an emission factor $f \in \mathbb{R}^p$ scaling the emission rates is considered in the state vector. Since an emission factor of 0.5 results in the same relative change as an emission factor of 2, the emission factors are transformed to the variable

$$\mathbf{u} = \ln(\mathbf{f}) \quad \text{and} \quad \mathbf{u}^b = \ln(\mathbf{f}^b), \quad (2.9)$$

where f^b corresponds to the emission factor of the background. The cost function from Eq. (2.8), extended to include the scaled emission factors, then reads:

$$\begin{aligned} J(\delta \mathbf{x}_0, \delta \mathbf{u}) = & \frac{1}{2} [\delta \mathbf{x}_0]^T \mathbf{B}^{-1} [\delta \mathbf{x}_0] \\ & + \frac{1}{2} \sum_{j=0}^N [\mathbf{d}_j - G_j(\delta \mathbf{x}_0)]^T \mathbf{R}^{-1} [\mathbf{d}_j - G_j(\delta \mathbf{x}_0)] \\ & + \frac{1}{2} (\delta \mathbf{u})^T \mathbf{K}^{-1} (\delta \mathbf{u}). \end{aligned} \quad (2.10)$$

Here, $\mathbf{K} \in \mathbb{R}^{p \times p}$ is the emission error covariance matrix and $\delta \mathbf{u} = \mathbf{u} - \mathbf{u}^b$ represents perturbations in the emissions. The optimized emission rates can be expressed as

$$\mathbf{E}_j = \mathbf{f} \cdot \mathbf{E}_j^b = \exp(\mathbf{u}) \mathbf{E}_j^b, \quad (2.11)$$

where $\mathbf{E}_j^b$ are the background emission rates at time t_j .

In practice, the convergence of the minimization can be slow and inaccurate, since the condition number of the Hessian of the cost function can be very large (Haben et al., 2011). To reduce the condition number of the Hessian, a common approach is the introduction of the state variables

$$\mathbf{v} := \mathbf{B}^{-1/2} \delta \mathbf{x}_0 \quad \text{and} \quad \mathbf{w} := \mathbf{K}^{-1/2} \delta \mathbf{u}, \quad (2.12)$$

where $\mathbf{B} = \mathbf{B}^{1/2} \mathbf{B}^{T/2}$, and $\mathbf{K} = \mathbf{K}^{1/2} \mathbf{K}^{T/2}$. This allows to rewrite the cost function as

$$J(\mathbf{v}, \mathbf{w}) = \frac{1}{2} \mathbf{v}^T \mathbf{v} + \frac{1}{2} \mathbf{w}^T \mathbf{w} + \frac{1}{2} \sum_{j=0}^N [\mathbf{d}_j - G_j(\delta \mathbf{x}_0)]^T \mathbf{R}^{-1} [\mathbf{d}_j - G_j(\delta \mathbf{x}_0)]. \quad (2.13)$$

In order to minimize this cost function, its gradient is computed. It can be expressed using the adjoint model operator (Talagrand, 1997) and, following Elbern et al. (2007), it reads:

$$\nabla J(\mathbf{v}, \mathbf{w}) = \begin{pmatrix} \mathbf{v} \\ \mathbf{w} \end{pmatrix} - \begin{pmatrix} \mathbf{B}^{1/2} & 0 \\ 0 & \mathbf{K}^{1/2} \end{pmatrix} \sum_{j=0}^N G_j^T \mathbf{R}^{-1} [\mathbf{d}_j - G_j(\delta \mathbf{x}_0)]. \quad (2.14)$$

Here, G_j^T is defined with respect to the adjoint of the tangent-linear operators M_j and H_j

and, thus, is itself also linear,

$$G_j^T := \{M_1^T \dots M_{j-1}^T M_j^T H_j^T\}. \quad (2.15)$$

The adjoint integration of a state can be considered an evolution ‘backwards in time’ since M_j^T performs the model integration from t_j to t_{j-1} . H_j^T maps from the observation phase space into the model phase space.

For detailed derivations and additional information on 4D-var, consider Elbern et al. (1997), Bouttier and Courtier (2002), Rabier and Liu (2003), Courtier et al. (1994), and Lahoz et al. (2010). The 4D-var based data assimilation is well established and employed in several operational models. For example, in the IFS (Integrated Forecasting System) from the European Center for Medium range Weather Forecasting (ECMWF, 2024) for numerical weather prediction and in the context of chemistry transport models, for example, in the Sulfur Transport Eulerian Model (STEM; Zhang et al., 2008), and in the EUROpean Air pollution Dispersion-Inverse Model (EURAD-IM; Elbern et al., 2007). The latter is capable of emission factor optimization, used in the following evaluations and discussed in detail in the next section.

In comparison to the EnKF mentioned previously, the 4D-var method has advantages and disadvantages. Developing the adjoint model for the calculation of the gradient increases the complexity in the development of a 4D-var algorithm compared to an EnKF. Furthermore, the 4D-var algorithm relies on the hypothesis that the model is without errors, which is not a necessary assumption for an EnKF. However, the key advantages are the ability to incorporate observations according to their temporal distribution, and the physically consistent time evolution of the model state. Furthermore, emission factor corrections for SO_2 are shown to be slightly more accurate with 4D-var than EnKF, especially with high spatio-temporal variations in the emission source, whereas the EnKF shows advantages for strong emission corrections (Hu et al., 2023). A direct comparison of these algorithms can be found in Kalnay et al. (2007), and Buehner et al. (2010a,b).

2.2. The atmospheric chemistry transport model

EURAD-IM

The EUROpean Air pollution Dispersion-Inverse Model (EURAD-IM; Elbern et al., 2007) is a Eulerian meso-scale chemistry transport model. It is capable of 3D-var (3D-variational data assimilation) and 4D-var data assimilation, where the latter includes procedures to jointly optimize initial values and emission factors, as described in Sect. 2.1. It is part of the Copernicus Atmosphere Monitoring Service (CAMS) regional ensemble (Marécal et al., 2015; Colette et al., 2025) and has been applied in a variety of studies related to air quality, episode scenarios (De Souza Fernandes Duarte et al., 2021; Franke et al., 2024; Gama et al., 2019; Franke et al., 2022; Deroubaix et al., 2024), and observation instrument studies (Erraji et al., 2024; Tillmann et al., 2022).

The case study and the optimal observation network evaluations in this thesis are performed using the EURAD-IM.

The model domain covers Europe with a Lambert conformal conic projection with 15 km $\times$ 15 km horizontal resolution for the entire domain. The vertical resolution spans 30 layers defined on σ -coordinates

$$\sigma_k = \frac{p_k - p_{top}}{p_{bot} - p_{top}}, \quad (2.16)$$

with p_{top} and p_{bot} signifying the pressure at the top and bottom of the model domain, respectively. p_k is the pressure of the layer with index k . In the presented analysis, the pressure at the model top is defined as $p_{top} = 100$ hPa.

2.2.1. Forward model

Figure 2.1 shows a simplified schematic of the EURAD-IM forward model including inputs and outputs. The meteorological and terrestrial data is retrieved from the Weather Research and Forecasting Model (WRF) version 3.7 and the related preprocessors (WPS) (Skamarock et al., 2008) using the initial and boundary values from the Integrated Forecasting System (IFS) (Hersbach et al., 2025). The chemical boundary values for the EURAD-IM are obtained from the Composition-IFS (C-IFS) (Inness et al., 2019). The anthropogenic emission inventory used in this thesis are taken from the CAMS-REG-AP (version 3.1) inventory (Kuenen et al., 2022) and supplemented with the reported emission values from Umweltbundesamt Deutschland (2019) for Germany. The used inventories contain emission data for CO, NO_x, SO₂, NMVOCs (non-methane volatile organic compounds), NH₃ (ammonia), PM₁₀, and PM_{2.5}.

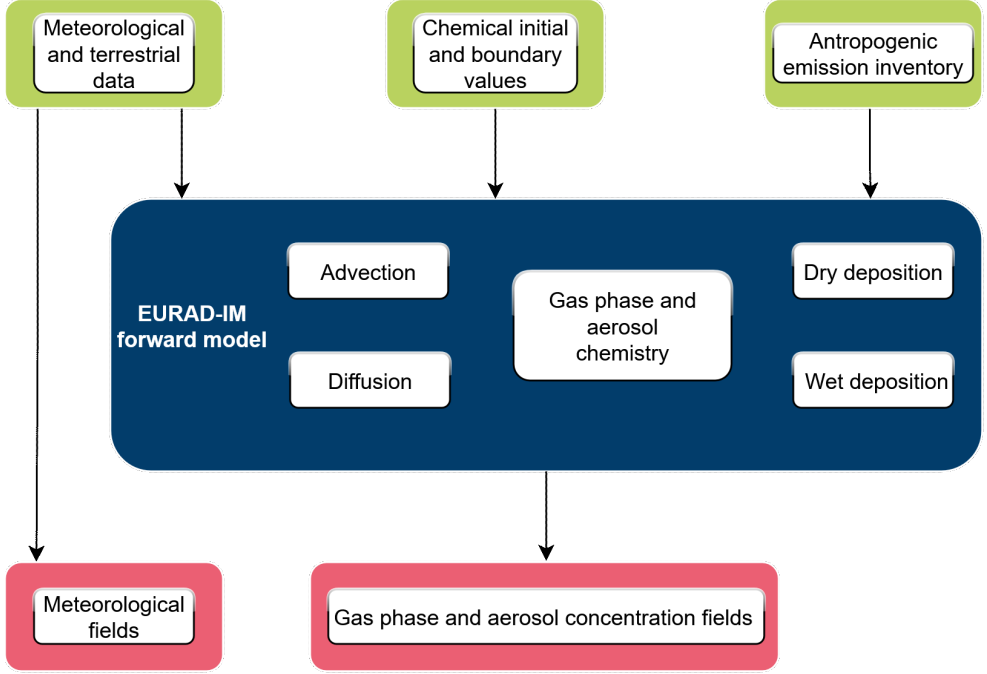


Figure 2.1. Simplified schematic of the EURAD-IM forward model. The inputs are shown in green, the outputs in red and the forward model itself as a blue box containing various processes.

The forward model of EURAD-IM solves the partial differential equations for the individual chemical constituents (Seinfeld and Pandis, 2016):

$$\frac{\partial c_i}{\partial t} = -\nabla(\mathbf{v}c_i) + \nabla\left(\rho\mathbf{K}\nabla\frac{c_i}{\rho}\right) + A_i + E_i - S_i. \quad (2.17)$$

Here, c_i denotes the mixing ratio of a chemical species, $\mathbf{v}$ is the wind velocity vector, ρ describes the air density, $\mathbf{K}$ is the operator governing eddy diffusion, A_i is the chemical generation of the species, E_i is the emission flux, and S_i is the deposition flux. The model is capable of describing about 110 different chemical species with 265 reactions.

Equation (2.17) is solved using operator splitting (McRae et al., 1982; Elbern et al., 2007), where each process is treated in sequence. The evolution of a state vector $\mathbf{x}$ from time t to $t + \Delta t$ is described by

$$\mathbf{x}(t + \Delta t) = T_h T_v D_h D_v C D_v D_h T_v T_h \mathbf{x}(t), \quad (2.18)$$

where T_h denotes horizontal transport and T_v vertical transport. D_h is the horizontal diffusion and D_v describes the vertical diffusion, and C is the chemical transformation, including the change due to emissions and deposition. Due to this splitting, the time step for the

transport and diffusion is $\Delta t_{T,D} = \Delta t/2$, while the time step for the chemical solver is even further dynamically reduced. For this thesis, the forward model of the EURAD-IM uses the Walcek scheme for advection (Walcek, 2000), and the methodology from Pleim and Chang (1992) for diffusion. The chemistry mechanism covering the gas-phase is the RACM-MIM (Regional Atmospheric Chemistry Mechanism - Mainz Isoprene Mechanism; Geiger et al., 2003; Stockwell et al., 1997), and the mechanism describing the aerosol thermodynamics is MADE (Modal Aerosol Dynamics model for Europe; Ackermann et al., 1998). The deposition of chemical components is split into dry- and wet- deposition. The dry deposition is implemented according to the method of Zhang et al. (2003) for the gas-phase, and the method by Petroff and Zhang (2010) for aerosol species. The wet deposition follows the model by Roselle and Binkowski (1999).

The common output of the EURAD-IM forward model are gas phase and aerosol concentration fields in an hourly resolution for each model grid cell, including the vertical layers.

2.2.2. Data assimilation scheme

The EURAD-IM is capable of assimilating various types of observations, depending on the use case. In the application within the CAMS regional ensemble, hourly, in situ, ground-based observations are used (Colette et al., 2025), while Erraji et al. (2024) directly assimilated in situ observations from drones.

The 4D-var data assimilation implemented in EURAD-IM follows the method presented in Sect. 2.1 with the cost function Eq. (2.13) and its gradient Eq. (2.14). Figure 2.2 details the data assimilation cycle in the EURAD-IM.

Each data assimilation cycle proceeds as follows: First, the EURAD-IM forward model (described in Sect. 2.2.1) is used to propagate the initial state under consideration of the emissions to the end of the assimilation window. The initial state and emission factors of the first iteration of the data assimilation cycle are the so called background states. In subsequent iterations, the considered initial state and emission factors are the results of the optimization of the previous iteration. The forecast is then compared to the observations and the cost function according to Eq. (2.13) is computed. The adjoint of the forward model is used to compute the gradient of the cost function, Eq. (2.14). As shown in Elbern et al. (2007), the adjoint model is comprised of the adjoint of the operators in Eq. (2.18) and capable of performing the backward time iterations. The obtained gradient is then used to minimize the cost function by means of the Limited-memory-Broyden-Fletcher-Goldfarb-Shanno algorithm (L-BFGS; Liu and Nocedal, 1989). From this minimization, the updated initial state and updated emission factors are obtained and used as input for the next iteration. Due to the high dimensionality of the cost function and the tangent linear approach, the minimization is not guaranteed to find the global minimum. Therefore, multiple data assimilation cycles are performed until a termination criterion is met. The initial state and emission

factors for the analysis forecast are selected from the iteration with the lowest costs and, together with their resulting analysis forecast, form the model output. When subsequent assimilation cycles are performed, the model state of the analysis at the end of the previous assimilation window and the optimized emission factors are used as background state for the next assimilation cycle.

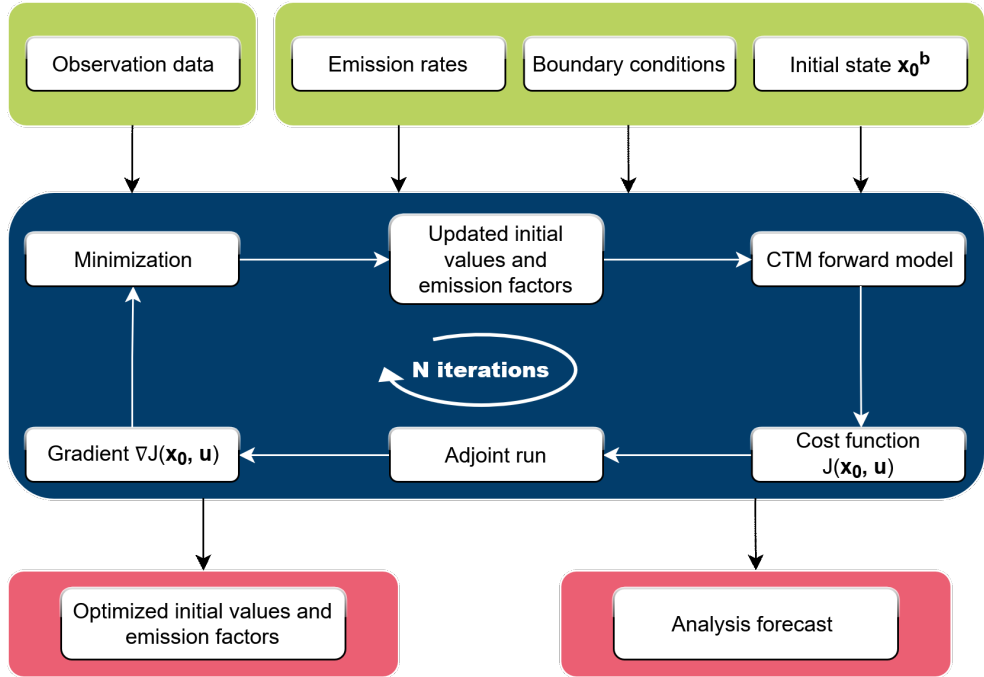


Figure 2.2. Flowchart of the 4D-var assimilation process implemented in the EURAD-IM. The input data is shown in green, the assimilation cycle itself in blue, and the output in red. x_0^b denotes the background model state at initial time, x_0 the initial state for each iteration, and u the transformed emission factor.

2.3. Theory of observing system simulation experiments

Observing System Simulation Experiments (OSSEs) (Arnold and Dey, 1986) are a type of experiment used to analyse the benefit and effectiveness of the assimilation of different observing systems on the forecasts. Instead of being restricted to existing observations, an OSSE can examine future observations (Lahoz and Schneider, 2014). These are extracted as artificial observations from a model output, the nature run, serving as the true atmospheric state, and their impact on data assimilation models is evaluated.

An OSSE is initiated with the generation of the nature run, which, in the context of air quality modeling, is the model output of a CTM. A different CTM, or a different CTM configuration, is then used to assimilate these artificial observations. The ability of the analysis to replicate the nature run quantifies the value of the assimilated observations for the accuracy of the simulation.

Since the costs of developing and deploying satellite instruments exceed the cost of OSSE model simulations, OSSEs are commonly used to measure the effectiveness of, and improve, potential satellite observation configurations.

For example, the influence of satellite observations on the respective tropospheric model concentrations from the Copernicus Sentinel-5P (ESA, 2025b) and the Sentinel-4 (ESA, 2025a) missions has been evaluated by Quesada-Ruiz et al. (2020) for O_3 , and by Timmermans et al. (2019) for NO_2 observations with an OSSE setup. Hereby, the observations are extracted from the performed nature run such that they resemble measurements from the respective satellite instruments. Quesada-Ruiz et al. (2020) showed that the model analysis benefits most at around 200-500 hPa, while the impact is insignificant on the low troposphere or the surface. Timmermans et al. (2019) distinguished that Sentinel-4 geostationary NO_2 observations benefit the modelled NO_2 columns during the entire day, while the Sentinel-5P low-Earth orbit observations benefit the forecast for 3-6 h after the overpass. Yumimoto (2013) evaluated the impact of different satellite deployment options, orbits and instruments, for CO observation on air quality forecasting over East Asia. They showed that observations from a multi-spectral geostationary satellite with near infrared and thermal infrared instruments showed about 30 % better impact due to higher observation frequency and 90 % better impact due to the combination of instruments on 60 h CO forecasting than observations from low-earth orbit thermal infrared observations. Chen et al. (2025) investigated the effects of different ground-based spatial observation densities for aerosol species on air quality forecasting. They concluded that assimilating data with aerosol component information outperforms the assimilation of PM_{10} and $PM_{2.5}$ with respect to 48 h forecasts, and that high density observation networks perform worse than lower density networks.

A general review on the OSSE methodology can be found in Masutani et al. (2010), while Timmermans et al. (2015) offer a review with a focus on OSSEs in air quality modeling ap-

plications. For the purpose of this work, the OSSE methodology is described in detail. A schematic overview, adapted from Timmermans et al. (2015), is shown in Fig. 2.3. The first step of an OSSE is the generation of the nature run (NR). It serves as ground truth, from which the artificial observations are extracted, and for the comparison with the assimilation runs to evaluate the OSSE. Therefore, it is important that the nature run and the model generating it are capable of accurately representing the investigated species and the atmospheric state as realistically as possible, as systematic errors can influence the entire experiment. The criteria to verify the realism of the nature run are described subsequently. Depending on the experiment setup, the nature run is generated from the output of a CTM (Model 1) with or without data assimilation. Any assimilated data is not used in the later steps of the OSSE. When designing an OSSE with the goal of investigating a specific instrument, the spatio-temporal resolution of the nature run must match or exceed the proposed instrument's capabilities. This ensures the generation of sufficient data to accurately replicate the instrument's measurements.

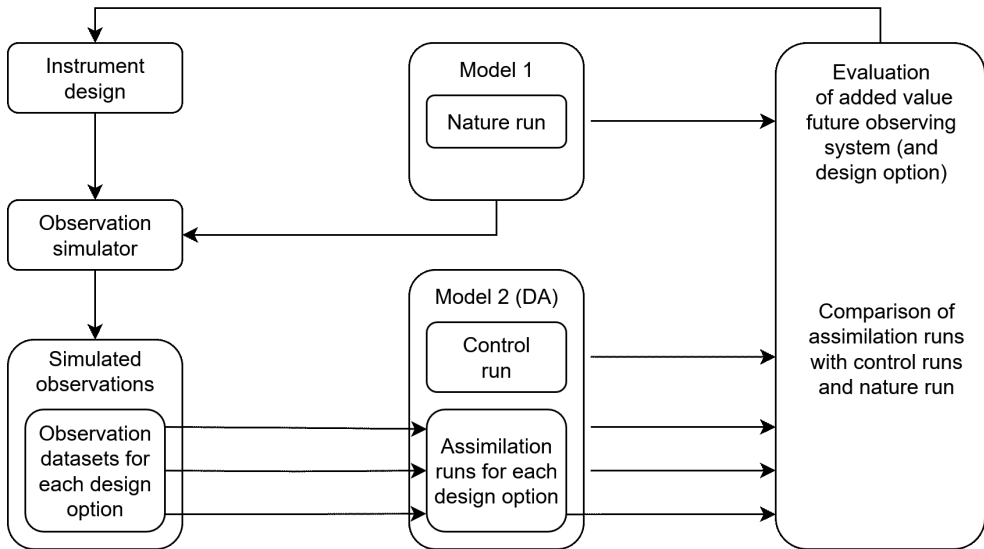


Figure 2.3. Schematic of the OSSE framework, adapted from Timmermans et al. (2015). ‘DA’ is short for data assimilation.

Artificial observations are sampled from the nature run according to the specifics of the investigated measurement instrument and observation configuration. This includes the expected error of the simulated measurement device as well as spatial and temporal resolution. Depending on the goal of the OSSE, instrument restrictions need to be accounted for. In the case of satellite based observations, for example, the instruments measure electromagnetic radiation in different frequency bands, which is altered by the atmospheric column. Therefore, atmospheric phenomena, such as cloud cover, can interfere with these measurements

and need to be accounted for in the extraction of the observations. A detailed description on the simulation of satellite observations for OSSEs can be found in Timmermans et al. (2015).

The flexibility in the design of the artificial observation simulation allows for the intercomparison of the impact of a wide range of observation systems and combination of instruments on model results for the same atmospheric state. Furthermore, the placement and number of observations can be varied to assess the respective influence on the model simulation (Masutani et al., 2010), as examined by Chen et al. (2025) for example.

The observation dataset generated from the observation systems is used in the data assimilation procedure of a second CTM or second CTM realization (Model 2). This analysis is the assimilation run (AR) of the OSSE. Furthermore, a separate model output is generated from Model 2, in which the same observations, if any, are assimilated as in the generation of the nature run. It is labeled control run (CR) and serves as baseline comparison for the impact of the observation datasets on the model output and for the evaluation of the realism of the nature run. The difference and statistics between the CR and NR should be comparable to the differences between an operational model setup and real observations, when averaged with respect to the spatio-temporal resolution of the model. This ensures that the impact of the artificial observations on the analysis of the assimilation run is comparable to the impact of real observations. Conclusions about the observation configurations thus translate to real setups. Depending on the use case, the NR can be compared to real observations directly in order to evaluate its validity. An example for this can be found in Claeysman et al. (2011), where the bias, RMSE, and correlation of air quality monitoring stations for various species are shown to be similar to the nature run. In Timmermans et al. (2019) the 1 h averaged diurnal cycle of O_3 and NO_2 , and in Quesada-Ruiz et al. (2020) the 1 h averaged diurnal cycle of O_3 of the NR is directly compared to the averaged diurnal cycle extracted from observations.

The ARs are compared to the NR and the CR to quantify the impact of the investigated observation configurations on the modeled atmospheric state. The NR and the CR serve hereby as upper and lower bounds for the added value provided by the observation systems. This allows for quantitative estimations on the capability of Model 2 with the added observations to replicate the NR, which represents the true state of the atmosphere. The comparison of different ARs, generated by different observations, can furthermore provide estimates about the impact of the different observation systems. This information can then be used to optimize the instrument design and observation system attributes in further experiments to improve the AR with respect to the NR.

To ensure the validity of an OSSE, it is important that Model 1 differs from Model 2. When the models are too similar, or even identical, the informative value of the OSSE is decreased. Certain deficiencies or systematic errors in simulating the state of the atmosphere are circumvented, since they are included in the artificial observations, and the analysis of

their assimilation will be overly optimistic (Masutani et al., 2010). This overconfidence in the skill of the model and the underestimated influence of the data is referred to as the ‘identical twin problem’ (Arnold and Dey, 1986). This can be avoided by either using an entirely different model for the NR than for the AR and CR, or by changing key aspects of Model 1 to obtain Model 2. The latter ought to be achieved such that the NR differs from the CR while still retaining realistic atmospheric variations. A common choice in air quality applications is the change of the emission inventory between Model 1 and Model 2 (Edwards et al., 2009; Chen et al., 2025). When Model 2 is generated by changing aspects of Model 1, it is important to consider the influence of these modifications on the results of the OSSE.

3. Data clustering to optimise the representativity of observational data in air quality data assimilation: a case study with EURAD-IM (version 5.9.1 DA)

The following chapter is published as: "Data clustering to optimise the representativity of observational data in air quality data assimilation: a case study with EURAD-IM (version 5.9.1 DA)", authored by A. Hermanns, A.C. Lange, J. Kowalski, H. Fuchs, and P. Franke in Geoscientific Model Development, in 2025. DOI: <https://doi.org/10.5194/gmd-18-9417-2025> under the Creative Commons 4.0 License (<https://creativecommons.org/licenses/by/4.0/> (last access: 15. December 2025)).

The authors contributed to the article as follows:

- Contributed to conception and design: AH, PF, and ACL.
- Contributed to acquisition of data: AH.
- Contributed to analysis and interpretation of data: AH, PF, ACL, JK, and HF.
- Drafted and/or revised the article: AH, PF, ACL, JK, and HF.
- Confirmed the submitted version for publication: AH, PF, ACL, JK, and HF.

The publication addresses the issue of how to generate a representative analysis in an operational air quality data assimilation setting, where the set of available observations is fixed. In data assimilation, a fraction of the available data is withheld from the assimilation procedure for the independent validation of the analysis. This splitting of the data, if not performed carefully, can result in biases in the observations used for assimilation. This could lead to a less accurate analysis of the atmospheric state than is possible with the available data. Thus, the publication proposes a method for optimally sampling the available data to create representative assimilation and validation data sets.

The available observations are characterized by intrinsic properties: the mean and variance of the measured annually averaged diurnal cycle, observation latitude, longitude, and altitude. Based on these properties, the observation locations are subject to clustering algorithms. Two algorithms are investigated, k-means and k-means soft constrained, each using a different subsets of the intrinsic properties. The resulting clusters contain observations with common features, illustrating that the clustering is reasonable. Sampling each of these clusters equally generates a representative assimilation and validation data set.

These are tested using the EURAD-IM for simulating four months in 2016. The representativity of the data assimilation analysis is evaluated by calculating the difference in RMSE between the assimilation and the validation data set. It is demonstrated that the derived assimilation data sets improve the representativity of the analysis compared to an operationally used split.

The thesis analyzes the impact of different observational networks on the emission rate optimization in a 4D-var setup. This study initializes this assessment focusing on the optimal use of existing ground-based observational networks, especially illustrating the need for representative data sets for assimilation and validation. In subsequent sections, this analysis will be advanced to find optimal observational networks beyond existing infrastructures.



Data clustering to optimise the representativity of observational data in air quality data assimilation: a case study with EURAD-IM (version 5.9.1 DA)

Alexander Hermanns^{1,2}, Anne Caroline Lange¹, Julia Kowalski³, Hendrik Fuchs^{1,2}, and Philipp Franke¹

¹Forschungszentrum Jülich, Institute of Climate and Energy Systems – Troposphere (ICE-3), Jülich, 52425, Germany

²Faculty of Mathematics and Natural Sciences, University of Cologne, 50923 Cologne, Germany

³Faculty of Mechanical Engineering, RWTH Aachen University, 52062 Aachen, Germany

Correspondence: Philipp Franke (p.franke@fz-juelich.de)

Received: 31 January 2025 – Discussion started: 16 May 2025

Revised: 8 August 2025 – Accepted: 15 October 2025 – Published: 3 December 2025

Abstract. In the field of air quality analysis, data assimilation is commonly used to integrate information on the atmospheric state provided by observations into the model. However, the analysis is largely dependent on the data available to the assimilation system. In order to obtain an accurate analysis of the true state of the atmosphere, the representativity of the utilized data becomes a fundamental requirement. Here, a method is presented that derives a representative split of the ground-based monitoring network data that depends only on the characteristics of the observation data. The core of the methods is a clustering algorithm to subdivide the data into subsets. Two clustering algorithms, k-means, and k-mean soft constraint, are tested and applied to air pollutant observations in Europe. The clusters are solely derived from observation intrinsic properties (such as geographic location and averaged concentrations). The resulting clusters reliably distinguish common features of the observational data, e.g. mean and variance of averaged air pollutant concentrations. Representativity of the observational data in the assimilation and validation subset is ensured by sampling each cluster individually. The method is tested using the assimilation system of the chemistry transport model EURAD-IM (European Air pollution Dispersion – Inverse Model) and evaluated for data from four months in 2016. A significant improvement of the analysis' representativity, quantified by the difference between the analysis' root mean square error with respect to the assimilation and validation dataset, is found in the results. Compared to an operational configuration, the largest improvement in the relative representativity measure is evaluated for CO with 16 %, for NO₂ with 4 %, and for O₃

with 1 %. A reduction in the relative representativity measure is observed for SO₂ with –5 %, for PM₁₀ with –2 % and for PM_{2.5} with –5 %, although these differences do not lead to significant deviations in absolute values given the overall error and the improvement for CO outweighing the changes in the other species.

1 Introduction

Air pollution is the main environmental health risk (World Health Organization, 2021). Therefore, its assessment and reliable prediction are essential to further develop policies aiming at improved air quality. Although observations of air pollutants via air quality monitoring networks, mainly consisting of ground based in situ observations, are routinely used to monitor and evaluate air pollution, they only show a subset of the atmospheric state. A common approach is to complement the observations by model simulations, which show a more complete representation of the atmospheric state but lack of accuracy due to insufficient input data (e.g., initial values, boundary values, and emission data). Data assimilation is a method that integrates information from observational data into the model state with the objective of obtaining an analysis that is, in theory, the best available representation of the atmospheric state. However, this requires representative observations to be assimilated in order to obtain more accurate analyses with respect to the (unknown) true atmospheric state and atmospheric parameters, such as

air pollutant emissions. The Copernicus Atmosphere Monitoring Service (CAMS, <https://atmosphere.copernicus.eu/regional-air-quality-production-systems>, last access: 7 August 2025) employs an ensemble of air quality models using data assimilation to provide daily air quality forecasts and re-analysis over Europe. An overview of the ensemble and the models used within can be found in Marécal et al. (2015) and Colette et al. (2025). In data assimilation applications, observational data are typically split into an assimilation and a validation set, where the former is used for the data assimilation procedure and the latter is withheld and used to evaluate the analysis with independent observations. This split of the available data is known as sub-sampling. Each subset of the observational data must be representative of the atmospheric state to obtain the highest accuracy of the analysis.

Various measures to estimate the representativity of observational data have been explored, yet no consensus on a universal approach has been reached (Kracht et al., 2017). Nappo et al. (1982) developed a general definition for representativity for meteorological applications: “a point measurement is representative of the average in a larger area (or volume) if the probability that the squared difference between point and area (volume) measurement is smaller than a certain threshold more than 90 % of the time”. This definition is employed in the work of Piersanti et al. (2015) to derive areas of representativity for air quality monitoring stations in Italy and to show their seasonal variability. Other methods determined the redundancy of air-quality observations in urban areas and their correlation with satellite data to define the representativity. For example, Ma et al. (2019) found PM₁₀ (particulate matter with an aerodynamic diameter of 10 µm or less) to be the primary air pollutant in Lanzhou, China, from July to December 2015, and identified the most representative observation site from a group of four. Su et al. (2022) analysed the spatial representativity of air quality monitoring networks estimated using environmental influence factors on PM_{2.5} concentrations (particulate matter with an aerodynamic diameter of 2.5 µm or less). With this, they applied k-means clustering to identify optimal locations for additional monitoring stations in under-represented areas, which resulted in additionally 40 monitoring stations and a 15 % increase in the representativity. Ignaccolo et al. (2008) converted discrete time series of pollutant observations in the Piemonte region, Italy, into spline coefficients and identify similarities using a Partitioning Around Medoid (PAM) clustering algorithm. They were able to show different spatial patterns in the clusters for various species depending on anthropogenic and geographical influences. The study by Joly and Peuch (2012) characterizes pollutant time series by 8 parameters derived from their daily-, monthly-, seasonally- and annually averaged diurnal cycles. With this, they were able to derive 10 classes of air pollution monitoring sites using linear discriminant analysis, which allow for the classification of new sites without the need to recalculate the classification. This classification is used by the Copernicus Atmosphere

Monitoring Service (CAMS) to filter observational data in order to improve their regional products (Peuch et al., 2022).

In contrast to most literature, this study focuses on generating a representative sub-sampling of the available observational data instead of assessing the representativity of individual observational data directly. For this, a method for the generation of the sub-samples is developed that bases on data clustering and grouping the available observation stations according to their characteristics. The advantage of this data based approach is that it can be applied to the input data of the assimilation system and does not require prior model evaluations. Here, the relative representativity of two datasets is determined by quantifying and subsequently comparing the difference in the RMSEs between the model analysis and the observations from the assimilation and validation data set. This measure is hereafter called AV-difference and is described in detail in Sect. 3.4.

Cluster analysis is a statistical data analysis method aiming to identify subgroups (clusters) of similar data points (objects) within a data set based on their features. The similarity of objects can be measured as the mutual distance in feature space, where the exact distance measure may vary depending on the employed method. Objects in the same cluster are considered more similar to each other than to objects in other clusters. Cluster analysis has been applied in atmospheric applications for various purposes. Some examples include the improved identification of patterns of air quality regime classifications for ozone (O₃), nitrogen dioxide (NO₂), sulfur dioxide (SO₂), and PM₁₀ for observation stations in Germany (Flemming et al., 2005). Lyapina et al. (2016) clustered European ozone observations based on monthly averaged diurnal cycles and showed that different O₃ regimes (i.e., groups of observation stations with similar statistical characteristics) are not geographically separated. Carro-Calvo et al. (2017) identified zones with coherent spatio-temporal patterns in observations of summer ozone concentrations and were able to link weather patterns to the occurrence of extreme values. Utilizing temperature and precipitation, Fovell and Fovell (1993) were able to identify four climate zones in the United States and further divided them into subzones by using a higher number of clusters. A review about clustering applications and methodologies in air pollution studies from 1980 to 2019 can be found in Govender and Sivakumar (2020).

As a common clustering algorithm, k-means was applied to derive correlations between O₃ and NO₂, nitrate radical (NO₃), night-time nitrous acid (HONO), and formaldehyde (HCHO) by Wang et al. (2022). In addition, (Borge et al., 2022) used k-means to identify zones of common air quality in Madrid. On global scales, the method was used to analyse aerosol regimes (Li et al., 2022).

2 Chemical transport model EURAD-IM

The EUROpean Air pollution Dispersion – Inverse Model (EURAD-IM Elbern et al., 2007; Franke et al., 2024) is used to evaluate the effect of different sampling strategies of observational data in a realistic modeling framework. EURAD-IM is a chemistry transport model capable of three- and four-dimensional variational (3D-/4D-var) data assimilation applied for air quality evaluation (De Souza Fernandes Duarte et al., 2021; Tillmann et al., 2022; Liu et al., 2024) and analyses (Franke et al., 2022; Erraji et al., 2024). Here, 4D-var is used for jointly optimizing initial values and emission factors (Elbern et al., 2007). The utilized adjoint model comprises all relevant processes that are included in the forward model, namely advection, horizontal, and vertical diffusion, chemical transformation, and secondary inorganic aerosol formation. The model domain covers Europe with a $15\text{ km} \times 15\text{ km}$ horizontal resolution and includes 30 vertical layers up to 100 hPa with the additional option of nesting domains ($5\text{ km} \times 5\text{ km}$ for central Europe, $1\text{ km} \times 1\text{ km}$ for smaller domains). Meteorological input data are retrieved from the Weather Research and Forecasting Model (WRF) version 3.7 (Skamarock et al., 2008) with Integrated Forecasting System (IFS) forecast data (Inness et al., 2019) taken as initial and boundary values. Chemical boundary values are provided by the complementary chemistry modules of the Composition Integrated Forecasting System (C-IFS). Anthropogenic emissions are taken from the German Environment Agency (Submission 2019, version 2) for Germany, due to the available data being directly aggregated onto the model grid, and from the Copernicus Atmosphere Monitoring Service's regional anthropogenic emission inventory (CAMs-REG-AP, version 3.1) for Europe. It is worth noting that the proposed sampling method is independent of the used data assimilation model and the EURAD-IM simulations solely serve as a validation of the proposed method.

3 Cluster analysis

The observational data used in this study are ground-based observations used within CAMS regional analysis. The data is pre-filter to the classes 1–7 according to Joly and Peuch (2012) from data accessible at the European Environmental Agency (EEA) via <https://eeaadmz1-downloads-webapp.azurewebsites.net/> (last access: 9 June 2025). They consist of measurements of hourly concentrations of carbon monoxide (CO; 382 stations), nitrogen dioxide (NO₂; 1743 stations), ozone (O₃; 1712 stations), sulfur dioxide (SO₂; 1005 stations), PM₁₀ (1029 stations), and PM_{2.5} (509 stations) from 40 EEA member and cooperating countries. In total, 6380 individual time series from 2279 stations are used after filtering the data as outlined in Sect. 3.3. An overview of the geographic distribution of the available observation stations for each species is shown in Fig. 1. The density of CO observa-

tions compared to the other measured species is substantially lower, even in highly populated areas.

The clustering process divides observation stations into subgroups (clusters), determined by the algorithm itself and the features used. Finally, by sub-sampling the clusters into assimilation and validation data, the different observational features can be included equally in both data sets. The representativity of these sets obtained by the clustering is analysed in this work. An overview of the methodology is shown in Fig. 2, where the individual components are detailed below.

3.1 K-means algorithm

In this study, clustering algorithms based on k-means (MacQueen, 1967), sometimes referred to as the Lloyd–Forgy algorithm (Forgy, 1965), are employed. K-means aims to split a number of n objects ($d_1 \dots d_n$) with feature vectors $f(d_i) \in F$, elements of the vector space F , into K clusters following the algorithm outlined below:

1. Initialize K random cluster centres (centroids) $C_1 \dots C_K$ as elements in F .
2. For all i , assign each object d_i to the closest cluster C_j by minimizing the squared Euclidean distance, Eq. (1).
3. For all clusters, compute the new cluster centres C'_j by averaging the feature vectors within the cluster.
4. Iterate step 2 and 3 until the cluster centres $C_1 \dots C_K$ do not change or until a fixed number of iterations is completed.

In the application of this work, objects are observation stations and the features are the station characteristics (e.g., latitude, longitude, altitude, and observed concentrations of the different species). The squared Euclidean distance between the features of d_i and the centroid C_j is defined as

$$\text{dist}_{\text{euclid}}(d_i, C_j) = \|f(d_i) - f(C_j)\|^2, \quad (1)$$

where $\|\cdot\|$ is the Euclidean norm. The elements of the feature vectors are normalized to a dimensionless value between $[0, 1]$ in order to ensure that each feature affects the distance value with the same weight. The number of clusters K is chosen before the computation. Different approaches exist to find the optimal number of clusters. In this study, the cluster hyperparameters (i.e. number of clusters K) are determined by selecting those that maximize the mean silhouette score $\bar{s}$, as defined by Rousseeuw (1987),

$$\bar{s} = \frac{1}{\#d_i} \sum_{d_i} \frac{b(d_i) - a(d_i)}{\max\{a(d_i), b(d_i)\}}, \quad (2)$$

where $a(d_i)$ is the averaged distance of the object d_i to all other objects in the same cluster C_j and $b(d_i)$ the averaged distance of d_i to all elements in the next closest cluster, $\#d_i$ is the total number of objects d_i .

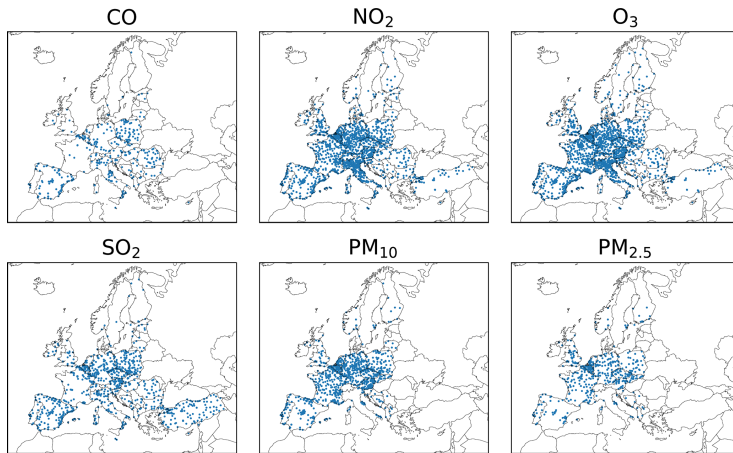


Figure 1. Distribution of available observation stations per species in Europe in 2016.

For $d_i \in C_I$,

$$a(d_i) = \frac{1}{\#C_I - 1} \sum_{d_j \in C_I} \text{dist}(d_i, d_j), \quad (3)$$

$$b(d_i) = \min_{J \neq I} \frac{1}{\#C_J} \sum_{d_j \in C_J} \text{dist}(d_i, d_j), \quad (4)$$

with $\#C_I$ denoting the number of objects in the cluster C_I . Although different choices to calculate the distance are possible, it is convenient to select the Euclidean distance Eq. (1) as distance functions in Eqs. (3) and (4). The mean silhouette score is an effective measure for the quality of the clustering. It returns values from $[-1, 1]$ with $\bar{s} \approx 1$ indicating an “optimal” clustering. For $\bar{s} \approx 0$ the result is inconclusive, since objects are equally distant to their assigned and second closest cluster, and $\bar{s} \approx -1$ indicates that d_i is assigned to the wrong cluster.

3.2 K-means soft constraint algorithm

The concentrations measured at an observational site contain valuable information in order to determine the similarity to other stations. Since not all species are measured at all stations, certain feature vectors do not contain valid values for all entries. Therefore, the k-means soft constraint algorithm (Wagstaff, 2004) is used for the clustering. It aims to identify clusters in data, where not all features have values for all objects. This avoids the need to fill-in missing values by interpolated values, which can lead to inaccuracies due to necessary assumptions. In the following, the method by Wagstaff (2004) is described.

The vector space of feature vectors F is separated into the features that are known for all objects, F_0 , and the features that are missing in some objects, F_m . Such that $F = F_0 \oplus F_m$.

Clustering is then performed for F_0 with constraints based on F_m . The constraint $\sigma(d_i, d_j)$ indicates the separation between the objects d_i and d_j in the subspace F_m . It is defined using the features $f_m(d_i)$ and $f_m(d_j)$ from the vector subspace F_m , with

$$\sigma(d_i, d_j) = \sqrt{\sum_{f_m \in F_m} (f_m(d_i) - f_m(d_j))^2}. \quad (5)$$

Features that are missing in both objects do not contribute to the constraint. The k-means soft constraint algorithm consists of the same steps performed in the classical k-means algorithm. However, it uses a modified distance function

$$\text{dist}_{\text{KSC}}(d_i, C_J, \omega) = (1 - \omega) \frac{\text{dist}_{\text{euclid}}(d_i, C_J)|_{F_0}}{V_{\max}} + \omega \frac{\Gamma_{d_i}}{\Gamma_{\max}}, \quad (6)$$

with $V_{\max}$ being the maximum variance considering all objects. Γ_{d_i} is the sum of all squared constraints for $d_i \in C_I$ with objects in the same cluster $d_j \in C_I$,

$$\Gamma_{d_i} = \sum_{d_j \in C_I} \sigma^2(d_i, d_j). \quad (7)$$

$\Gamma_{\max}$ is the sum of all squared constraints whether d_j is in the same cluster as d_i or not, i.e., $\Gamma_{\max} = \sum_{d_j} \sigma^2(d_i, d_j)$. Furthermore, a scaling parameter ω is introduced, weighting the influence of F_m in relation to F_0 . The first term of Eq. (6) is the squared Euclidean distance normalized by the maximum variance. Thus, for $\omega = 0$, the k-means soft constraint algorithm leads to the same result as a standard k-means algorithm in which only features in F_0 are included. The second term increases the distance of d_i to its assigned centroid

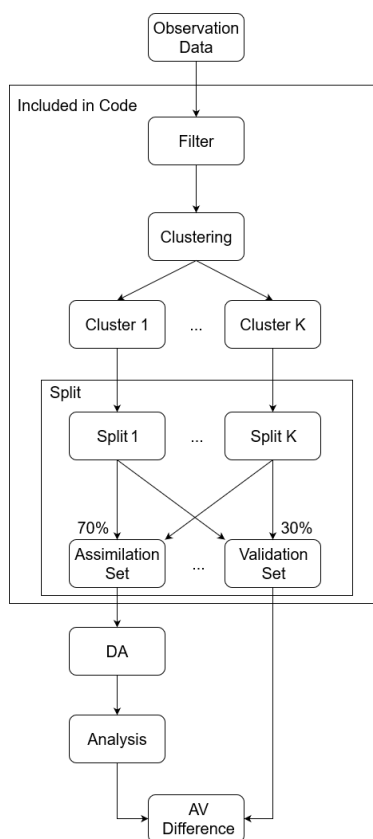


Figure 2. Flowchart of the proposed methodology. In the presented case, “Observation Data” is the observation data used in the CAMS-project. The elements in the box titled “included in code” are included in the provided software code (Hermanns, 2025), with the KSC algorithm as the “Clustering”. The box “split” is the extraction of the assimilation and validation set from a clustering result. “DA” is short for data assimilation. The “AV-difference” is the representativity measure and detailed in Sect. 3.4.

C_I scaled by the constraints. It is important to note that this limitation prevents the algorithm from converging in many applications due to an artificially increased distance between objects within the same cluster compared to objects in other clusters.

3.3 Data preparation and experimental design

Before the application of the data clustering to observational data, the data has to be prepared to avoid influence of outliers which can negatively affect the clustering quality (Nowak-Brzezińska and Gaibei, 2022). This is achieved by filtering the data with empirically determined thresholds.

Values below $0 \mu\text{g m}^{-3}$ or exceeding $50\,000 \mu\text{g m}^{-3}$ for CO and $5000 \mu\text{g m}^{-3}$ for all other species are removed from the observation data to exclude outliers. Time series for species with a relative standard deviation below 0.01 in the assimilation period are excluded, as it is assumed that the hourly concentrations of air pollutants usually have a greater temporal variability. Moreover, observations are removed for intervals where consecutive observations are constant for more than two hours. A summary of the filtering procedures and their effect on the amount of observational data is shown in Table 1.

Two different classes of features have been used to cluster the observational data. The first set of features in the clustering algorithms are the mean and the variance of the annually averaged diurnal cycles for each species. While the diurnal cycles vary throughout the year and can exhibit different behaviours depending on the season, weather conditions or human behaviour, an average over the entire year is not susceptible to short time-scale variations. This is done to better compare the test configurations with the reference, which is the configuration used in the CAMS reanalysis for 2016 (discussed later on). The placement of observations in the reference is static for the entire year and to ensure a fair comparison, the test configurations and thus their clustering features are also chosen to be static. The mean and variance of the annually averaged diurnal cycle are chosen to characterize the differences in the measurement sites. While other features are also suitable for clustering as was shown e.g., in Joly and Peuch (2012), restricting the number of features is important to apply meaningful clustering. In high dimensions, the “nearest neighbour” problem cannot be solved in all cases (Beyer et al., 1999).

The annually averaged diurnal cycle is calculated from the filtered observational data by computing the average for each hour of the day for the entire year. From this annually averaged diurnal cycle, the mean and the variance are calculated and form one set of the features used in the clustering algorithms. A filter is implemented where the standard deviation of these values is calculated for each species individually and measurement stations where the mean or the variance of the annually averaged diurnal cycle exceed 7.5 times the species-specific standard deviation are excluded.

The second set of features used in the clustering algorithms are the geographic coordinates, which are first converted to the Cartesian coordinate grid used in the EURAD-IM based on a Lambert conformal grid projection. To ensure that all features have the same weight, each feature is separately normalized to an interval of $[0, 1]$.

Two experiments are performed to evaluate the two clustering algorithms on the data. The first experiment applies the k-means clustering algorithm and is named Clustering-Diurnal (CD) in the following. The features consist of the mean and variance of the annually averaged diurnal cycle of CO, NO₂, O₃, SO₂, PM₁₀, and PM_{2.5} observations from 2279 air quality monitoring stations while the measurement

stations themselves are the objects. The feature vectors contain a maximum of 12 valid elements for each object. Since not all stations provide measurement data for all species, missing values are ignored for the calculation of the cluster centres. In the second experiment, the k-means soft constrained algorithm is applied, named KSC in the following. Here, the features are the geographical coordinates and the altitude for each measurement station as well as the mean and variance of the annually averaged diurnal cycle. The geographical coordinates and altitude are available for all objects (i.e., the measurement stations) and thus correspond to the features from the feature space F_o (cf. Eq. 6). The mean and variance of the annually averaged diurnal cycle are not available for all species for all objects and are used as constraining features of the feature space F_m . As the validation of a species at a measurement station may be influenced by the assimilation of measured concentrations of other species at the same location due to their chemical coupling, all observations from one measurement station are assigned to either the assimilation or the validation set (data bundling).

The optimal parameters used in both clustering algorithms are determined by maximising the silhouette score (Eq. 2). The number of clusters K is varied from 2 to 9 and the k-means soft constrained parameter ω from 1 to 0, with decreasing increments towards higher silhouette scores. In the case of the KSC experiment, the algorithm is not deterministic and $\bar{s}$ can vary depending on the initial choice for the cluster centers. For the CD experiment, the optimal number of clusters is $K = 2$ and for the KSC experiment, $K = 8$ and $\omega = 1.0 \times 10^{-6}$ are found to generate the largest silhouette score.

To ensure the equal representation of each cluster in the validation and assimilation data sets, the data split is applied randomly on each individual cluster. With this, 70 % of the data is assigned to the assimilation data set, while the remaining 30 % are taken as validation data. The data split is designed to ensure a random 70/30-split for each species in each cluster while maintaining the data bundling. The resulting data splits are referred to as observation configurations in the following sections. To verify that no unintended correlations between the assimilation and validation data sets have been introduced, which could affect the clustering process, each observation configuration is validated using a random forest classification algorithm (Breiman, 2001). Furthermore, the two experiments are compared with an observation configuration that was operationally used within the regional CAMS reanalysis for 2016 (REF experiment). For this, the same data filtering as described above is applied. The REF experiment contains the same observation stations as the KSC and CD experiments. The experiments differ only in the assignment of the observation stations to the assimilation/validation set used in the data assimilation runs. Note that the CAMS REF configuration applied a bundling of NO_2 with O_3 and PM_{10} with $\text{PM}_{2.5}$ observations (i.e., these species are always assigned to the same data set) before selecting semi-

randomly, where validation stations are placed near assimilation stations. Here, spatially isolated observations are used for the assimilation.

The impact of the station clustering on the representativity of the assimilation and validation data sets is evaluated using the 4D-var data assimilation system of EURAD-IM (Elbern et al., 2007). The data assimilation procedure optimizes the initial conditions as well as the emission factors for each grid box to obtain an improved analysis of the atmospheric state. The evaluation is performed on data from four different months in 2016 (January, March, June, and September) to identify potential seasonal effects. The assimilation is performed on 30 consecutive days within each month with an assimilation window of 24 h.

3.4 Definition of the representativity measure

The difference in the RMSE (Root-Mean-Square Error) of the analysed model state calculated for the assimilation and validation observations (hereafter AV-difference) is assumed as representativity measure with low values indicating a higher degree of representativity. A set of available observations (OBS) is split into an assimilation and validation set ($A + V$), hereafter observation configuration. To evaluate the split in terms of representativity, consider two distinct observation configurations with assimilation sets (A_1 and A_2) and validation sets (V_1 and V_2), with

$$\text{OBS} \rightarrow \begin{cases} A_1 + V_1, \\ A_2 + V_2. \end{cases} \quad (8)$$

Each assimilation set A_i is used in the data assimilation system M (e.g., the EURAD-IM as in this study), generating an analysis P_i

$$M : A_1 \mapsto P_1 \quad (9)$$

$$M : A_2 \mapsto P_2. \quad (10)$$

Note that $P_1 \neq P_2$ due to the different observation configurations. Amongst several possible options, the RMSE is chosen as an error metric. Here, the RMSE of the analysis is calculated for each simulation day for the assimilation and validation sets for individual hourly data pairs.

$$E_i^A := \text{RMSE}(P_i, A_i) \quad (11)$$

$$E_i^V := \text{RMSE}(P_i, V_i) \quad \text{for } i \in \{1, 2\}. \quad (12)$$

Then, the AV-difference is the absolute difference between E_i^A and E_i^V . To compare the observation configurations, the mean value of the AV-difference is calculated over the simulated days:

$$\text{AV}_i^{\text{diff}} := \overline{|E_i^A - E_i^V|}. \quad (13)$$

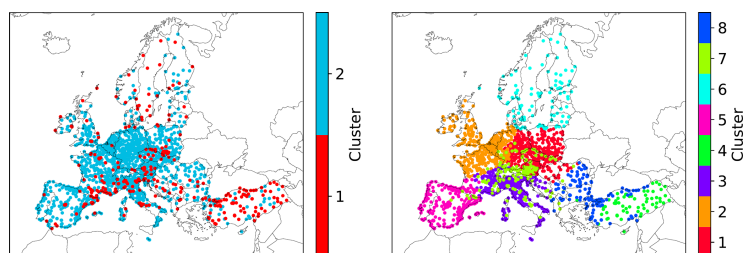


Figure 3. Distribution of the clusters for the CD (left) and KSC (right) experiment considering all observational sites in Europe regardless of the observed species. Different clusters are grouped by color.

A perfect representativity is achieved if the AV-difference vanishes, i.e., the RMSE of the analysis compared to the assimilation data is equal to the RMSE compared to the validation data. A low AV-difference indicates that the modeled concentrations at the assimilation stations are representative for the concentrations in unobserved regions, where validation stations are located. In turn, a high AV-difference shows an overfitting (or overconfidence) in observed areas compared to other regions. The model state is thus less representative, which limits the reliability of the analysis. The aim of this study is to evaluate the effect of different observation configurations on the representativity of the analysed atmospheric state. Although the analysis includes estimates of emission corrections of air pollutants, the evaluation of the accuracy of these corrections is not considered in this study.

An evaluation of the analysis' results in January shows that inaccuracies in the meteorological fields (especially in the wind fields) can lead to unrealistically strong emission corrections in certain locations. This is caused by a lateral displacement of the emission plume due to the erroneous meteorology, which results in a large RMSE at the surrounding observation stations. To account for this, stations with a RMSE larger than $4500 \mu\text{g m}^{-3}$ as well as stations with emission correction factors larger than 15 in their vicinity are excluded from the evaluation of the representativity. These thresholds are chosen empirically by investigating the model analysis and remove $\sim 2\%$ of observations. Both filtering criteria are rigorously applied for all analysed months. However, the former only applies in the January analysis. The emission correction factor based filtering affects all months for locations with a priori small emission strengths.

4 Results

4.1 Analysis of the data clustering

The grouping of observation stations into clusters for the CD and KSC experiments is shown in Fig. 3. The locations of the observational sites for each pollutant species are shown in Fig. 1. Further, Fig. 4 shows the normalized mean and

Table 1. Summary of the filtering procedures and the number of removed time series from the evaluation in percent.

Method	Removed time series [%]
Outlier removal	0.0
Relative standard deviation	< 0.1
Constant intervals	0.4
Annually averaged diurnal cycle outlier	0.3
RMSE outlier (January and CO only)	0.8
Emission factor outlier	2.2

variance of the annually averaged diurnal cycle for the two clusters of the CD experiment for the observed species.

In the CD experiment, 86 % of the stations (1349 stations in total) are assigned to cluster 2. This cluster contains observations with generally lower mean and variance of the annually averaged diurnal cycle (Fig. 4) than observations in cluster 1 (227 stations), except for O_3 . However, there is no spatial pattern in the two clusters since the geographical locations of the observation stations are not included as feature in the clustering process.

The clusters resulting from the KSC experiment are predominantly influenced by geographical location (Fig. 3). However, the boundaries between the clusters are not clearly defined, which would be seen in a k-means clustering that only takes the geographic location as features. This signifies the influence of the annually averaged diurnal cycle on the clustering results. Some agreement of the cluster structure with the Köppen-Geiger classification of climate zones in Europe (Peel et al., 2007) is found, showing that the clusters selected by the clustering algorithm are meaningful. Besides the geographical location, the clustering is influenced by the altitude of the observation station and the annually averaged diurnal cycle. This is illustrated in Fig. 5 for the altitude distribution for all clusters and in Fig. 6 for the annually averaged diurnal cycle of NO_2 averaged for all clusters used in the KSC experiment. Observation stations in clusters 4 and 7 show higher mean altitudes of 1024 and 964 m, re-

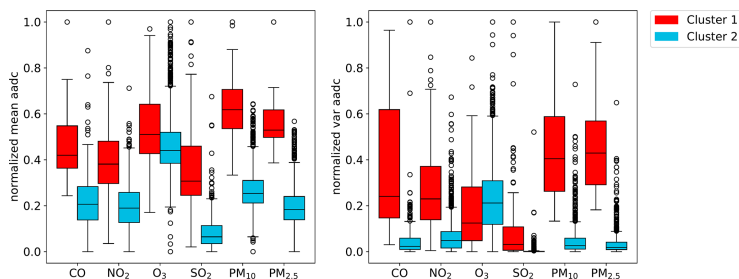


Figure 4. Comparison of the normalized mean (left) and variance (right) of the annually averaged diurnal cycle (aadc) per cluster generated in the CD experiment for each observed species. The boxes extend from the first to third quartile, the median is marked by the solid line, the whiskers extend to $1.5 \times$ the interquartile range. Outliers are marked as circles.

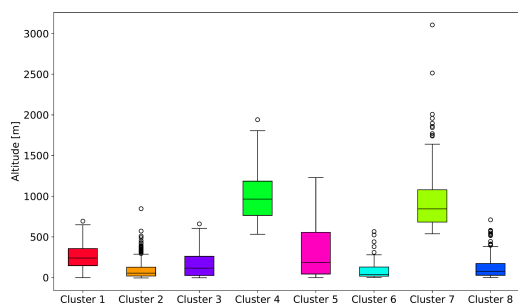


Figure 5. Distribution of the stations' altitude for all clusters obtained in the KSC experiment. The boxes extend from the first to third quartile, the median is marked by the solid line, the whiskers extend to $1.5 \times$ the interquartile range. Outliers are marked as circles.

spectively, while stations in the other clusters are generally located below 500 m. Cluster 4 is mostly located in the central region of Turkey and cluster 7 is retrieved in the Alps and other mountainous regions mainly in Central Europe. Due to the higher altitude of the observation stations in cluster 7, the annually averaged diurnal cycle shows the lowest mean ($10.9 \mu\text{g m}^{-3}$) and variance ($(3.3 (\mu\text{g m}^{-3})^2)$) for NO₂ concentrations among all clusters. However, the NO₂ mean diurnal cycle in this cluster is similar to that of cluster 6 (mean $12.3 \mu\text{g m}^{-3}$ and variance $4.5 (\mu\text{g m}^{-3})^2$), which mostly covers Scandinavia and therefore includes mostly remote stations. The clusters 1, 2, 3, and 5 show a similar mean and variance of the annually averaged diurnal cycle for NO₂ concentrations, while clusters 4 and 8 show the highest values (mean: $37.2 \mu\text{g m}^{-3}$; $28.5 \mu\text{g m}^{-3}$; variance: $75.8 (\mu\text{g m}^{-3})^2$ $23.3 (\mu\text{g m}^{-3})^2$). This illustrates the ability of the clustering used in the KSC experiment to group stations according to the atmospheric chemical observations.

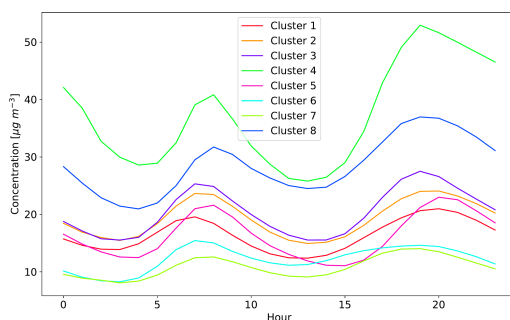


Figure 6. Annual mean diurnal cycles of the NO₂ concentration for all clusters generated in the KSC experiment. Each curve represents the annually averaged diurnal cycle for all stations within one cluster.

4.2 Evaluation of the representativity

The RMSEs for the assimilation and validation data sets for the REF, CD and KSC experiment are exemplary shown in Fig. 7 for 4 March 2016. Further, the AV-difference and the relative AV-difference is given for each experiment. The latter is calculated by dividing the AV-difference by the mean of the assimilation and validation RMSE. The REF experiment shows the lowest RMSE for the assimilation data set ($211.7 \mu\text{g m}^{-3}$) but the highest RMSE for the validation data set ($353.8 \mu\text{g m}^{-3}$). This likely indicates an overfit of the analysis to the assimilation data or, in other words, a weak representation of unobserved regions. The observation configuration used in the KSC experiment yields the lowest AV-difference of $36.5 \mu\text{g m}^{-3}$ for the CO concentration and thus provides the most representative model state. This corresponds to a relative AV-difference of 15 %. Furthermore, the figure illustrates that a representative model output is a trade-off between an increase in the RMSE for the assimilation

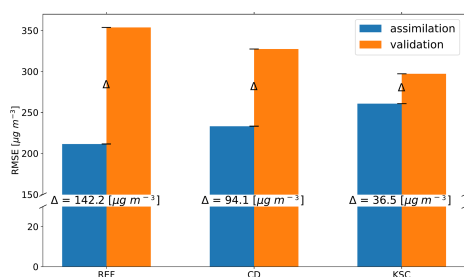


Figure 7. Comparison of the RMSE for CO of the assimilation and validation data sets for 4 March 2016, for the REF, CD and KSC experiment. The central number corresponds to the AV-difference for each experiment.

data set and a decrease in the RMSE for the validation data set.

In order to assess the representativity for the whole assimilation period, the values of the mean AV-difference are averaged over all analysed days for all months. The results for CO and NO₂ are discussed in the following.

Table 2 compares the observation configurations from the different experiments in terms of the mean RMSE for the assimilation and validation data sets and the mean AV-difference considering all simulation days. The mean AV-difference is calculated using the absolute value of the differences on each day, thus,

$$\overline{\text{RMSE}}_{\text{assim}} - \overline{\text{RMSE}}_{\text{val}} \neq \text{AV-difference}. \quad (14)$$

The evaluation of the mean AV-difference for CO shows a significant improvement in the representativity with a reduction from 87.0 μg m⁻³ (relative AV-difference of 31 %) in the REF experiment to 49.5 μg m⁻³ (18 %) and 40.6 μg m⁻³ (15 %) in the CD and the KSC experiment configurations, respectively. The observation configuration in the KSC experiment yields the largest improvement with a decrease of the AV-difference of ~53 % which corresponds to a reduction in the relative AV-difference by 16 %. This indicates a reduction of the overfitting of the analysis.

The RMSE for NO₂ is one order of magnitude smaller than for CO with a mean assimilation RMSE of 16.2 μg m⁻³ in the REF experiment. The same trend of an increasing mean RMSE for the assimilation data set and a decreasing mean RMSE for the validation data set can be seen, which leads to a reduction of the AV-difference for NO₂ of ~56 % (CD) and ~50 % (KSC). The decrease in the relative AV-difference is less substantial with a decrease of 5 % for the CD experiment and a decrease of 4 % for the KSC experiment. This is due to the generally lower RMSE for NO₂ than for CO.

The RMSE for NO₂, O₃, SO₂, PM₁₀, and PM_{2.5} are an order of magnitude smaller than those for CO, and are therefore deemed to be of lesser significance for the evaluation of the configuration in the different experiments as the abso-

lute difference is smaller. This is reflected in the relative AV-differences, which exhibit lesser changes compared to CO. Furthermore, the RMSE target reference values in the quality control of the CAMS regional services are 16 μg m⁻³ for O₃, PM₁₀ and PM_{2.5} and 22 μg m⁻³ for NO₂ (Gauss et al., 2024), making AV-differences of ~1 μg m⁻³ negligible. The corresponding values of the mean RMSEs and AV-differences for the other species are given in Table C1 in the Appendix.

Note that the sum of $\overline{\text{RMSE}}_{\text{assim}} + \overline{\text{RMSE}}_{\text{val}}$ shows no significant difference between the observation configurations. This indicates that the quality of the analysis is not affected by the different experiments. Furthermore, while the evaluation shows fluctuations for each season, the general result holds true for each season individually, see Fig. A1 in the Appendix.

4.3 Validation of the experiments for the year 2017

To evaluate the generalizability of the findings in the experiments, the generated data split for the year 2016 is applied to data from March 2017. Here, the generated assimilation-validation data split that has been derived for the 2016 data is used to divide the 2017 data. Thus, the data clustering is not influenced by the data from 2017. Stations that are not included in both data sets are excluded to ensure comparability of the results. This leads to a reduction of the number of stations of approximately 9 % of observations in the assimilation data set.

Table 3 shows the mean RMSE for assimilation and validation stations and the mean AV-difference for CO and NO₂ concentrations in March 2017 for the REF and KSC experiments. To save computing time, the CD experiment has been excluded from this validation. For CO, the KSC experiment shows an improved overall representativity of the model analyses by ~9 % for March 2017. This is likely due to the geographic information included in the KSC clustering algorithm, which is independent of the actual observations. However, this improvement is smaller than for the periods in 2016. Compared to the average AV-difference for 2016 (87.0 μg m⁻³ d⁻¹), the mean AV-difference for the REF experiment is significantly lower in 2017 (35.3 μg m⁻³ d⁻¹), leaving less potential for improvement. This is reflected in the relative AV-difference, where an insignificant change of 1 % can be observed. The reduction of the mean AV-difference between the REF and the KSC experiments for NO₂ is more similar to the reduction obtained for the 2016 periods with ~65 % reduction in March 2017 compared to ~50 % in 2016. With a reduction in the relative AV-difference by 8 %. An overview of the results for the other species is given in Table D1 of the Appendix.

The split of the available observation data derived for 2016 cannot be carried over directly to the year 2017. This is expected since clustering is inherently sensitive to the addition and removal of individual objects. Furthermore, the charac-

Table 2. AV-difference and RMSE for the assimilation and validation data sets for CO and NO₂ averaged over all simulated days in 2016 for the REF, CD, and KSC experiments.

CO	$\overline{\text{RMSE}}_{\text{assim}} [\mu\text{g m}^{-3}]$	$\overline{\text{RMSE}}_{\text{val}} [\mu\text{g m}^{-3}]$	$\overline{\text{AV-difference/day}} [\mu\text{g m}^{-3}]$	relative $\overline{\text{AV-difference}}$
REF	240.3	324.4	87.0	31 %
CD	272.8	290.6	49.5	18 %
KSC	264.1	295.4	40.6	15 %

NO ₂	$\overline{\text{RMSE}}_{\text{assim}} [\mu\text{g m}^{-3}]$	$\overline{\text{RMSE}}_{\text{val}} [\mu\text{g m}^{-3}]$	$\overline{\text{AV-difference/day}} [\mu\text{g m}^{-3}]$	relative $\overline{\text{AV-difference}}$
REF	16.2	17.8	1.6	9 %
CD	16.6	17.0	0.7	4 %
KSC	16.7	17.3	0.8	5 %

Table 3. AV-difference and RMSE for the assimilation and validation data sets for CO and NO₂ averaged for March 2017 for the REF and the KSC experiment.

CO	$\overline{\text{RMSE}}_{\text{assim}} [\mu\text{g m}^{-3}]$	$\overline{\text{RMSE}}_{\text{val}} [\mu\text{g m}^{-3}]$	$\overline{\text{AV-difference/day}} [\mu\text{g m}^{-3}]$	relative $\overline{\text{AV-difference}}$
REF	233.9	265.3	35.3	14 %
KSC	240.2	250.4	32.1	13 %

NO ₂	$\overline{\text{RMSE}}_{\text{assim}} [\mu\text{g m}^{-3}]$	$\overline{\text{RMSE}}_{\text{val}} [\mu\text{g m}^{-3}]$	$\overline{\text{AV-difference/day}} [\mu\text{g m}^{-3}]$	relative $\overline{\text{AV-difference}}$
REF	16.6	18.8	2.3	13 %
KSC	17.1	17.7	0.8	5 %

teristics of observation stations are not static, as shown in Figs. B1 and B2 in the Appendix.

5 Summary and conclusions

In this study, two methods to choose representative assimilation and validation sets are developed and tested. They utilize k-means clustering algorithms and variations thereof and only require input from observations. Thus, they are independent of the assimilation model. The methods are tested in the CD and KSC experiment, where in the former, only the diurnal cycles of the observed concentrations are considered, and the latter experiment additionally accounts for the geographical locations and altitudes of the observational sites. They are compared to a REF experiment based on an observation configuration provided within the CAMS project.

The results demonstrate that the observation configurations derived in the experiments influence the analysis in data assimilation applications, which were conducted using the EURAD-IM. The quality of the analysis obtained using a configuration is quantified by its representativity. Here, the relative representativity of two datasets is determined by quantifying and subsequently comparing the difference in the RMSEs between the model analysis and the observations from the assimilation and validation data set. Investigating each month separately confirms these results, showing that the generated configurations outperform the split in the REF

experiment. This is most notable for CO, where a limited number of observations is available. The clustering method used in the KSC experiment is capable of identifying clusters of observation stations that show different atmospheric chemical properties. Improvements in the representativity for CO (NO₂) are determined with a ~ 53 % (~ 50 %) reduction of the difference between the RMSE of the analysis compared to the assimilation and validation data set, which is regarded as the representativity measure. To account for the difference in the magnitude of the RMSE for the different species, the representativity measure is scaled with the average RMSE of the species. This yields a 16 % improvement for CO and only a 4 % improvement for NO₂. The less computationally demanding split in the CD experiment is also found to enhance the representativity, although to a lesser extend than in the KSC experiment. The clusters in the KSC experiment contain more information about the characteristics of their assigned observation stations (e.g., diurnal cycle and station altitude). Therefore, the KSC experiment is preferred for the selection of representative assimilation and validation data sets. However, the representativity for the other modelled species is not improved. This is due to the relative scarcity of the CO observations (see Fig. 1) and a resulting strong influence on the clustering. Furthermore, all species observations from a specific measurement site are categorized either into the assimilation or the validation set, called bundling. This is done to prevent advantages in the data as-

simulation due to the correlation between different species. A less restrictive bundling, as for example in the CAMS observational data set (REF experiment), could mitigate this disadvantage. In the REF experiment NO₂ is bundled with O₃ and PM₁₀ is bundled with PM_{2.5}.

The results have been obtained by considering observations for the entire year of 2016. This allows to generate one clustering result that can be applied to all analysis episodes within this year. It is assumed that a clustering using only data from the assimilation time period will further improve the representativity as the stations' characteristics are considered more accurate. However, this will require more computational resources. Nonetheless, the results clearly show that a clustering on a broader period already provides benefits in terms of representativity compared to the observation configuration in the REF experiment.

The KSC experiment shows an improved representativity (~9 % compared to the REF experiment for CO) for an observation period with data from 2017 that has been split by copying the result from 2016 without subjecting the 2017 data to the clustering algorithm. The decrease in the degree of improvement found in the comparison to the 2016 period is attributed to the fact that the diurnal cycles of the observations in 2016 are not representative for the measurements made in 2017 (see Appendix B). Thus, the clustering and split are not optimal for the assimilation period compared to a data split that considers the observations from 2017.

A distinct advantage of the proposed method to generate observation configurations is that it guarantees the availability of validation data in all regions (clusters), which is crucial for species with sparse observations. It is worth noting that a k-means procedure that is solely based on longitude, latitude, and altitude of the stations can yield a clustering configuration similar to that of the cluster derived using the k-means soft constrained algorithm. However, in the case of localized variable mean concentrations of the observed species within one region, the k-means soft constrained algorithms is of advantage as it takes this difference into account. It is also important to consider the occurrence of extreme events, which may not be adequately captured by a purely geographical k-means approach.

Further investigations are envisioned to gain a deeper understanding of the influence of individual observation stations on the performance of the data assimilation algorithm (one aspect of observability) and the inclusion of other types of observations to optimize the knowledge return from observation networks. Satellite observations are another common and valuable source of information for data assimilation systems. Extending the presented method to analyse the influence of these additional data on the representativity potentially provides insight into the optimal application of the data for assimilation.

Appendix A: Seasonality of the AV-difference

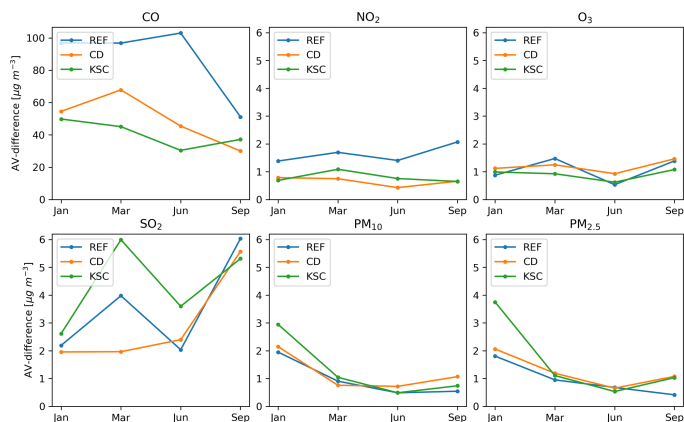


Figure A1. The evolution of the AV-difference for the REF, CD and KSC configurations for each species is shown over the four evaluated months of 2016.

Appendix B: Comparison of data from 2016 to 2017

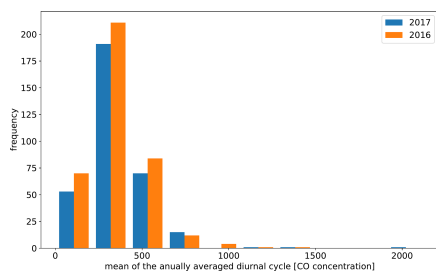


Figure B1. Histogram of the mean of the annually averaged diurnal cycles for CO observations in 2016 and 2017.

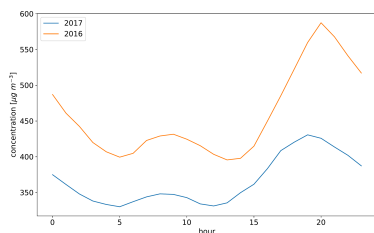


Figure B2. Annually averaged diurnal cycle for CO from the observation station "PL0575A" in 2016 and 2017.

Appendix C: Mean AV-differences in 2016 for SO₂, O₃, PM₁₀, and PM_{2.5}

Table C1. AV-difference and RMSE for the assimilation and validation data sets for SO₂, O₃, PM₁₀ and PM_{2.5} averaged over all simulated days in 2016 for the different observation configurations.

SO ₂	$\overline{\text{RMSE}}_{\text{assim}} \left[\mu\text{g m}^{-3} \right]$	$\overline{\text{RMSE}}_{\text{val}} \left[\mu\text{g m}^{-3} \right]$	$\overline{\text{AV-difference}}/\text{day} \left[\mu\text{g m}^{-3} \right]$	relative $\overline{\text{AV-difference}}$
REF	10.4	10.4	3.6	35 %
CD	10.7	8.7	3.0	31 %
KSC	9.5	12.6	4.4	40 %
O ₃	$\overline{\text{RMSE}}_{\text{assim}} \left[\mu\text{g m}^{-3} \right]$	$\overline{\text{RMSE}}_{\text{val}} \left[\mu\text{g m}^{-3} \right]$	$\overline{\text{AV-difference}}/\text{day} \left[\mu\text{g m}^{-3} \right]$	relative $\overline{\text{AV-difference}}$
REF	22.5	23.5	1.1	5 %
CD	22.2	23.4	1.2	5 %
KSC	22.7	23.6	0.9	4 %
PM ₁₀	$\overline{\text{RMSE}}_{\text{assim}} \left[\mu\text{g m}^{-3} \right]$	$\overline{\text{RMSE}}_{\text{val}} \left[\mu\text{g m}^{-3} \right]$	$\overline{\text{AV-difference}}/\text{day} \left[\mu\text{g m}^{-3} \right]$	relative $\overline{\text{AV-difference}}$
REF	14.3	14.7	1.0	7 %
CD	14.0	14.7	1.2	8 %
KSC	14.6	14.2	1.3	9 %
PM _{2.5}	$\overline{\text{RMSE}}_{\text{assim}} \left[\mu\text{g m}^{-3} \right]$	$\overline{\text{RMSE}}_{\text{val}} \left[\mu\text{g m}^{-3} \right]$	$\overline{\text{AV-difference}}/\text{day} \left[\mu\text{g m}^{-3} \right]$	relative $\overline{\text{AV-difference}}$
REF	11.0	11.3	1.0	9 %
CD	10.5	11.3	1.2	11 %
KSC	10.7	11.9	1.6	14 %

Appendix D: Mean AV-differences in 2017 for SO₂, O₃, PM₁₀, and PM_{2.5}

Table D1. AV-difference and RMSE for the assimilation and validation data sets for SO₂, O₃, PM₁₀ and PM_{2.5} averaged over all simulated days in 2017 for the REF and KSC observation configurations.

SO ₂	$\overline{\text{RMSE}}_{\text{assim}} \left[\mu\text{g m}^{-3} \right]$	$\overline{\text{RMSE}}_{\text{val}} \left[\mu\text{g m}^{-3} \right]$	$\overline{\text{AV-difference}}/\text{day} \left[\mu\text{g m}^{-3} \right]$	relative $\overline{\text{AV-difference}}$
REF	6.8	6.8	2.1	31 %
KSC	6.0	8.1	3.0	43 %
O ₃	$\overline{\text{RMSE}}_{\text{assim}} \left[\mu\text{g m}^{-3} \right]$	$\overline{\text{RMSE}}_{\text{val}} \left[\mu\text{g m}^{-3} \right]$	$\overline{\text{AV-difference}}/\text{day} \left[\mu\text{g m}^{-3} \right]$	relative $\overline{\text{AV-difference}}$
REF	20.3	21.2	0.9	4 %
KSC	20.3	21.3	1.1	5 %
PM ₁₀	$\overline{\text{RMSE}}_{\text{assim}} \left[\mu\text{g m}^{-3} \right]$	$\overline{\text{RMSE}}_{\text{val}} \left[\mu\text{g m}^{-3} \right]$	$\overline{\text{AV-difference}}/\text{day} \left[\mu\text{g m}^{-3} \right]$	relative $\overline{\text{AV-difference}}$
REF	13.2	14.0	1.0	7 %
KSC	13.1	13.4	0.8	6 %
PM _{2.5}	$\overline{\text{RMSE}}_{\text{assim}} \left[\mu\text{g m}^{-3} \right]$	$\overline{\text{RMSE}}_{\text{val}} \left[\mu\text{g m}^{-3} \right]$	$\overline{\text{AV-difference}}/\text{day} \left[\mu\text{g m}^{-3} \right]$	relative $\overline{\text{AV-difference}}$
REF	10.1	10.3	0.7	7 %
KSC	9.9	10.7	1.0	10 %

Code and data availability. The routines for the k-means soft constraint clustering, including the data preparation can be found at <https://doi.org/10.5281/zenodo.14711881> (Hermanns, 2025). Instructions for running the code are included in the readme file (Hermanns, 2025).

Author contributions. Contributed to conception and design: AH, PF, ACL. Contributed to acquisition of data: AH. Contributed to analysis and interpretation of data: AH, PF, ACL, JK, HF. Drafted and/or revised the article: AH, PF, ACL, JK, HF. Approved the submitted version for publication: AH, PF, ACL, JK, HF.

Competing interests. The contact author has declared that none of the authors has any competing interests.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications makes every effort to include appropriate place names, the final responsibility lies with the authors. Views expressed in the text are those of the authors and do not necessarily reflect the views of the publisher.

Acknowledgements. This work was partially performed as part of the Helmholtz School for Data Science in Life, Earth and Energy (HDS-LEE) and received funding from the Helmholtz Association of German Research Centres. The authors gratefully acknowledge the Earth System Modelling Project (ESM) for funding this work by providing computing time on the ESM partition of the supercomputer JUWELS (Jülich Supercomputing Centre, 2021) at the Jülich Supercomputing Centre (JSC).

Financial support. The article processing charges for this open-access publication were covered by the Forschungszentrum Jülich.

Review statement. This paper was edited by Yongze Song and reviewed by three anonymous referees.

References

- Beyer, K., Goldstein, J., Ramakrishnan, R., and Shaft, U.: When Is “Nearest Neighbor” Meaningful?, in: Database Theory – ICDT’99, edited by Beeri, C. and Buneman, P., Springer Berlin Heidelberg, Berlin, Heidelberg, 217–235, ISBN 978-3-540-49257-3, 1999.
- Borge, R., Jung, D., Lejarraaga, I., de la Paz, D., and Cordero, J. M.: Assessment of the Madrid region air quality zoning based on mesoscale modelling and k-means clustering, *Atmospheric Environment*, 287, 119258, <https://doi.org/10.1016/j.atmosenv.2022.119258>, 2022.
- Breiman, L.: Random Forests, *Machine Learning*, 45, 5–32, <https://doi.org/10.1023/A:1010933404324>, 2001.
- Carro-Calvo, L., Ordóñez, C., García-Herrera, R., and Schnell, J. L.: Spatial clustering and meteorological drivers of summer ozone in Europe, *Atmospheric Environment*, 167, 496–510, <https://doi.org/10.1016/j.atmosenv.2017.08.050>, 2017.
- Colette, A., Collin, G., Besson, F., Blot, E., Guidard, V., Meleux, F., Royer, A., Petiot, V., Miller, C., Fermond, O., Jeant, A., Adani, M., Arteta, J., Benedictow, A., Bergström, R., Bowdalo, D., Brandt, J., Briganti, G., Carvalho, A. C., Christensen, J. H., Couvidat, F., D’Elia, I., D’Isidoro, M., Denier van der Gon, H., Descombes, G., Di Tomaso, E., Douros, J., Escribano, J., Eskes, H., Fagerli, H., Fatahi, Y., Flemming, J., Friese, E., Frohn, L., Gauss, M., Geels, C., Guarnieri, G., Guevara, M., Guion, A., Guth, J., Hänninen, R., Hansen, K., Im, U., Janssen, R., Jeofrion, M., Joly, M., Jones, L., Jorba, O., Kadantsev, E., Kahnert, M., Kaminski, J. W., Kouznetsov, R., Kranenburg, R., Kuenen, J., Lange, A. C., Langner, J., Lannuque, V., Macchia, F., Manders, A., Mircea, M., Nyiri, A., Olid, M., Pérez García-Pando, C., Palamarchuk, Y., Piersanti, A., Raux, B., Razingerg, M., Robertson, L., Segers, A., Schaap, M., Siljamo, P., Simpson, D., Sofiev, M., Stangel, A., Struzewska, J., Tena, C., Timmermans, R., Tsikerdekis, T., Tsyro, S., Tyuryakov, S., Ung, A., Uppstu, A., Valdebenito, A., van Velthoven, P., Vitali, L., Ye, Z., Peuch, V.-H., and Rouil, L.: Copernicus Atmosphere Monitoring Service – Regional Air Quality Production System v1.0, *Geosci. Model Dev.*, 18, 6835–6883, <https://doi.org/10.5194/gmd-18-6835-2025>, 2025.
- De Souza Fernandes Duarte, E., Franke, P., Lange, A. C., Friese, E., Juliano da Silva Lopes, F., João da Silva, J., Souza dos Reis, J., Landolfo, E., Santos e Silva, C. M., Elbern, H., and Hoelzemann, J. J.: Evaluation of atmospheric aerosols in the metropolitan area of São Paulo simulated by the regional EURAD-IM model on high-resolution, *Atmospheric Pollution Research*, 12, 451–469, <https://doi.org/10.1016/j.apr.2020.12.006>, 2021.
- Elbern, H., Strunk, A., Schmidt, H., and Talagrand, O.: Emission rate and chemical state estimation by 4-dimensional variational inversion, *Atmos. Chem. Phys.*, 7, 3749–3769, <https://doi.org/10.5194/acp-7-3749-2007>, 2007.
- Erraji, H., Franke, P., Lampert, A., Schuldt, T., Tillmann, R., Wahner, A., and Lange, A. C.: The potential of drone observations to improve air quality predictions by 4D-Var, *Atmos. Chem. Phys.*, 24, 13913–13934, <https://doi.org/10.5194/acp-24-13913-2024>, 2024.
- Flemming, J., Stern, R., and Yamartino, R. J.: A new air quality regime classification scheme for O₃, NO₂, SO₂ and PM₁₀ observations sites, *Atmospheric Environment*, 39, <https://doi.org/10.1016/j.atmosenv.2005.06.039>, 2005.
- Forgy, E.: Cluster Analysis of Multivariate Data: Efficiency versus Interpretability of Classification, *Biometrics*, 21, 768–769, 1965.
- Fovell, R. G. and Fovell, M.-Y. C.: Climate Zones of the Conterminous United States Defined Using Cluster Analysis, *Journal of Climate*, 6, 2103–2135, <http://www.jstor.org/stable/26198599>, 1993.
- Franke, P., Lange, A. C., and Elbern, H.: Particle-filter-based volcanic ash emission inversion applied to a hypothetical sub-Plinian Eyjafjallajökull eruption using the Ensemble for Stochastic Integration of Atmospheric Simulations (ESIAS-

- chem) version 1.0, *Geosci. Model Dev.*, 15, 1037–1060, <https://doi.org/10.5194/gmd-15-1037-2022>, 2022.
- Franke, P., Lange, A. C., Steffens, B., Pozzer, A., Wahner, A., and Kiendler-Scharr, A.: European air quality in view of the WHO 2021 guideline levels: Effect of emission reductions on air pollution exposure, *Elementa: Science of the Anthropocene*, 12, 00127, <https://doi.org/10.1525/elementa.2023.00127>, 2024.
- Gauss, M., Mortier, A., Blake, L., Valdebenito, A., Negri, C., Petiot, V., Joly, M., Fermond, O., Douros, J., Tsiokerdekis, A., Eskes, H. J., Bennouna, Y., Thouret, V., Meleux, F., Raux, B., Ung, A., Sofiev, M., and Palamarchuk, J.: Documentation for the Evaluation and Quality Control (EQC) of the CAMS regional services – Status 2024, https://atmosphere.copernicus.eu/sites/default/files/custom-uploads/EQC-regional/CAMS283_2021SC2_D83.4.1.1-2024_202407_v2.pdf (last access: 10 July 2025), 2024.
- Govender, P. and Sivakumar, V.: Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019), *Atmospheric Pollution Research*, 11, 40–56, <https://doi.org/10.1016/j.apr.2019.09.009>, 2020.
- Hermanns, A.: KSC – Observational Data Clustering Preprocessor, Zenodo [code], <https://doi.org/10.5281/zenodo.14711881>, 2025.
- Ignaccolo, R., Ghigo, S., and Giovenali, E.: Analysis of air quality monitoring networks by functional clustering, *Environmetrics*, 19, 672–686, 2008.
- Inness, A., Ades, M., Agustí-Panareda, A., Barré, J., Benedictow, A., Blechschmidt, A.-M., Dominguez, J. J., Engelen, R., Eskes, H., Flemming, J., Huijnen, V., Jones, L., Kipling, Z., Massart, S., Parrington, M., Peuch, V.-H., Razinger, M., Remy, S., Schulz, M., and Suttie, M.: The CAMS reanalysis of atmospheric composition, *Atmos. Chem. Phys.*, 19, 3515–3556, <https://doi.org/10.5194/acp-19-3515-2019>, 2019.
- Joly, M. and Peuch, V.-H.: Objective classification of air quality monitoring sites over Europe, *Atmospheric Environment*, 47, 111–123, <https://doi.org/10.1016/j.atmosenv.2011.11.025>, 2012.
- Jülich Supercomputing Centre: JUWELS Cluster and Booster: Exascale Pathfinder with Modular Supercomputing Architecture at Jülich Supercomputing Centre, *Journal of Large-scale Research Facilities*, 7, <https://doi.org/10.17815/jlsrf-7-183>, 2021.
- Kracht, O., Santiago, J. L., Martin, F., Piersanti, A., Cremona, G., Righini, G., Vitali, L., Delaney, K., Basu, B., Ghosh, B., Spangl, W., Brendle, C., Latikka, J., Kousa, A., Parjala, E., Meretoja, M., Malherbe, L., Letinois, L., Beauchamp, M., Lenartz, F., Hutsemekers, V., Nguyen, L., Hoogerbrugge, R., Eneroth, K., Silvergren, S., Hooyberghs, H., Viaene, P., Maiheu, B., Janssen, S., Roet, D., and Gerboles, M.: First outcomes of the fair-mode & aquila intercomparison exercise on spatial representativeness, in: 18. International Conference on Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes (HARMO 18), edited by: Di Sabatino, S., Trini Castelli, S., and Brattini, E., Bologna, Italy, 355–359, <https://ineris.hal.science/ineris-01863209>, 2017.
- Li, J., Hendricks, J., Righi, M., and Beer, C. G.: An aerosol classification scheme for global simulations using the K-means machine learning method, *Geosci. Model Dev.*, 15, 509–533, <https://doi.org/10.5194/gmd-15-509-2022>, 2022.
- Liu, L., Hohaus, T., Franke, P., Lange, A. C., Tillmann, R., Fuchs, H., Tan, Z., Rohrer, F., Karydis, V., He, Q., Vardhan, V., Andres, S., Bohn, B., Holland, F., Winter, B., Wedel, S., Novelli, A., Hofzumahaus, A., Wahner, A., and Kiendler-Scharr, A.: Observational evidence reveals the significance of nocturnal chemistry in seasonal secondary organic aerosol formation, *npj Climate and Atmospheric Science*, 7, 1–11, <https://doi.org/10.1038/s41612-024-00747-6>, 2024.
- Lyapina, O., Schultz, M. G., and Hense, A.: Cluster analysis of European surface ozone observations for evaluation of MACC reanalysis data, *Atmos. Chem. Phys.*, 16, 6863–6881, <https://doi.org/10.5194/acp-16-6863-2016>, 2016.
- Ma, M., Chen, Y., Ding, F., Pu, Z., and Liang, X.: The representativeness of air quality monitoring sites in the urban areas of a mountainous city, *Journal of Meteorological Research*, 33, 236–250, 2019.
- MacQueen, J.: Some methods for classification and analysis of multivariate observations, in: *Proceedings of the Fifth Symposium on Math. Statistics and Probability*, University of California Press, 1, 281–297, <https://api.semanticscholar.org/CorpusID:6278891> (last access: 18 November 2025), 1967.
- Marécal, V., Peuch, V.-H., Andersson, C., Andersson, S., Arteta, J., Beekmann, M., Benedictow, A., Bergström, R., Bessagnet, B., Cansado, A., Chéroux, F., Colette, A., Coman, A., Curier, R. L., Denier van der Gon, H. A. C., Drouin, A., Elbern, H., Emili, E., Engelen, R. J., Eskes, H. J., Foret, G., Frieze, E., Gauss, M., Giannaros, C., Guth, J., Joly, M., Jaumouillé, E., Josse, B., Kadygrov, N., Kaiser, J. W., Krajsek, K., Kuenen, J., Kumar, U., Liora, N., Lopez, E., Malherbe, L., Martinez, I., Melas, D., Meleux, F., Menut, L., Moinat, P., Morales, T., Parmentier, J., Piacentini, A., Plu, M., Poupkou, A., Queguiner, S., Robertson, L., Rouil, L., Schaap, M., Segers, A., Sofiev, M., Tarasson, L., Thomas, M., Timmermans, R., Valdebenito, Á., van Velthoven, P., van Versendaal, R., Vira, J., and Ung, A.: A regional air quality forecasting system over Europe: the MACC-II daily ensemble production, *Geosci. Model Dev.*, 8, 2777–2813, <https://doi.org/10.5194/gmd-8-2777-2015>, 2015.
- Nappo, C. J., Caneill, J. Y., Furman, R. W., Gifford, F. A., Kaimal, J. C., Kramer, M. L., Lockhart, T. J., Pendergast, M. M., Pielke, R. A., Randerson, D., Shreffler, J. H., and Wyngaard, J. C.: The Workshop on the Representativeness of Meteorological Observations, June 1981, Boulder, Colo., *Bulletin of the American Meteorological Society*, 63, 761–764, <http://www.jstor.org/stable/26222836>, 1982.
- Nowak-Brzezińska, A. and Gaibei, I.: How the Outliers Influence the Quality of Clustering?, *Entropy*, 24, <https://doi.org/10.3390/e24070917>, 2022.
- Peel, M. C., Finlayson, B. L., and McMahon, T. A.: Updated world map of the Köppen-Geiger climate classification, *Hydrol. Earth Syst. Sci.*, 11, 1633–1644, <https://doi.org/10.5194/hess-11-1633-2007>, 2007.
- Peuch, V.-H., Engelen, R., Rixen, M., Dee, D., Flemming, J., Suttie, M., Ades, M., Agustí-Panareda, A., Ananasso, C., Andersson, E., Armstrong, D., Barré, J., Bousserez, N., Dominguez, J. J., Garrigues, S., Inness, A., Jones, L., Kipling, Z., Letertre-Danczak, J., Parrington, M., Razinger, M., Ribas, R., Vermoote, S., Yang, X., Simmons, A., de Marcilla, J. G., and Thépaut, J.-N.: The Copernicus Atmosphere Monitoring Service: From Research to Operations, *Bulletin of the American Meteorological Society*, 103, E2650–E2668, <https://doi.org/10.1175/BAMS-D-21-0314.1>, 2022.

- Piersanti, A., Vitali, L., Righini, G., Cremona, G., and Ciancarella, L.: Spatial representativeness of air quality monitoring stations: A grid model based approach, *Atmospheric Pollution Research*, 6, 953–960, <https://doi.org/10.1016/j.apr.2015.04.005>, 2015.
- Rousseeuw, P. J.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, *Journal of Computational and Applied Mathematics*, 20, 53–65, [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7), 1987.
- Skamarock, W., Klemp, J., Dudhia, J., Gill, D., Barker, D., Duda, M., Huang, X.-Y., Wang, W., and Powers, J.: A Description of the Advanced Research WRF Version 3, <https://doi.org/10.13140/RG.2.1.2310.6645>, 2008.
- Su, L., Gao, C., Ren, X., Zhang, F., Cao, S., Zhang, S., Chen, T., Liu, M., Ni, B., and Liu, M.: Understanding the spatial representativeness of air quality monitoring network and its application to PM_{2.5} in the mainland China, *Geoscience Frontiers*, 13, 101370, <https://doi.org/10.1016/j.gsf.2022.101370>, 2022.
- Tillmann, R., Gkatzelis, G. I., Rohrer, F., Winter, B., Wesolek, C., Schuldt, T., Lange, A. C., Franke, P., Frieze, E., Decker, M., Wegener, R., Hundt, M., Aseev, O., and Kiendler-Scharr, A.: Air quality observations onboard commercial and targeted Zeppelin flights in Germany – a platform for high-resolution trace-gas and aerosol measurements within the planetary boundary layer, *Atmos. Meas. Tech.*, 15, 3827–3842, <https://doi.org/10.5194/amt-15-3827-2022>, 2022.
- Wagstaff, K.: Clustering with Missing Values: No Imputation Required, in: *Classification, Clustering, and Data Mining Applications*, edited by: Banks, D., McMorris, F. R., Arabie, P., and Gaul, W., Springer Berlin Heidelberg, Berlin, Heidelberg, 649–658, ISBN 978-3-642-17103-1, 2004.
- Wang, X., Wang, S., Zhang, S., Gu, C., Tanvir, A., Zhang, R., and Zhou, B.: Clustering Analysis on Drivers of O₃ Diurnal Pattern and Interactions with Nighttime NO₃ and HONO, *Atmosphere*, 13, <https://doi.org/10.3390/atmos13020351>, 2022.
- World Health Organization: WHO global air quality guidelines: particulate matter (PM_{2.5} and PM₁₀), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide, World Health Organization, ISBN 9789240034228, 2021.

4. Benchmark experiment

The previous chapter presents a method for the selection of a representative subset from a given set of ground observations for data assimilation. This allows for optimal usage of the available data in an operational setting and a representative assessment of the model performance.

In order to determine the optimal placement of observations for emission rate optimization in the context of air quality modeling using data assimilation, observing system simulation experiments (OSSEs) (see Sect. 2.3) are used. These experiments are employed to quantify the impact of different observation systems on the model output. They address the spatial placement of observations, the variation of the number of observations and the choice of observation type and serve to evaluate the impact of these observational settings on the emission estimation. From these results, the importance of individual observations can be inferred and observation systems can be designed to improve the model results.

To establish a baseline for the comparison of different air quality monitoring networks, the following chapter describes the setup, execution and the results of a benchmark OSSE for the EURAD-IM. Emissions are a main driver of atmospheric pollution, yet their composition, exact amount, and their spatial- and temporal distribution is not precisely known. Thus, in an operational setting, the accuracy of the model analysis with respect to the emission rate corrections can only be quantified implicitly by the comparison of the model analysis with the assimilation and validation observations. In the benchmark OSSE, the emission rates in the nature run (introduced in the following section) are subjected to a known perturbation. The quality of the observation configuration can thus be evaluated in terms of its ability to reproduce this perturbation, based on the observations available for the 4D-var data assimilation of EURAD-IM.

During the course of the benchmark experiment, the aerosols are excluded from the optimization of the initial values and emission factors but are considered in the chemistry and transport of the forward model. This is done to exclude influences of the complex aerosol chemistry on the analysis and to focus on the gas-phase.

4.1. OSSE setup

The OSSEs are performed for the period from the 10 March to the 24 March 2016. This episode was chosen due to the low occurrence of meteorological phenomena which influ-

ence the emission-concentration relationship (e.g. washout of air pollutants by rain, strong non-linearities due to an ozone episode). The period is chosen to cover two weeks in order to limit the computational demand of the model simulations to a reasonable and feasible amount. The following section describes the realization of the different OSSE components. For an overview of the used setup of simulations and assimilated observations, see Table 4.1.

Table 4.1. Summary of the OSSE components and their realization in the benchmark OSSE. The model resolution is $15 \text{ km} \times 15 \text{ km}$ horizontally for all simulations.

OSSE component	Model	Emission factor	DA
Nature run	EURAD-IM	2	no
Control run	EURAD-IM	1	no
Assimilation run	EURAD-IM	1	yes
Observations	In situ ground, hourly concentrations CO, NO ₂ , SO ₂ , O ₃		

Nature run (NR)

The nature run for the OSSEs is generated using the forward model of the EURAD-IM with a $15 \text{ km} \times 15 \text{ km}$ horizontal resolution, while the emission rates of the chemical species CO, NO_x, SO_x, and anthropogenic NMVOCs are multiplied by a factor of 2. No data assimilation is performed during the generation of the nature run. The nature run simulation is started at the 10 March from concentration fields derived from an analysis of the whole year. By multiplying the emission rates (shown in Fig. 4.1) by a constant factor, the objective of the OSSE is to replicate these factors in the analysis of the assimilation runs. This enables the quantification of the impact of the observation system on the emission rate corrections. Furthermore, by changing the emission rates for the NR only, the identical twin problem as described in Sect. 2.3, can be avoided when using the EURAD-IM for the generation of the NR, the control run and the assimilation runs. This is explicitly evaluated in Sect. 4.2.

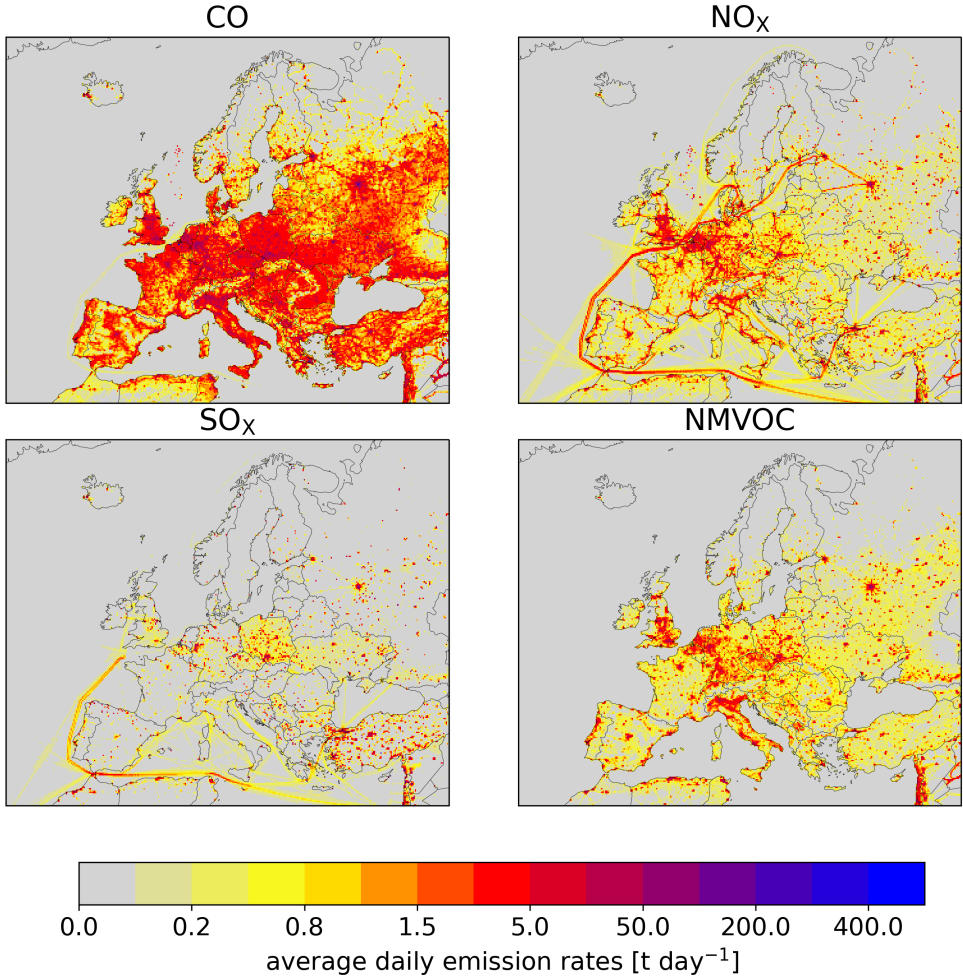


Figure 4.1. Plot of the emission inventory used in the benchmark OSSE. The emissions for CO are displayed in the top left, NO_x top right, SO_x bottom left and NMVOC bottom right. The values are shown as average daily emission over the year 2016 in t day^{-1} . The pixels correspond to the used model grid with a $15 \text{ km} \times 15 \text{ km}$ horizontal resolution. The inventory data are taken from CAMS-REG-AP (version 3.1) inventory (Kuenen et al., 2022) and supplemented with the reported emission values from Umweltbundesamt Deutschland (2019) for Germany.

Control run (CR)

The control run is generated using the same forward model of the EURAD-IM as is used in the generation of the nature run. However, the emission rates are unperturbed, i.e. no emission factor is applied to the emission inventory data. No data assimilation is performed

and the initial concentration fields are the same as for the NR.

Assimilation runs (ARs)

The assimilation runs are computed using the EURAD-IM 4D-var data assimilation scheme including the forward model and the adjoint model. Here, the emission rates are the same as for the control run. For each day, the model analysis is consecutively generated and consists of a 24 h assimilation window. The atmospheric state and emission rate corrections from the previous day's analysis are used as the initial conditions for the next day. In each assimilation run a different configuration of artificial observations (extracted from the nature run) is assimilated. Unless otherwise stated, the assimilation runs considers 15 iterations. Both, emission rate and initial value optimization are performed. The latter is included to replicate the operational usage of the model. Furthermore, differences in the initial values are encountered from the first assimilation day onward due to inaccuracies in the emission rate corrections.

Artificial observations

For the benchmark OSSEs, the artificial observations are generated as ground observations for the chemical species CO, NO₂, O₃, and SO₂. They are extracted from the nature run with an hourly resolution day by day and are designed to resemble observations from a ground-based monitoring network ensuring comparability to a real observation system. The observation operator is equal to the unit operator $H = 1$. Species and domain specific representation errors are applied to the measurements. The values are $\sqrt{5} \times 30$ ppb for CO, $\sqrt{5} \times 1.8$ ppb for NO₂, $\sqrt{5} \times 2$ ppb for O₃, and $\sqrt{5} \times 5$ ppb for SO₂, where $\sqrt{5}$ is a domain specific representation factor. No additional Gaussian perturbations, resembling measurement errors, are applied to the observations. Such an additional consideration of a measurement error was tested and showed to have only limited effect on the general conclusions of the experiments, see Appendix A.1.

The main benchmark ARs are performed with observations arranged in regular grids that align with the grid used for the model simulations. This is done to cover the model domain homogeneously and to avoid inducing biases in the evaluations by placing observations based on additional information. Furthermore, all four species are simultaneously observed at each designated observed grid cell to avoid improving the analysis due to chemical interactions and resulting correlations between observed species. In total, 6 regular grid observation configurations are generated. They are classified by a grid parameter, which indicates the separation of neighbouring observations in the x-y plane of the model grid. Thus, it is inversely related to the density of the observations. A grid with grid parameter i is constructed by placing observations in every i -th grid cell along a line of constant model y-coordinates. This is repeated in every i -th grid line in model x-direction including an offset in y-direction

of $i/2$ (rounded half of the grid parameter) grid cells for each line. The different regular grid observations configurations are shown in Fig. 4.2. For grid parameter 10, the observation grid is extended over the ocean and other water bodies. This is done to test the influence of observations over these areas. However, tests have shown that the observations over the ocean have limited influence on the assimilation result. Thus, the other observation grids are restricted to land surfaces only, to reduce the overall number of observations. For most regular grid configurations, the number of ground observations is larger than for real air quality monitoring networks. Extending the networks to cover water surfaces would exceed the memory capabilities of the supercomputing cluster. The configuration with grid parameter 1 has the maximum number of observations, with each species being observed in every model grid cell (excluding ocean cells). It serves as the theoretical upper limit for the number of observations. Table 4.2 details the number of observations and the observation densities in the regular grid configurations. The latter is calculated by comparing the number of observations against the theoretical maximum number, given by the grid parameter 1 configuration.

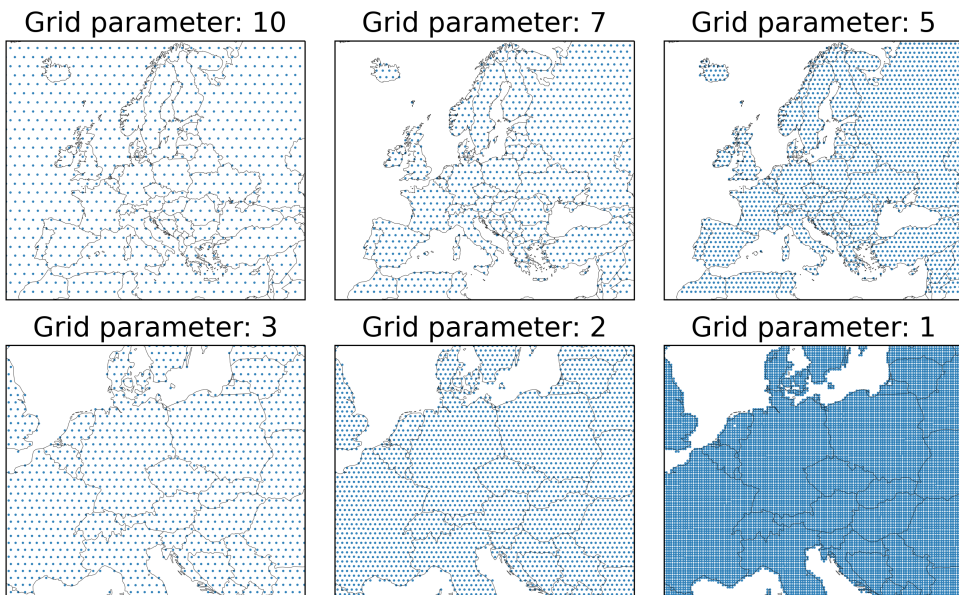


Figure 4.2. Configuration of the six regular grid observation networks. The graphics for grid parameters 10, 7, and 5 show the whole model domain. For grid parameters 3, 2, and 1, only central Europe is shown to improve readability. Each dot represents a ground-based observation site of hourly concentration measures for CO, NO₂, O₃ and SO₂.

To examine the effects of irregular grid configurations, two observation configurations based on Liu and Rabier (2002) are generated. They show that, for a 1-D data assimilation

Table 4.2. Number of observations and the observation density over the land-surface depending on the grid parameter. Each observation constitutes 24 h time series for CO, NO₂, O₃ and SO₂. For grid parameter 10, the number of observations in parentheses is the total number of observations and the plain number corresponds to the observations over land.

Grid parameter	Number of observations	Observation density [obs cell ⁻¹]
10	541 (1015)	0.010
7	1095	0.020
5	2165	0.040
3	5984	0.111
2	13330	0.249
1	53538	1.000

framework, optimal thinning of observations is achieved when the correlation between adjacent observations is less than, or equal to 0.15. In this case, optimal thinning is defined as the trade-off between a too low observation density and introducing correlations between observation errors. In the tested configurations, the observations are placed such that the correlation between the time series of neighbouring observed concentrations is less than 0.15 in the first and 0.50 in the second case. They are abbreviated as NC15 and NC50. Only observations over land surfaces are considered and the observation locations for individual species are calculated independent of each other. The NC50 observation configuration was chosen to investigate the effect of higher correlations between adjacent observations on the model analysis. Additionally, the observation configuration derived in Chapter 3 is included in the evaluation and called ‘KSC’ configuration. It serves as a comparison to an operational network. Table 4.3 contains the species specific numbers of observations and the observation densities for the irregular grid configurations. The layouts of these observation configurations are shown in Fig. 4.3. The correlation based configurations are shown for 10 March 2016. In the NC15 and NC50 configuration, the largest number of observations is present for SO₂. The concentrations of SO₂ are largely dominated by strong point sources and less by background concentrations. This causes localized concentration differences and generally reduces the correlation between neighbouring observations. The lowest number of observations is assigned for O₃, due to the strong spatial correlation of O₃ concentrations. It is mostly formed through photochemical reactions (Krupa and Manning, 1988) induced by sunlight, which exhibits strong spatial patterns and translates to the spatial concentration distribution. NO₂ concentrations are influenced by emissions, however, photochemical reactions increase spatial correlations and the diurnal cycle is dominated by human activity. This combination leads to stronger spatial correlations than for SO₂ and less observations. The number of CO observations is slightly lower than for NO₂ but larger than for O₃. This can be explained by the long lifetime of CO in the troposphere (Khalil and Rasmussen, 1990) and the relative spatial homogeneity of the emissions. Both effects

increasing the spatial correlation.

Table 4.3. Averaged number of observation locations and the density of observations over land surface for the irregular grid observation configurations, distinguished by trace gas. The averages are computed over the entire analysis episode from 10 to 24 March 2016. When applicable, the minimum to maximum number of observation are shown in brackets.

Species	Number of observations	Observation density [obs cell ⁻¹]
NC50		
CO	6251 [4580, 7143]	0.119
NO ₂	7209 [5893, 7966]	0.136
SO ₂	9356 [7656, 10436]	0.177
O ₃	4463 [3411, 5646]	0.084
NC15		
CO	2943 [2158, 3508]	0.056
NO ₂	3000 [2400, 3535]	0.057
SO ₂	4831 [3769, 5518]	0.091
O ₃	2018 [1566, 2548]	0.038
KSC		
CO	246	0.005
NO ₂	1038	0.020
SO ₂	612	0.012
O ₃	1068	0.020



Figure 4.3. Configuration of the irregular grid observation networks for the four observed species CO, NO₂, O₃ and SO₂. The grids for the correlation based networks, NC15 and NC50, are the grids for the first day of the assimilation experiment, 10 March 2016. Each dot represents an observation location, where hourly observations are extracted from the NR analogous to operational available observations.

4.2. Assessment of the nature run

According to Timmermans et al. (2015), the nature run and subsequently the artificial observations need to replicate a realistic atmosphere to ensure the validity of the OSSE results. To assess the realism of the nature run and gauge the effectiveness of the artificial observations, the CR is compared against the NR and against real ground observations. Figure 4.4 shows the spatially averaged time series of hourly concentrations for the observed species CO, NO₂, O₃, and SO₂ for the NR, the CR, and real ground-based measurements at the location of the ground measurements used in Chapter 3. Table 4.4 shows the correlation, bias and RMSE between the control run and the real observations, and between the control run and the nature run for the time series shown in Fig. 4.4.

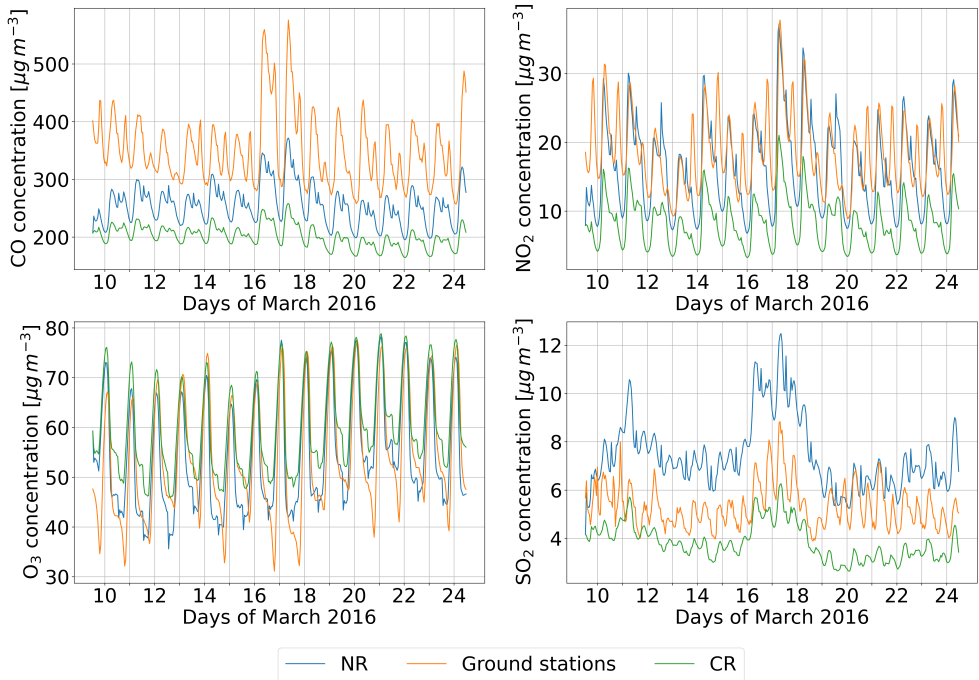


Figure 4.4. Comparison of the spatially averaged time series of hourly concentrations of the nature run (blue), control run (green), and of real observations (orange) at ground measurement sites for the experiment horizon from 10 to 24 March 2016. Each plot shows the time series of an observed species CO (top left), NO₂ (top right), O₃ (bottom left), and SO₂ (bottom right). The x-axis is marked at 12 UTC of the indicated day.

In the case of CO, the time series of the ground observations shows higher concentration values compared to the NR and the CR. The bias of the CR to the ground observations is $-148.9 \mu\text{g m}^{-3}$ (-42 %), while the bias of the CR to the NR is $-53.7 \mu\text{g m}^{-3}$ (-21 %). The concentration maxima between 16 and 18 March 2016 are captured in all three time series,

Table 4.4. Correlation coefficient, bias, and RMSE of the hourly averaged mean concentrations at ground measurement sites of the control run compared with the ground observations and the nature run. Bias and RMSE are shown in $\mu\text{g m}^{-3}$ and in %. The percentage is calculated with respect to the mean of the ground observation and NR time series, respectively.

Species	Control run to ground observations			Control run to nature run		
	Corr.	Bias	RMSE	Corr.	Bias	RMSE
CO	0.80	-148.9 ~ -42 %	156.1 ~ 45 %	0.93	-53.7 ~ -21 %	56.5 ~ 22 %
NO ₂	0.69	-10.3 ~ -55 %	11.0 ~ 59 %	0.98	-8.1 ~ -49 %	8.6 ~ 53 %
O ₃	0.85	6.2 ~ 11 %	8.7 ~ 16 %	0.99	5.7 ~ 10 %	6.3 ~ 12 %
SO ₂	0.64	-1.5 ~ -28 %	1.7 ~ 31 %	0.92	-3.6 ~ -48 %	3.7 ~ 50 %

while they are least pronounced in the CR and most noticeable for the ground observations. These maxima are caused by a low average wind speed and the resulting accumulation of air pollutants. The relatively low bias between the CR and the NR is due to the atmospheric lifetime of CO, which is estimated at an average of about 2 months with a range of 1.5 to 3 months (Khalil and Rasmussen, 1990). Thus the CO concentrations are dominated by the background concentrations and are not as strongly influenced by the shift in the emission rates, induced in the NR, as other species.

For NO₂ the bias and the RMSE between CR and real observations are comparable to the ones between the CR and NR, with a bias of ~ -55 % for the ground observation comparison and ~ -49 % compared to the NR. The RMSE shows values of ~ 59 % and ~ 53 %, respectively. The negative bias is expected for the NR, since the increase in NO_x emissions leads to an increase in the NO₂ concentrations. The time series of the ground observations shows variations in the early mornings of the simulated days. It exhibits two maxima, in the morning and afternoon, corresponding to emissions during the rush-hours. These peaks are also present in the CR and the NR, although the morning maximum concentrations are of lower magnitude than the evening spikes. The mean lifetime of atmospheric NO₂ is on the order of hours (Liu et al., 2016) and thus the average concentrations are strongly influenced by the NO_x emissions. This is a reason for the relatively large bias in the NO₂ concentrations between the NR and the CR.

In the comparison for O₃, the relative bias and RMSE are lower than for the other observed species and positive. The relative bias between CR and ground observations is ~ 11 % and between CR and NR ~ 10 %, while the relative RMSE is ~ 16 % and ~ 12 %, respectively. The low magnitude of the bias is due to O₃ being the only observed species with no

direct emissions and thus no direct increase of the model concentrations in the NR, indicating a lower sensitivity of O_3 concentrations with respect to the emissions of precursor species. However, the time evolution shows differences due to increased emissions of precursor species (Atkinson, 2000). The differences are largest during the night, while the daily maxima in the early afternoon show similar concentrations in all three time series. This nighttime difference is related to the increased NO concentrations in the NR, which reacts with O_3 to form NO_2 . During nighttime, NO_2 cannot re-form O_3 due to the missing solar radiation for photolysis. This causes an effective reduction in the nighttime O_3 concentrations. Thus, O_3 is the only evaluated species where the concentrations are reduced due to the increase in the emission rates.

In the case of SO_2 , the bias between the CR and the NR is larger than of the CR to the ground observations, with relative values of $\sim -48\%$ and $\sim -28\%$, respectively. This is due to the strong impact of the SO_x emission rate modification on the atmospheric concentrations of SO_2 . The species is mostly emitted by strong point sources and has a seasonally varying lifetime of several days (Lee et al., 2011). This allows for medium range transport and thus an increase in the overall model concentrations.

The relationship of the concentrations to the emission rate perturbations differ depending on the species (cf. Tab. 4.4). A low sensitivity of the concentrations to emission perturbations reduces the magnitude of the innovation vector and thus the effect on the emission rate correction.

Generally, the time series of the spatially averaged concentrations in the NR have a strong correlation with the time series in the CR (values from 0.92 to 0.99). This is expected, since the temporal evolution of the atmospheric state is calculated with the same model in both cases. The difference between the NR and the CR is due to the changed emission rates, which does not impact the diurnal cycles of the trace gas concentrations directly. In summary, the differences between the CR and the NR of the NO_2 and O_3 concentrations are similar to those of the available ground observations. In the case of CO, the NR concentrations show lower values than the concentrations in the real observations and in the case of SO_2 they exceed them. However, the investigated case of a general emission rate increase by a factor of 2 is not a case expected in an operational setting. A heterogeneous perturbation of the emission rates, while more realistic, might have the downside that different perturbations in neighbouring model grid cells may compensate each other, thus, reducing the impact on the concentrations in the NR. This would hinder the evaluations, since the emission rate corrections could be limited. It can be concluded that the nature run resembles sufficiently realistic observations for the proposed OSSE and capture the influence of the modified emission rates.

4.3. Analysis of the benchmark OSSEs

To evaluate the performance of the artificial observation networks in the benchmark OSSEs, the analyses of the assimilation runs are compared to the nature run. Hereby, the emission rate corrections for CO, NO_x, SO_x, and NMVOC serve as the main evaluation parameter. As detailed in Sect. 4.1, the nature run is generated by applying a constant perturbation of a factor of 2 to the emission rates. This allows for a direct evaluation of the emission rate corrections, which is not possible in an operational setting, where the true emissions are not known. Furthermore, the model concentrations of the observed species CO, NO₂, O₃, and SO₂ are used for the evaluation. This is more closely related to the evaluation of the model analysis in an operational setting. Thus, the concentrations from the ARs are compared with the NR concentrations to validate the assimilation result, as discussed in Chapter 3.

4.3.1. Emission rate correction

Formally, the emission rate perturbations are achieved using an emission factor f . Thus, the error of the emission rates, Eq. (2.11), is calculated as the difference between the emission rates of the model analysis and the expected emission rates from the NR:

$$\Delta E_p(\boldsymbol{\chi}) = E_p(\boldsymbol{\chi}) - E_{NR}(\boldsymbol{\chi}) \quad (4.1)$$

$$= f_p(\boldsymbol{\chi}) \times E^{inv}(\boldsymbol{\chi}) - f_{NR}(\boldsymbol{\chi}) \times E^{inv}(\boldsymbol{\chi}) \quad (4.2)$$

$$= (f_p(\boldsymbol{\chi}) - f_{NR}(\boldsymbol{\chi})) \times E^{inv}(\boldsymbol{\chi}) \quad (4.3)$$

$$= (f_p(\boldsymbol{\chi}) - 2) \times E^{inv}(\boldsymbol{\chi}), \quad (4.4)$$

where f_p is the corrected emission factor after the model analysis, E^{inv} is the emission inventory data and $E_p(\boldsymbol{\chi})$ and $E_{NR}(\boldsymbol{\chi})$ are the emission rates of the analysis and the nature run. $\boldsymbol{\chi}$ denotes the surface coordinates in the model space.

With this definition, the RMSE of the emission rate correction of the assimilation run can be expressed as:

$$\text{RMSE} = \sqrt{\frac{1}{N} \int \Delta E_p^2(\boldsymbol{\chi}) d\boldsymbol{\chi}}, \quad (4.5)$$

where N is the number of grid cells in the analysis output and the integration is performed over all surface model grid cells (i.e. without grid cells over the ocean). Furthermore, equation (4.1) allows for the definition of a maximum emission rate RMSE which is characterized by unchanged emission factors in the model analysis, $f_p(\boldsymbol{\chi}) = 1$ and thus, $\Delta E_{p,max}(\boldsymbol{\chi}) = E^{inv}(\boldsymbol{\chi})$. The maximum RMSE values for the different emission species are shown in Table 4.5.

Table 4.5. Values for the maximum of the emission rate RMSE for the investigated emission species in the OSSE setup.

Emission species	maximal RMSE $\text{T d}^{-1} \text{ cell}^{-1}$
CO	6.98
NO_x	1.90
SO_x	3.96
NMVOC	1.18

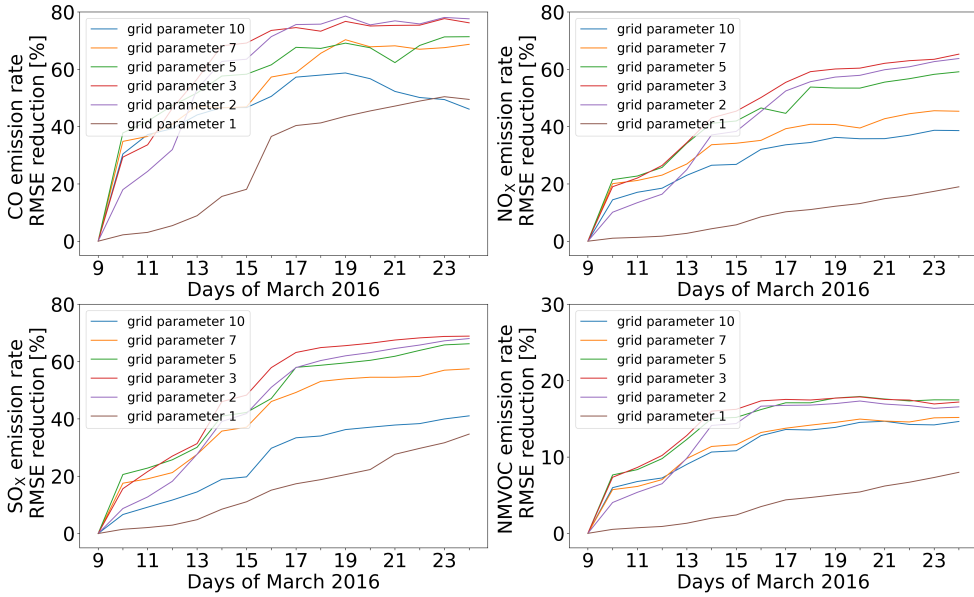


Figure 4.5. Evolution of the relative emission rate RMSE reductions in [%] of the maximum RMSE for the regular observation grid configurations from 9 to 24 March 2016. The values for the emission rates of CO (top left), NO_x (top right), SO_x (bottom left), and NMVOC (bottom right) are shown. The y-axis is limited to values less than 30 % for NMVOC and 80 % for the other species.

Figure 4.5 shows the evolution of the relative emission rate RMSE reductions for CO, NO_x , SO_x and NMVOC emissions in the regular grid configurations for the full period from 10 to 24 March 2016. The RMSEs are normalized by the maximum RMSEs for each species (cf. Tab. 4.5), which accounts for differences in the magnitude of the concentrations in the subsequent evaluations. The diagrams show that the emission rate RMSE decreases for all species during the assimilation window and approaches a configuration specific plateau. This plateau is approximately reached around 17 March (8 days after the analysis' start), and subsequent assimilation episodes yield less additional reduction of the RMSE. The highest RMSE reductions are present in the analysis of the observation configurations with grid parameter 3 and 2, with negligible differences in their final values. The relative RMSE at

the last day of the analysis, 24 March, 2016, is reduced by 76 % and 78 %, respectively, for CO, 65 % and 64 % for NO_x, 69 % and 68 % for SO_x, and 17 % each for NMVOC. The lowest RMSE reductions and thus the worst performance with respect to the emission rate corrections are calculated for the analysis of the observation configuration with grid parameter 1. It shows the least improvement per day of the analysed assimilation runs, and the lowest final emission rate RMSE reduction, except for the emission rates of CO. For NO_x the final value is 19 %, for SO_x a value of 35 %, and for NMVOC a value of 8 % is calculated. In the analysis of CO, the grid parameter 10 assimilation run initially outperforms the grid parameter 1 configuration with a higher RMSE reduction of up to 35 %. However, after 20 March its RMSE reduction decreases and falls behind that of the grid parameter 1 configuration at 23 March. The final values for CO are 49 % for grid parameter 1 and 46 % for grid parameter 10. The reason for this decrease in RMSE reduction is the increasing overestimation of strong emission sources in the grid parameter 10 assimilation run. The emission rate correction is focused around the area of the observations and in the grid parameter 10 configuration, the most relevant emission sources are not well covered. Discrepancies in the model concentrations are then adjusted by the overestimation of less impactful emission sources, instead of a lower, homogeneous adjustment. This is showcased in Appendix A.2.

At the start of the OSSE, the relative performance of the analyses using different observation configurations varies.

Figure 4.6 illustrates the relative emission rate RMSE reduction for the regular and irregular grid benchmark tests in relation to the observation density for 17 March 2016. This specific date is chosen since the emission rate corrections reach a plateau after which further assimilation days yield less additional improvement. Furthermore, in the subsequent work, a restriction of the assimilation episode allows for a substantial reduction in computational costs. The curves for the regular grid configurations show that the analysis of the assimilation runs reach higher RMSE reduction values for increasing observation densities up to a maximum, after which they decrease again. For NO_x, SO_x, and NMVOC the maxima are reached at densities of $\sim 0.1 \text{ obs cell}^{-1}$, which corresponds to the grid parameter 3 configuration. For CO the maximum is slightly shifted and occurs around $\sim 0.25 \text{ obs cell}^{-1}$, corresponding to the grid parameter 2 configuration. This shift is likely related to the lifetime of CO and the resulting greater importance of background concentrations over emission rates for the model concentration. Additionally, CO emissions are not as strongly dominated by point sources as other species. Therefore, additional observations do introduce as much additional information to the data assimilation and do not increase the condition number as much as in the case of SO_x for example, where point sources dominate. The assimilation run with the highest possible ground observation density, achieved with the observation configuration characterized by grid parameter 1, yields the lowest RMSE reduction values for the regular grid configurations with 40 % for CO, 10 % for NO_x, 17 % for SO_x, and 4 % for NMVOC. The condition number of the Hessian matrix of the cost function gradient scales

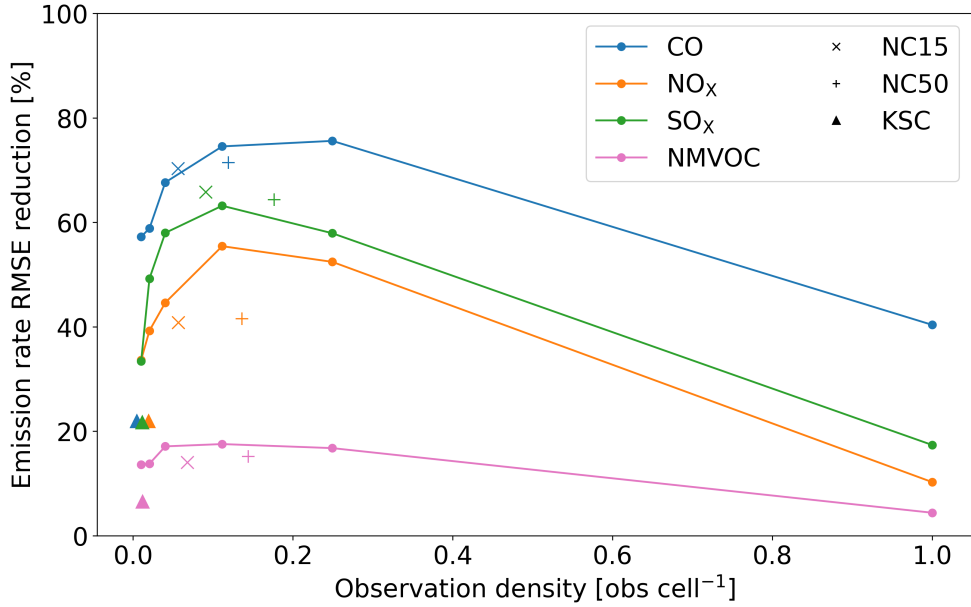


Figure 4.6. Emission rate RMSE reductions in [%] of the maximum RMSE for the benchmarking configurations at 17 March 2016 depending on the observation density for CO (blue), NO_x (orange), SO_x (green), and NMVOC (pink) according to Tab. 4.2 and Tab. 4.3. The observation densities of NO₂ and SO₂ are shown for NO_x and SO_x, respectively. Since no observations for NMVOC are used, the average observation density is shown. The RMSE of the regular grid configurations are indicated by dots, those for the correlation based configurations NC15 and NC50 by "x" and "+"-markers, respectively, and those for the KSC configuration by triangles. The observation density for the grid parameter 10 configuration is restricted to observations over land.

with the number of observations (Haben et al., 2011). Due to the extremely large number of observations in this configuration, the minimization of the cost function is of low accuracy and efficiency. For a spatial comparison of the emission rate corrections see Appendix A.3. Generally, the emission rates are overestimated in some areas, which is due to the compensation of missing observations between the observation locations and due to inaccuracies in previous assimilation days. In the grid parameter 1 configuration this effect of overestimation is nearly unobserved. However, the optimization is much slower, compared to the other configurations leading to a suboptimal result. A higher number of iterations could improve this result but would increase the computational costs.

The analysis of the KSC configuration shows similar RMSE reduction values for the emission species CO, NO_x, and SO₂ at 22 %. For NMVOC the RMSE reduction is evaluated at 6 %. These values approximately coincide with the trend of the curve, defined by the emission rate RMSE reductions of the regular grid configurations, when extended to lower densities,

considering the fluctuations in the observation density across the spatial domain.

The RMSE of the assimilation runs with correlation based irregular grid configurations show some agreement with the curve from the regular grid configurations. For CO the deviation to the curve is smallest with values of 70 % for the NC15 observation configuration and 71 % for the NC50 configuration. In the case of NO_x the configurations yield lower RMSE reduction values than the regular configuration of similar density, 41 % and 42 %, respectively. This is related to the high observation density in high emission areas. Due to the correlation based determination of the observation locations, those regions are more densely observed, leading to a local configuration similar to grid parameter 1, which hinders the optimization. For SO_x they achieve higher RMSE reductions than any assimilation run with a regular grid observation configuration, regardless of the observation density. Reaching values of 66 % for the NC15 configuration and 64 % for the NC50 configuration. In the case of NMVOC the relative RMSE reductions are 14 % and 15 %, respectively. However, as shown in Fig. 4.3, the density of the irregular grid configurations shows large spatial differences, which diminish the comparability based on the ground observation density.

In general, the correction of the emission rates of NMVOC emissions are limited due to the absence of direct NMVOC observations in the data assimilation. The emission rate corrections for NO_x and SO_x show better results, since they are closely related to the observed species NO₂ and SO₂, respectively, while for CO emissions, direct observations of the emitted species are assimilated.

4.3.2. Trace gas concentration

Figure 4.7 shows the evolution of the RMSE reduction (relative to the maximum RMSE as determined by the difference between the NR and CR) of the concentrations for CO (top left), NO₂ (top right), SO₂ (center left), O₃ (center right), and NMVOC (bottom left). Although NMVOC concentrations are not directly observed, the comparison of the RMSE reduction is included for completeness. As in the analysis of the emission rate RMSEs, the RMSEs for the concentration are reduced during the analysis for all species. In the case of O₃, an initial decrease in the RMSE reduction after the first day is calculated (-28 % to -18 %), which is increased during the next days. This is due to the low influence of the emission rate perturbation on the O₃ concentrations (cf. Tab. 4.4). The innovation vector for O₃ is smaller than for the other species, which are thus of more importance for the minimization. The reductions of the concentration RMSE for NO₂, SO₂, and NMVOC approach an observation configuration specific plateau after approximately 7 to 10 days of data assimilation. However, the RMSEs show fluctuation after this initial assimilation period. In the case of CO and O₃ concentrations, the plateau is approximately reached after 3 days, which is earlier than for other species. This is due to the initial overestimation of the emission rates, which is reduced in later days. Since the CO model concentrations are less

dependent on the emissions than other species, this has less influence on the concentrations. In the case of O_3 this earlier plateau is caused by the low emission rate dependence of the model concentrations. Thus further emission rate corrections for other species have limited effect on the O_3 concentrations and the initial value correction is sufficient to compensate the residual error. Overall, the different observation configurations show only negligible influence on the RMSE for O_3 concentrations except for the grid parameter 1 configuration. At 24 March, the final RMSE reductions are spread over values of 39 % to 32 %, with grid parameter 1 as outlier at 24 %.

For CO, NO_2 , and SO_2 the highest RMSE reductions are calculated for the grid parameter 3 and 2 assimilation runs consistent with the RMSE reduction of the emission rates. Final values of 66 % and 64 % are reached for CO, 72 % and 70 % for NO_2 , 68 % each, for SO_2 , and 19 % and 18 % for NMVOC.

In contrast to the emission rate optimization, the RMSE reduction of the grid parameter 1 assimilation run is not the lowest among the different configurations for SO_2 with a final value of 41 %. Instead, the grid parameter 10 configuration shows the lowest RMSE reduction with 36 %. This is due to the grid parameter 10 configuration covering less of the model domain, thus insufficiently covering the elevated concentrations released by emission point sources, which leads to an insufficient initial value optimization. In the case of CO, the lowest final reduction is also shown by the grid parameter 10 configuration, with 55 %, which coincides with the emission rate RMSE reduction (cf. Fig. 4.5). For NO_2 and NMVOC the lowest reduction value is calculated for the grid parameter 1 configuration assimilation run, with final values of 40 %, and 8 % respectively.

Among the species, for which the corresponding emission rates in the NR have been changed, the RMSE of the concentrations for CO and NMVOC show the least difference depending on the observation configurations. The RMSE reduction values at 24 March, 2016, differ only by approximately 11 %. For NO_2 and SO_2 the RMSE reductions vary by more than ~ 32 %. Further, the difference in the RMSE reductions of the O_3 concentrations for the different observation configurations is 15 %. This can be explained by the different lifetime of the trace gasses. CO persists in the troposphere for several months (Khalil and Rasmussen, 1990) and thus, the perturbed emission rates show a low signal compared to background concentrations. Furthermore, as shown in Tab. 4.4, the relative RMSE of the hourly averaged concentrations of CO between the CR and the NR is lower than for NO_2 and SO_2 , with ~ 22 % for CO and ~ 53 % for NO_2 and ~ 50 % for SO_2 . This shows the lower sensitivity of the CO concentration in the NR to the change in emission rates. SO_2 has lifetimes ranging from several hours to days, depending on the season (Lee et al., 2011), but the change in the atmospheric concentrations are strongly influenced by the change in the emission rates, as shown in Fig. 4.4. This is due to strong point sources being the main emission sources. Thus, the large differences in the emission rate optimization for SO_x between the assimilation runs correlate with the differences in the analysis concentrations of

SO₂.

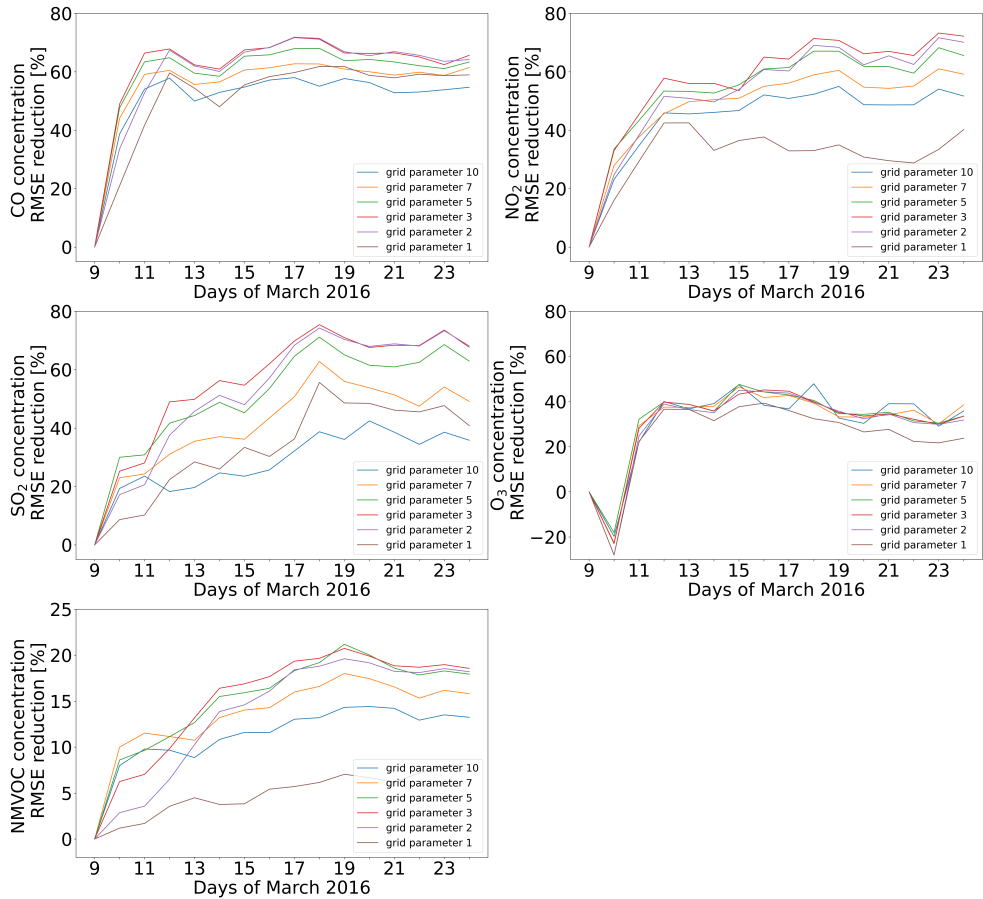


Figure 4.7. Evolution of the concentration RMSE reductions of CO (top left), NO₂ (top right), SO₂ (center left), O₃ (center right), and NMVOC (bottom left) relative to the maximum RMSE for the regular grid benchmark configurations from 9 to 24 March 2016. The y-axes are limited to values less than 80 %, except for NMVOC, where the upper limit is decreased to 25 %. For O₃, the lower limit is extended to -20 %.

Figure 4.8 illustrates the reductions of the concentration RMSE for CO, NO₂, SO₂, NMVOC, and O₃ of the regular and irregular grid benchmark tests in relation to the observation density for 17 March 2016. Similarly to the emission rate optimization, the regular grid configurations lead to an enhanced reduction of the trace gas concentration RMSE with increasing observation density, up to a maximum, after which the performance of the AR show a declining trend. For NO₂, SO₂, NMVOC, and O₃ this maximum is at $\sim 0.1 \text{ obs cell}^{-1}$, which corresponds to the observation configuration with grid parameter 3. For CO, the maximum is shifted to a higher observation density of $\sim 0.25 \text{ obs cell}^{-1}$, corresponding to grid

parameter 2. Similarly as for the emission rate corrections (see Fig. 4.6), the relatively homogeneous concentration field leads to a higher correlation between adjacent observations. Thus, the increase in the condition number is not as strong as for other species, if observations are added to the assimilation. The RMSE reductions at these maxima are 72 % for CO, 64 % for NO₂, 70 % for SO₂, 19 % for NMVOC, and 45 % for O₃. The lowest RMSE reductions are as follows: for CO at 58 % (grid parameter 10), for NO₂ at 33 % (grid parameter 1), for SO₂ at 32 % (grid parameter 10), for NMVOC at 6 % (grid parameter 1), and for O₃ at 36 % (grid parameter 1). The RMSE reductions of the NMVOC and O₃ concentrations shows a low dependency on the observation density.

The results of the assimilation runs with the correlation based grid configurations show lower RMSE reductions than the regular grid configurations with similar observation densities, except for O₃, where the values are comparable. For the KSC configuration, the concentration RMSE reductions of CO, NO₂, SO₂, and NMVOC show good agreement with the curve from the regular grid observations. In the case of O₃, the results outperform the all other configurations with a RMSE reduction of 48 %. A potential reason is the low performance in other species, which allows for more optimization potential for O₃.

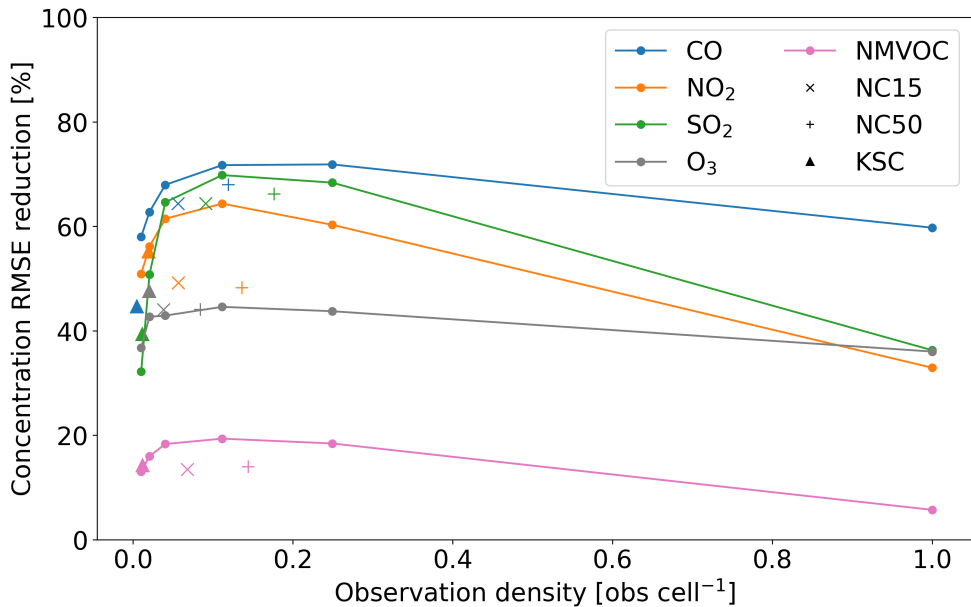


Figure 4.8. Reductions of the concentration RMSEs relative to the maximum RMSEs for the benchmarking configurations at 17 March 2016 depending on the observation density for CO (blue), NO₂ (orange), SO₂ (green), NMVOC (pink), and O₃ (grey). The RMSE values of the regular grid configurations are indicated by dots, those for the correlation based configurations NC15 and NC50 by "x" and "+"-markers, respectively, and those for the KSC configuration by triangles. The observation density for the grid parameter 10 configuration is restricted to observations over land. For NO_x and SO_x the observation densities of NO₂ and SO₂ respectively are shown. Since no observations for NMVOC are used, the average observation density is shown.

4.4. Conclusion of the benchmark experiment

In order to serve the development of observation configurations, the data from the benchmark runs are to be published, awaiting approval. Thus, they can serve as benchmark for further assessments and help to overcome the issue of having the same model for the NR and the CR/AR. It is shown that an optimal amount of observations exists, which is lower than the theoretical maximum. This is consistent with the results from Chen et al. (2025), where observation networks with high observation density showed suboptimal performance. Thus, either supplementation of existing observation configurations or thinning of observations (especially for satellite observations) is necessary to achieve optimal results. Observation networks based on the correlation of observed time series between adjacent observations (Liu and Rabier, 2002) show mixed performances depending on the investigated chemical species. They can create observation networks that are too dense around emission sources,

hindering their optimization. Further restrictions to the observation density and improvements to the selection criteria are thus needed to account for the differences in the species characteristics.

It is important to point out that the lowest possible errors are not necessarily achieved by the observation configurations tested in this benchmark. While the grid parameter 2 and 3 configuration generally result in the lowest RMSEs for the emission rate corrections and the concentrations, their high observation density for the entire model domain introduces observations of low impact on the optimization. This can decrease the effectiveness of the 4D-var minimization. Furthermore, it is not guaranteed that all emission sources require such a high observation density to be adequately captured.

The analysed RMSEs for the realistic observation configuration (KSC) can be used as indicators for the performance of the EURAD-IM in a real use case. This is of special interest for the emission rate corrections, since validation with observations is not possible, as no direct emission observations are available. However, the high spatial variance of the observation density in the KSC configuration creates spatial variance in the accuracy of the corrections which needs to be considered in the evaluation. A direct interpretation of the benchmark result as ‘model error’ is thus not feasible. The comparison with the KSC configuration shows that the emission rate and concentration optimization is in agreement with the observation density based evaluation of the benchmark tests. However, in accordance with the low observation density of the KSC configurations, the residual RMSE in the optimization shows great potential for improvement. In a non idealized application, or with real observation data, the inclusion of observations errors can lead to a shift in the optimal observation density. As demonstrated with a Gaussian perturbation of the observation data (cf. Appendix A.1), the optimization requires additional assimilation days to achieve comparable results. Furthermore, the introduced heterogeneity of the perturbed observations is likely to degrade the analysis for higher observation densities.

The next chapter extends this OSSE by proposing a methodology to generate an optimal observation configuration from all available data by including the results from the benchmark experiment.

5. Improving observational impact

The aim of this chapter is to establish and test a method to generate an observation configuration by determining, a priori, the observation locations that will have the largest impact on the emission rate optimization. To this end, the adjoint sensitivity field, based on the adjoint model evolution is used. It quantifies the area of influence of a single observation. From this, an ‘observation impact’ measure is postulated and an algorithm to generate a targeted observation configuration is developed. The method based on this algorithm is first tested in an OSSE akin to the benchmark OSSE in Chapt. 4, where the observation configuration is comprised solely of ground-based observations. Then the applicability to vertically integrated satellite observations is tested.

5.1. Surrogate model

To examine the impact of an observation on the emission rate correction, without the need to calculate a full model analysis and to save compute costs, a surrogate model is developed. As described in Sect. 2.2, the gradient of the cost function with respect to the logarithmic emission factor $\mathbf{u}$, the lower component of Eq. (2.14), is indicative of the change in the emission rate corrections. However, the gradient cannot be calculated prior to the data assimilation and can thus not be used for the intended purpose. The goal of this surrogate model is the approximation of said gradient component:

$$\nabla_{\mathbf{u}} J_{obs} = - \sum_{j=0}^N \tilde{M}_{j,ef}^T \mathbf{H}_j^T \mathbf{R}^{-1} [\mathbf{y}_j - G_j(\mathbf{x}_0)]. \quad (5.1)$$

The first component $\tilde{M}_{j,ef}^T$ quantifies the spatial influence of an assimilated observation on the emission rate correction. This is approximated by adjoint sensitivity fields, which are calculated for each model cell.

These sensitivity fields are calculated using the adjoint transport modules of the EURAD-IM. When focusing on a single model grid cell, the procedure is described as follows: For each observation location (labeled $\mathbf{x}_0$) a normalized signal is released at every hour. This can be expressed as:

$$\mathbf{H}_j^T \mathbf{R}^{-1} [\mathbf{y}_j - G_j(\mathbf{x}_0)] := \mathbf{1} \quad (5.2)$$

The released signal is propagated by the adjoint model operator pertaining to the trans-

port $\tilde{\mathbf{M}}_{j,tr}^T$ ‘backward in time’ to the initial time $t = t_0$. The individual sensitivity fields are independent from each other, thus, they do not interact.

The resulting sensitivity field $S(\boldsymbol{\chi}_0, \xi)$ describes the area to which an observation at grid cell $\boldsymbol{\chi}_0$ is sensitive to, within the specified assimilation window. Thus, $S(\boldsymbol{\chi}_0, \xi)$ defines the grid cells at coordinate ξ , where emission rate changes impact the model concentrations at the observation location $\boldsymbol{\chi}_0$. Herein, the strength of the signal defines the relative importance of the individual grid cell. This procedure is repeated for all model cells to generate the full adjoint sensitivity field for all possible observation locations. To ensure the comparability, the field is normalized according to

$$\bar{S}(\boldsymbol{\chi}, \xi) = S(\boldsymbol{\chi}, \xi) \left[\int S(\boldsymbol{\chi}, \eta) d\eta \right]^{-1}. \quad (5.3)$$

Figure 5.1 shows an example of a normalized sensitivity field for an observation location in Germany for 10 March 2016 including wind barbs. The sensitivity field mostly extends in an eastern direction with a maximum extension of approximately 20 grid cells, corresponding to 300 km. The direction of the extension is caused by the adjoint transport model. Since the location of the observation was dominated by north-easterly winds on 10 March 2016, the sensitivity fields extend mostly towards the North-East (i.e. upwind). The largest sensitivity value of 0.13 is found in the grid cell East of the observation location, while the observation point itself shows the second-highest value of 0.11.

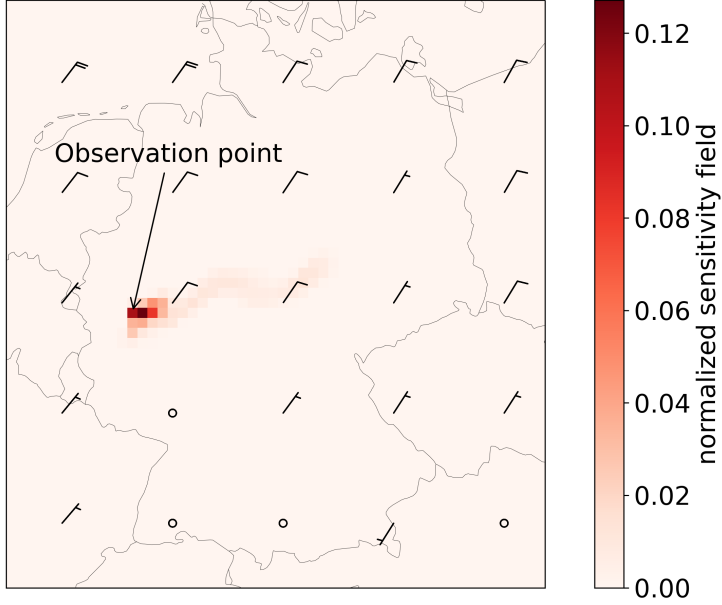


Figure 5.1. Example of a normalized sensitivity field for an observation location in Germany for 10 March 2016. The wind barbs show the mean wind direction and speed at the ground level, where each half flag depicts 5 kn and each full flag 10 kn. A circle indicates winds below 5 kn.

As the assimilation system optimizes emission rates by applying a factor to the background emission field, the gradient of the cost function is more sensitive to strong emission sources (i.e. multiplying a strong emission source by a factor has more impact on related concentrations than multiplying a small emission source by the same factor). To account for this influence of the background emission rates on the optimization, the adjoint sensitivity field is multiplied with the emission inventory. This yields:

$$S(\boldsymbol{\chi}, \boldsymbol{\xi}) \times E^{\text{inv}}(\boldsymbol{\xi}) = \sum_{j=0}^N \tilde{M}_{j,ef}^T, \quad (5.4)$$

where $\tilde{M}_{j,ef}^T$ is the tangent linear adjoint model operator pertaining to the emissions and N is the total number of model time steps.

For the second constituent of Eq. (5.1), the relation

$$BG_{RMSE}(\boldsymbol{\chi}) \propto \sum_{j=0}^N [\mathbf{y}_j - G_j(\mathbf{x}_0)] \quad (5.5)$$

is proposed here. $BG_{RMSE}(\boldsymbol{\chi})$ is the RMSE of an observation at location $\boldsymbol{\chi}$ to the background model concentration (here, the control run concentration). The right hand side of Eq. (5.5)

is the forcing term of the observations y_i with respect to the model state. This is approximated as the discrepancy between the observation and the background state of the model. The RMSE is chosen as error metric. It is noted that other error measures as well as other background model states can be chosen. Choosing the concentration from the control run enables the calculation of the sensitivity field prior to the assimilation.

To quantify the impact of an observation the components of the surrogate model are combined to:

$$O(\boldsymbol{\chi}) = \int \bar{S}(\boldsymbol{\chi}, \boldsymbol{\xi}) \times E^{inv}(\boldsymbol{\xi}) \times BG_{RMSE}(\boldsymbol{\chi}) d\boldsymbol{\xi}. \quad (5.6)$$

The quantity $O(\boldsymbol{\chi})$ is henceforth referred to as the ‘observation impact’.

The integration over the surrogate model quantifies the emission rates impacted by an observation at coordinate $\boldsymbol{\chi}$. The inclusion of BG_{RMSE} scales the potential impact of the observations by its departure from the model state. A low value for BG_{RMSE} indicates that the information provided by the observation is already incorporated in the model. Thus, a high value of $O(\boldsymbol{\chi})$ indicates a strong influence of the observation at grid coordinate $\boldsymbol{\chi}$ on the gradient of the cost function with respect to the emission rate, Eq. (5.1).

It is important to point out that the absolute value of $O(\boldsymbol{\chi})$ is only meaningful in the comparison of observations of the same kind and for the same species, since Eq. (5.6) does not account for different measurement units of different observing systems (e.g. satellite column observations vs. ground measurements). Further, different species show large differences in emission strength, which highly influences $O(\boldsymbol{\chi})$ via the term E^{inv} .

5.2. Algorithm for the generation of targeted observation configurations

To generate a targeted observation configuration based on the surrogate model and the resulting observation impact, Eq. (5.6), an algorithm is developed. It is outlined in Tab. 5.1 and combines the observation impact with the results from Chap. 4 by ensuring that the maximum observation density does not exceed the optimal value.

The desired number of observations is user specified before the execution of the algorithm. First, in step 1, the observation impact according to Eq. (5.6) is calculated for all model cells. In step 2, the observation with the largest impact value is selected for the observation network and excluded for the following iterations. Step 3 ensures that the highest observation density is in the range of the optimal observation density as indicated by the results from Chapt. 4. This is achieved by forming a 5 by 5 grid cell box, with the selected observation in the center. The best performances in the regular grid experiments are achieved by the observation configurations using grid parameter 3 and 2, which corresponds to a 2 to 1 grid

Table 5.1. Description of the algorithm to generate a targeted observation configuration based on the observation impact of individual observations.

-
- | | |
|----|--|
| 1. | Calculate the observation impact for all available observations |
| 2. | Add the observation with the largest impact value to the observation network |
| 3. | Exclude a 5 by 5 model grid cell box around the chosen observation |
| 4. | Iterate step 2 and 3 until 80% of the desired observations are chosen |
| 5. | Find the center of the largest area without observations |
| 6. | Add the observation with the largest impact value in a 10 model grid cell radius around that center to the observation network |
| 7. | Iterate step 5 and 6 until the number of desired observations is reached |
-

cell separation between adjacent assimilated observations. Therefore, the maximum local observation density is in accordance with the best performing benchmark configurations. Since the observation impact scales with the emission inventory data, areas with generally lower emission rates are underrepresented in steps 2 and 3 of the algorithm. Thus, step 4 limits the determination of observation locations by the highest residual observation impact values to 80 % of the desired number of observations. The remaining 20 % of observations are allocated to areas exhibiting the lowest observation density. In step 5, the center of the largest area without observations is identified. In step 6 of the algorithm, an area with a radius of 10 model grid cells around the center is defined. The observation with the largest observation impact within this area is added to the observation network. This selection criterion is introduced to ensure that the chosen observation provides meaningful information to the system while the sparsity of the network is reduced as much as possible. However, the choice of the radius can be user determined. In the final step 7, the targeted observation network is completed by iterating steps 5 and 6 until the desired number of observations is reached. The introduction of the restriction in step 4 and the subsequent algorithm steps ensure that information pertaining to the entire model domain is available for the process of data assimilation. This is relevant for the representativity of the model result, as demonstrated in Chapt. 3. In practice, the algorithm is executed for each observed species individually, leading to species specific observation networks. When different observation instruments are combined, the algorithm needs to be executed for each instrument separately and consecutively. To ensure that the targeted observational network is not too dense, the defined exclusion zone around observations (step 3 of the algorithm) should be applied to the existing observation configuration before adding further observations from other instruments. The separate execution is necessary, since the absolute value of the observation impact is not comparable for different instruments.

5.3. Analysis of a targeted ground observation network

To analyse the performance of a targeted observation configuration generated by the algorithm outlined in Tab. 5.1, an OSSE is performed. The test configuration is called OI (observation impact) henceforth. Here, the OSSE configuration from Chap. 4 is used to ensure comparability to the benchmark experiment. The nature run and the model used for the assimilation run are identical to the previous experiments, while the experiment period is reduced to 8 assimilation days, from 10 March to 17 March 2016. This is done to reduce the computational cost of performing the data assimilation. Sect. 4.3 shows that the largest improvement occurs within this initial experiment period and no significant changes in the RMSE reductions occur after 8 days of data assimilation. The location of the virtual observations are determined using the algorithm described in Tab. 5.1 with a total of 541 observations per species over land. This number is chosen to be equal to the observations in the grid parameter 10 benchmark configuration and comparable to the KSC configuration (cf. Tab. 4.3, 741 observations on average). Furthermore, the observations over the ocean surfaces from the grid parameter 10 configuration are added for comparability. Thus, the configuration consists of a total of 1015 observed 24 hours time series per species. In contrast to the regular grid benchmark configurations, the locations of the different trace gas observations are chosen individually for each species. All time series consist of hourly ground observations for 24 hours of the chemical species CO, NO₂, O₃ and SO₂. Their locations for 10 March 2016 are shown in Fig. 5.2. For O₃, no emission data is given and Eq. (5.6) is modified by setting $E^{inv}(\xi) = 1$. In the case of NO₂ and SO₂ observations, the emission data from NO_x and SO_x are used, respectively. For CO, NO₂, and O₃ the observations are mostly clustered around central Europe and the area near Moscow. For SO₂, the areas of Poland, Czech Republic, Southern Ukraine, and Turkey are most densely observed. This spatial focus is due to the relative strength of the emission inventory data in those areas and the resulting high observational impact values.

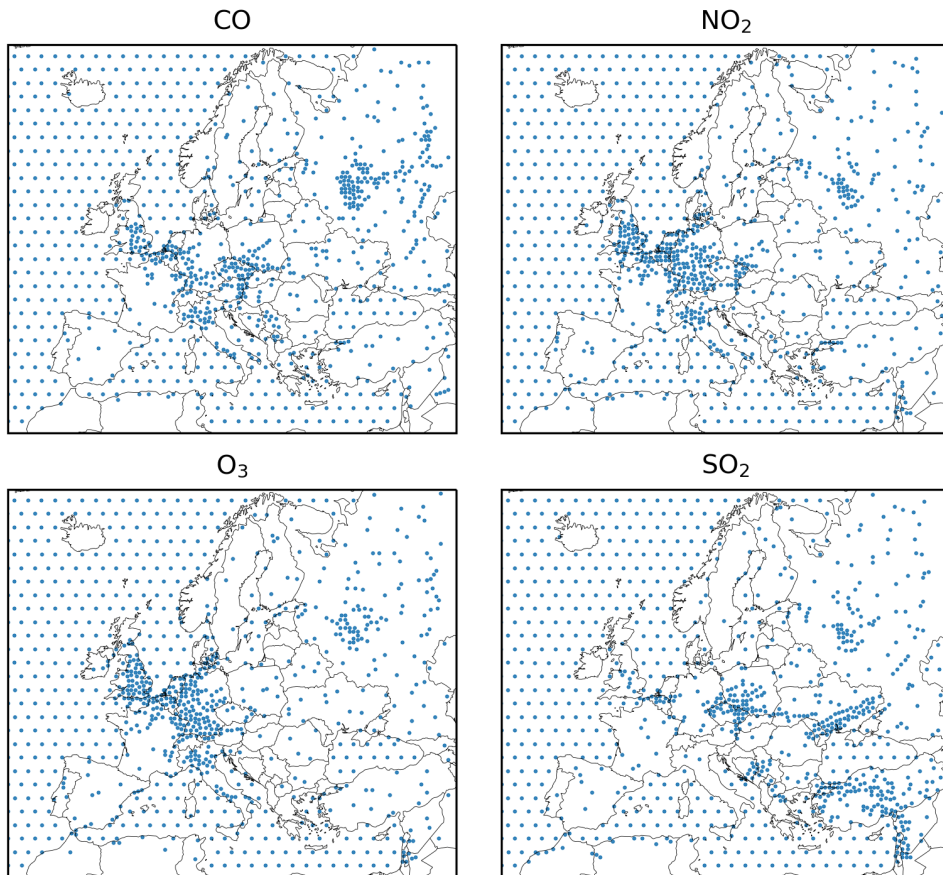


Figure 5.2. Observation configurations for CO, NO₂, O₃, and SO₂ generated using the observation impact algorithm on 10 March 2016. Each dot represents a time series of hourly ground-based concentration values.

5.3.1. Analysis of the OSSE result

The emission rate corrections of the assimilation run with the observation impact configuration are analysed by examining $\Delta E_p(\chi)$ as defined in Eq. (4.1) and the corresponding RMSE, Eq. (4.5). Figure 5.3 shows the evolution of the emission rate RMSE reductions for CO, NO_x, SO_x, and NMVOC for the assimilation period from 10 to 17 March 2016. The RMSEs are normalized by the maximum RMSEs for each species (cf. Tab. 4.5). The results from the assimilation run with the observation impact configuration are compared to the results of the regular grid benchmark configurations (dotted lines) as described in Chapt. 4. The emission optimization of the assimilation run with the observation impact configuration reaches higher RMSE reductions than the regular grid benchmark assimilation runs for most of the assimilation period. The final value of the emission rate RMSE reduction for

CO is 74 %, while the grid parameter 2 assimilation run reaches 76 %. In the case of NO_x , the final RMSE percentage is calculated at 55 %, which is equal to the RMSE reduction of the grid parameter 3 configuration which yields the largest reduction. For SO_2 , the assimilation run with the observation impact configuration reaches a RMSE reduction of 69 %, outperforming the 63 % of the grid parameter 3 assimilation run. The RMSE reduction for NMVOC is 16 %, which is slightly lower than the 18 % calculated for the grid parameter 3 configuration.

Overall, the observation impact assimilation run is able to optimize the emission rates within less analysis days than the benchmark ARs and outperforms them especially during the first analysis days. After 2 days, the calculated emission rate RMSE reduction for CO is 11 % higher than for the best regular grid benchmark AR. Furthermore, a plateau is reached at 14 March, after 5 days. This difference is most prominently observed in the analysis of the SO_x emission rates. After 2 days, a difference of 28 % is calculated and further episodes show a steady increase of the emission rate RMSE reduction. However, a plateau is not reached within the experiment period. In the case of NO_x and NMVOC, the RMSE reduction of the observation impact AR is 9 % and 2 % higher than the benchmark ARs, respectively.

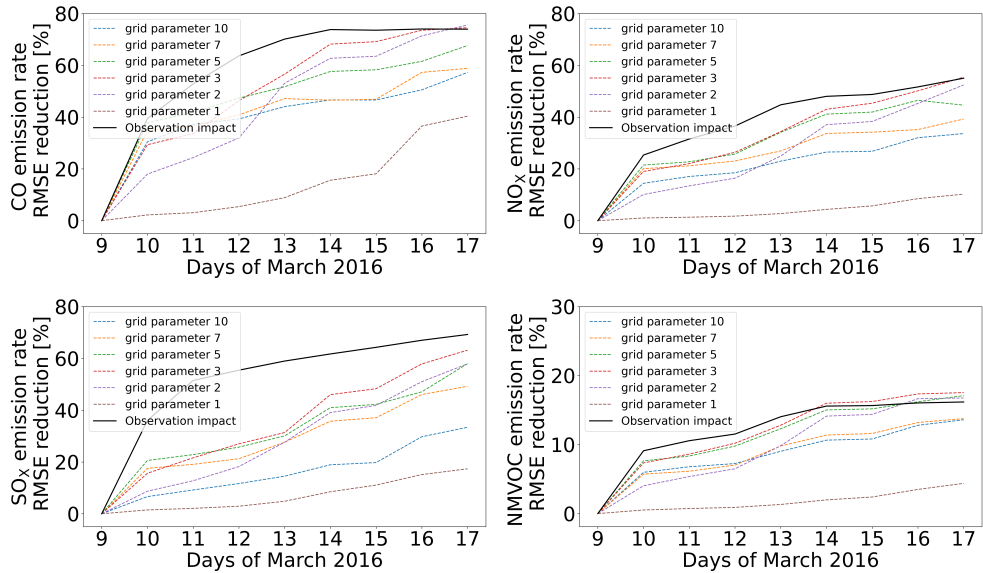


Figure 5.3. Evolution of the emission rate RMSE reductions in [%] of the maximum RMSE for the observation configuration derived using the observation impact algorithm (Eq. 5.6) from 9 to 17 March 2016. The values for the emission rates of CO (top left), NO_x (top right), SO_x (bottom left), and NMVOC (bottom right) are shown. The y-axis is limited to values less than 30 % for NMVOC, and 80 % for other species. The evolution of the regular grid configurations is shown as dashed lines for comparison. The observation impact configuration results are depicted as solid black line.

Figure 5.4 shows the emission rate RMSE reductions in dependence of the observation densities for the corresponding observation configurations for 17 March 2016. Since the observation impact configuration is designed to include the same number of observations as the regular grid configuration with grid parameter 10, the observation density is equal when averaged over the entire domain. However, the observation density shows strong local variations (cf. Fig. 5.2). In contrast to the irregular grid benchmark configurations, the emission rate RMSE reductions of the observation impact AR do not follow the curve described by the regular grid benchmark configurations. They match or, in the case of SO_x , outperform the highest RMSE reductions of the benchmark assimilation runs, despite containing up to 25 times less observations.

This general improvement of the observation impact configuration over the benchmark configurations has 3 main reasons: Firstly, the observations have an accurate coverage of the emission sources. This mostly benefits SO_x and NO_x , where strong localized emission sources dominate the emission rates. The regular grid benchmark configurations require a high number of observations to adequately cover them. Secondly, the background error influences the choice of observations. This enables the configuration to better capture differences in the background concentrations, benefiting the CO results. Thirdly, the efficient placement of the observations ensures the optimization to be effective. As demonstrated in Chapt. 4, high observation densities can degrade the effectiveness of the data assimilation. Thus, including the most relevant observation locations, with a low total number of observations, is beneficial.

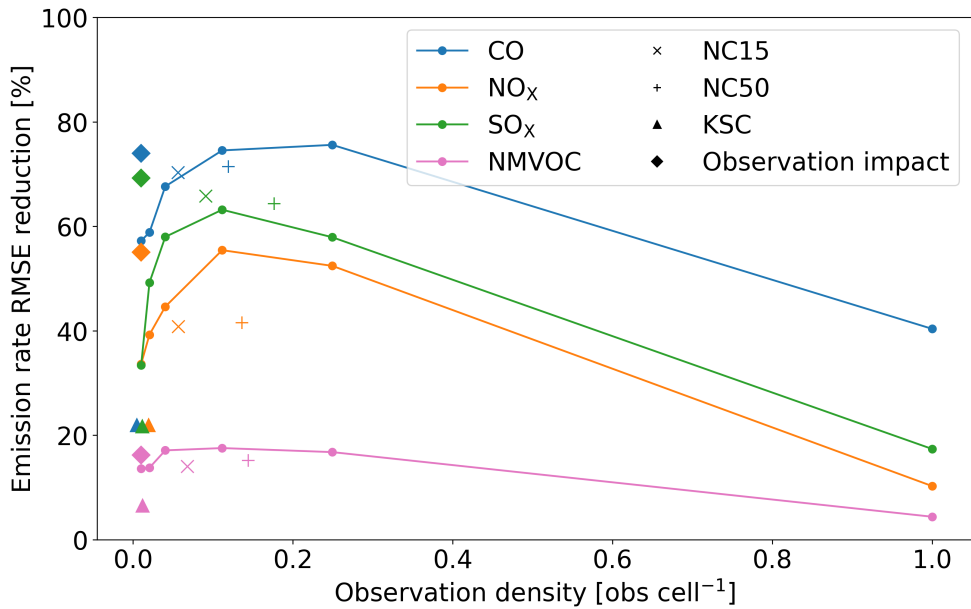


Figure 5.4. Emission rate RMSE reductions in [%] relative to the maximum RMSE at 17 March 2016 for CO (blue), NO_x (orange), SO_x (green), and NMVOC (pink). For NO_x and SO_x the observation densities of NO₂ and SO₂ are shown, respectively. Since no observations for NMVOC are used, the average observation density is indicated. The RMSE is shown for the regular grid configurations by dots, by "x"-markers for the correlation based configuration NC15, by "+"-markers for configuration NC50, by triangles for the KSC configuration, and by diamonds for the observation impact configuration. The observation density for the grid parameter 10 configuration is restricted to observations over land. This figure is identical to Fig. 4.6, except for the addition of the results for the observation impact configuration.

Figure 5.5 shows the concentration RMSE reduction of the species CO, NO₂, SO₂, O₃, and NMVOC for 17 March 2016 of the observation impact AR. The results of the other experiments (see also the description of Fig. 4.6) are added for comparison. Note that the indicator for the RMSE percentage of CO and SO₂ for the observation impact run overlap. The figure indicates the advantage of the observation impact AR compared to the benchmark ARs, where the former shows similar performance as the best regular grid configuration but with a lower observation density. The quantitative results are comparable to those of the emission rate corrections. For CO the observation impact AR achieves a RMSE reduction of 69 %, compared to 72 % in the grid parameter 3 and 2 AR. The RMSE reduction for NO₂ is equal to the value for the best performing benchmark configuration, with 64 %. For SO₂ the RMSE reduction is 68 %, compared to 70 % in the grid parameter 3 configuration. In the case of O₃ the observation impact AR achieves a concentration RMSE reduction of 44 %, and the grid parameter 3 AR 45 %. While the best performance is attributed to the KSC AR

with a RMSE reduction of 48 %. For NMVOC the RMSE reduction is evaluated at 17 %, compared to 19 % for the grid parameter 3 configuration. A comparison of evolution of the concentration RMSE percentages from 10 to 17 March 2016 is shown in Appendix B.1.

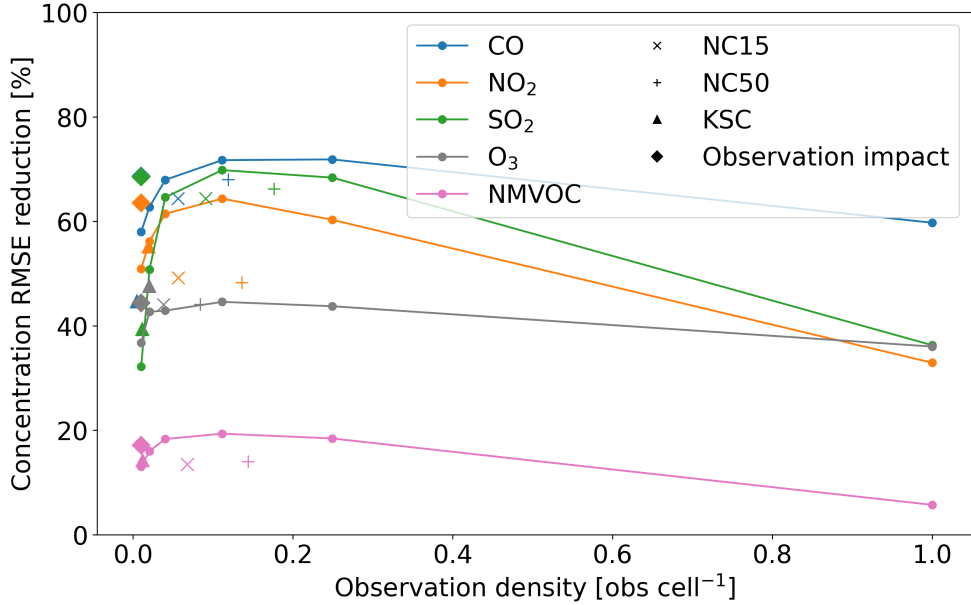


Figure 5.5. Reductions of the concentration RMSE in [%] relative to the maximum RMSE for the observation impact assimilation run, compared to the regular and irregular grid benchmark configurations at 17 March 2016. The RMSEs are shown for the species CO (blue), NO₂ (orange), SO₂ (green), O₃ (grey), and NMVOC (pink) in dependence to the observation densities. For NO_x and SO_x the observation densities of NO₂ and SO₂ are shown, respectively. Since no observations for NMVOC are used, the average observation density is shown. The RMSEs of the regular grid configurations are indicated by dots, by "x"-markers for the correlation based configuration NC15, by "+"-markers for configuration NC50, by triangles for the KSC configuration, and the RMSEs of the observation impact configuration by diamonds. The observation density for the grid parameter 10 configuration is restricted to observations over land. Note that the values for CO and SO₂ of the observation impact assimilation run overlap.

These results confirm the analysis of the emission rate RMSE reduction. The observation impact configuration is able to capture the emission sources and outperform the benchmark configurations with comparable density.

5.4. Application to satellite data

The previous section shows that the algorithm based on the observation impact (cf. Tab. 5.1) is able to select an observation configuration with good performance in terms of reproducing the nature run's emission rates. Since in an operational setting, the number of ground observations is limited and their locations cannot easily be changed, a common approach is to supplement the existing ground-based observational network by satellite observations. This approach is tested in the following.

5.4.1. Setup of the satellite OSSE

The OSSE framework established in Chapt. 4 is used for the evaluations in this section. The same nature run and assimilation run setup is employed with the experiment period extending 8 assimilation days, from 10 to 17 March 2016, according to Sec. 5.3.

The satellite observations used in the following experiments are created to resemble hourly measurements of the total tropospheric column-density of the respective species. They are extracted from the nature run concentration fields of CO, NO₂, SO₂, and O₃ by integrating the concentration over the entire model column without using an averaging kernel. For the calculation of the cost function an error of $R = 10\%$ of the total column-density is applied to all observations of CO and O₃. For NO₂ and SO₂, the error is set to $R = \max(3 \times 10^{15} \text{ molec cm}^{-2}, 20\%)$ of the total column-density. These error values were chosen to resemble those encountered in the operational assimilation of satellite data. The observation operator equals

$$H = \sum_k c(k)dk, \quad (5.7)$$

where k is the vertical layer index, $c(k)$ is the corresponding concentration, and dk the layer thickness. In contrast to Timmermans et al. (2015), no instrument simulation of the satellite is performed. Thus, limiting factors, such as cloud cover, signal to noise ratio, and distortions due to slanted columns are not considered. The present study aims to quantify the achievable emission rate RMSE reduction by using and incorporating satellite data. Introducing those aforementioned restrictions would reduce the available observations and limit the comparability to the benchmark experiment from Chapt. 4. Thus, the artificial satellite observations are designed to take instantaneous, hourly column-density measurements of the model domain over land.

For the purpose of this evaluation, different observation configurations are generated: one configuration of only satellite observations and supplemented configurations, either based on the observation impact or the KSC configuration. An overview of the supplemented observation configurations is shown in Tab. 5.2. OI+sat is generated from the observation impact configuration (OI), supplemented with 541 satellite observations per species. KSC+sat is a supplementation of the KSC configuration with 541 satellite observations per

species. Here, the ground observations over the ocean surfaces from the OI configuration are included as well to ensure comparability. Due to the different number of species specific observations in the KSC configuration (cf. Tab. 4.3), the total number of observations varies for each species. The Sat configuration consists only of 541 satellite observations. Thus, it can directly be compared to the original observation impact configuration. OI+grnd serves as a reference configuration. It is based on the original observation impact configuration, supplemented with 541 additional ground-based observations per species. It is intended to provide insights into the impact of adding observations to the optimal ground-based observations. The supplementation procedure is as follows: the exclusion areas from step 3 of the observation impact algorithm (cf. Tab. 5.1) are added to the base configuration. Then, the algorithm is executed with the desired observation type (cf. Tab. 5.2). For satellite observations, the background RMSE (cf. Eq. (5.5)) is determined with respect to the column-density of the NR and the CR calculated using Eq. (5.7).

Table 5.2. Summary of the observation configurations analysed in this section. The KSC configuration has been modified to include the grid based observations over the ocean, which are also present in the OI configuration.

Identifier	Base configuration	Additional observations	
		Ground	Satellite
OI+sat	OI	-	541
KSC+sat	KSC*	-	541
Sat	-	-	541
OI+grnd	OI	541	-

Figure 5.6 shows the observation locations for the different configurations of Tab. 5.2 for CO for 10 March 2016. The observation locations for NO₂, SO₂, and O₃ are shown in Appendix B.2. When comparing the OI+grnd and the OI+sat configuration, the overall observation locations are comparable with small deviations. For CO, additional ground observations are introduced in the area and to the east of Moscow, while the additional satellite observations instead cover the south of Ukraine and Moldova. For O₃, the additional ground observations in OI+grnd cover mostly central Europe and the Moscow area, while the additional satellite observations instead cover Bulgaria, Turkey, and the north of Tunisia and Algeria. In the case of NO₂ and SO₂ no mayor differences in the spatial placement of the supplemented observations is present. The difference between the observation impact configuration (cf. Fig. 5.2) and the Sat configuration coincides qualitatively with the difference between the OI+grnd and the OI+sat. In the KSC+sat configuration, the additional satellite observations for CO coincide largely with the placement of the observations in the sat configuration. This is due to the low spatial coverage of the ground observations. For NO₂ and SO₂ most additional satellite observations are placed in the Moscow area, the south of Ukraine, and Turkey (SO₂ only) since there is no ground observation coverage of

these areas. The satellite observations for O_3 in the KSC+sat configuration mostly cover North Africa and Turkey.

CO

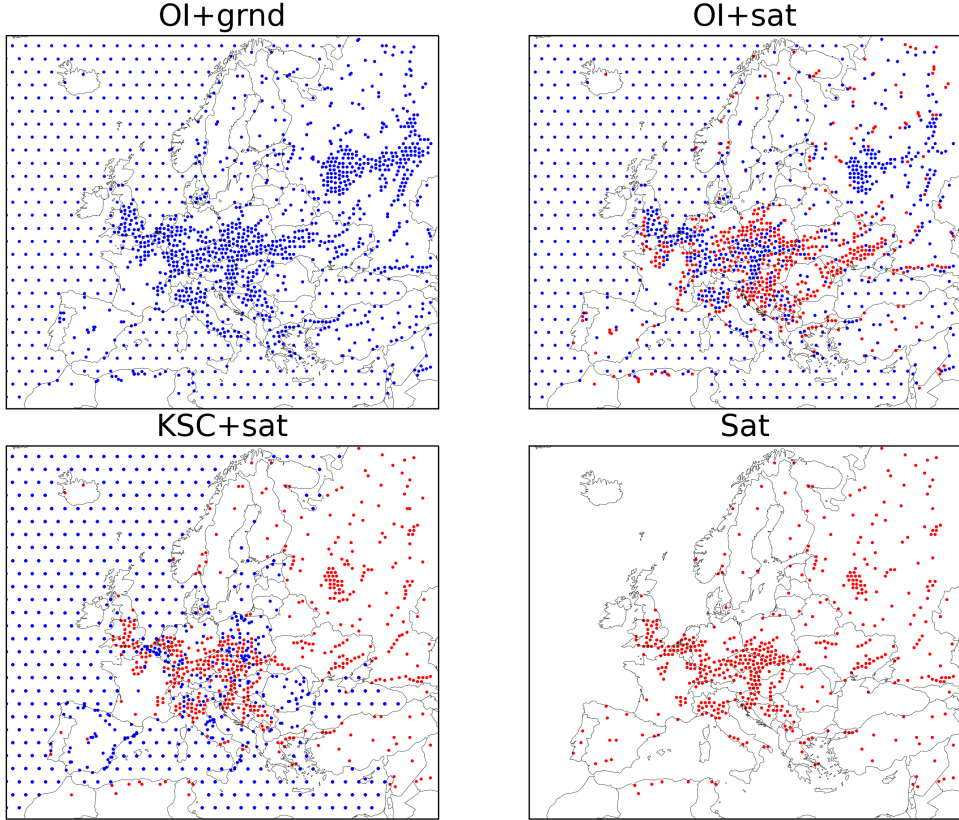


Figure 5.6. Configuration of the CO observation networks considered in this section for 10 March 2016. Each dot represents an observation site with the ground observations in blue and satellite observations in red.

5.4.2. Analysis of the satellite OSSE

Figure 5.7 compares the concentration (colored bars) and emission rate (circles) RMSE reductions for the tested observation configurations for 17 March 2016, i.e. after seven days of applying the data assimilation scheme.

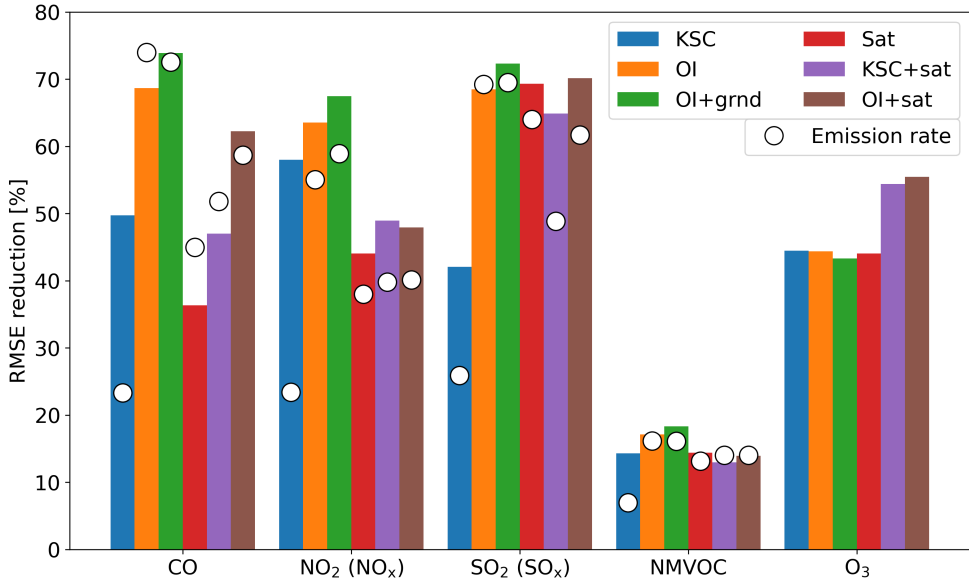


Figure 5.7. Concentration (bars) and emission (circles) RMSE reductions in [%] of the maximum RMSE for the supplemented and base configurations on 17 March 2016 for CO, NO₂ (NO_x emissions), SO₂ (SO_x emissions), NMVOC, and O₃. For O₃, no emission rates are available as it is a secondary pollutant. The y-axis is limited to values less than 80 %.

For CO, the Sat assimilation run shows the lowest concentration RMSE reduction (36 %) of all investigated configurations. The emission rate RMSE reduction (45 %) is larger than for KSC but lower than the other calculated emission rate RMSEs. This indicates that ground observations are superior to satellite observations for CO concentration optimization. However, the placement of the observations, based on the observation impact (Eq. (5.6)), still improves the emission rate optimization compared to the KSC. The KSC+sat assimilation run shows a decrease in the concentration RMSE reduction of 3 % compared to its base configuration and an increase in the emission rate RMSE reduction of 29 %. This is due to the placement of the additional observations which are targeted to sample the emissions sufficiently well, as mentioned above. The OI+sat configuration shows a decrease for both, the concentration and emission RMSE reduction compared to the OI configuration of 7 % and 15 %, respectively. The OI+grnd configuration shows improvement in the concentration RMSE reduction of 5 % and a decrease of 1 % in the reduction for the emission rate RMSE. This indicates that additional ground observations to the OI configuration have only slight effects, while additional satellite observations are detrimental to both the concentration and emission rate optimization. Generally, the initial value optimization due to satellite observations focuses on higher model layers than the optimization due to ground-based observations. This causes the above mentioned detrimental effect, since the emission rate

perturbations mostly affect the lower atmospheric layers. This effect will be examined in greater detail later on.

In the case of NO_2 and NO_x (for the emission rates), the Sat configuration shows the lowest RMSE reduction for the concentrations (44 %) and the second lowest emission rate RMSE reduction (45 %). The lowest emission rate RMSE reduction is calculated for the KSC configuration (23 %). Both, the KSC+sat and the OI+sat configuration show lower concentration RMSE reductions than their base configurations, With differences of 9 % and 16 %, respectively. Their overall performance is comparable to each other, reaching 49 % and 48 % concentration RMSE reduction, and 40 % emission rate RMSE reduction, each. The KSC+sat configuration, however, has a higher emission rate RMSE reduction than its base configuration by 17 %. OI+sat shows a lower value, by 15 %. The OI+grnd configuration improves both the concentration RMSE reduction and the emission rate reduction compared to the OI configuration by 3 % and 4 %, respectively. These results indicate that column integrated satellite observations for NO_2 are generally detrimental to the optimization of the modelled concentrations. While the realistic configuration (KSC) benefits from the additional satellite observations in the emission rate optimization, due to the improved representation of plumes from localized emission sources, the OI-based configurations show that additional ground observations are more beneficial. Here, the substitution of satellite observations hinders the emission rate optimization, while the base configuration is sufficient.

In the analysis of SO_2 concentrations and SO_x emission rates, the Sat configuration shows stronger concentration and emission rate RMSE reductions than for the other species, with 69 % and 64 %, respectively. This is due to the characteristics of SO_x and SO_2 . The largest concentration differences occur at the height of the planetary boundary layer (shown later) while the background concentrations are low. Thus, the information provided by the column density measurements is accurately resolved and sufficient information on the emission perturbations from the nature run is extracted. The KSC+sat configuration shows strong improvements compared to its base configuration. The concentration and emission RMSE reductions are improved by 23 %, each. The performance of the OI+sat configuration is slightly improved in the concentration correction (2 %), and degraded in the emission correction (7 %) reduction. Additional ground observations (OI+grnd configuration) cause a slightly larger improvement in the concentration RMSE reduction of 4 % and no change in the emission rate RMSE. This indicates, that additional observations do not improve the analysis of the OI configuration significantly, except for a degradation of the emission rate optimization by additional satellite observations. However, adding satellite observations to the KSC configuration greatly improve the results. This shows the importance of capturing the emission sources for SO_x and that column integrated satellite observations of SO_2 provide sufficient information for the optimization of both, the model state and the SO_x emission rates.

The optimization of NMVOC concentrations and emissions shows only a slight dependence on the used observation network. The concentration RMSE reductions span values from 13 % (KSC+sat) to 18 % (OI+grnd), while the emission rate RMSE reductions span values from 13 % (Sat) to 16 % (OI and OI+grnd). Only the emission rate RMSE reduction of the KSC configuration is calculated at 7 %. However, the best performance is achieved by the OI+grnd configuration. This low impact of the observation configuration on the NMVOC corrections is due to the absence of direct NMVOC measurements in these experiments. Thus, no constraints for the optimization are available and its effectiveness is reduced (Lange et al., 2023). The improvements in the emission rates is attributed to the improved performance in the other species, however the exact relationship is complex and not resolved in the analysis.

In the analysis of O_3 , the concentration RMSE reduction for the observation configuration with single instrument types perform similarly with values of 44 % for the KSC, Sat, and OI configuration, while the OI+grnd configuration reaches 43 %. The combination of ground and satellite observations show improved RMSE reductions with 54 % for KSC+sat and 55 % for OI+sat. This shows, that the ground and satellite observations contain relevant and non-redundant information for the initial value optimization of O_3 . The satellite observations are able to capture concentration changes in higher model layers which occur due to vertical transport of O_3 and relevant precursors in subsequent days (cf. Appendix B.3). As was shown by Erraji et al. (2024), the analysis benefits significantly from information about O_3 concentrations in higher layers. The ground observations are able to capture the effect of localized emission sources of O_3 precursor species more accurately. However, the optimization of NO_2 concentrations and NO_x emission rates is not as important for the O_3 concentration optimizations. This is shown in the difference of the NO_2/NO_x optimization for OI and OI+grnd compared to KSC+sat and OI+sat, where the former show better performance but lead to low O_3 concentration RMSE reductions. This highlights the importance of the higher altitude initial value corrections for O_3 .

To analyse the effect of the satellite observations on the initial value optimization, Figure 5.8 shows the component of the gradient of the cost function (Eq. (2.13)) for 10 March 2016 (first day of the experiments) pertaining to the initial value for CO, NO_2 , SO_2 , and O_3 , averaged along the model y-coordinate, for the Sat (left) and OI (center) assimilation run, as well as the concentration difference between the NR and CR (right), all normalized to $[-1, 1]$. These configurations are chosen to compare the influence of satellite and ground observations on the initial value optimization.

For CO, the gradient of the OI assimilation run is confined to the lower layers of the model atmosphere, while the gradient of the Sat assimilation run is extended over the entire layers. This is due to the satellite observations introducing information about the entire model column, while ground observations impact only the lower layers. However, in the OSSE, the differences between the NR and CR concentrations are solely caused by the emission

rate perturbation, which influences mainly the lower atmospheric layers. The extension of the observation influence to higher atmospheric layers is thus an inaccuracy compared to the expected state and explains the deterioration of the concentration RMSE reduction in the Sat assimilation run compared to the OI assimilation run. This effect is most pronounced for CO, due to the higher atmospheric background concentrations leading to stronger initial value optimization compared to other species.

For NO₂ the initial value gradient in the Sat assimilation run does not give meaningful values for altitudes above approximately 500 m. This is due to the majority of the NO₂ concentrations being confined to lower altitudes, close to the emission sources. However, the initial value gradient in the Sat assimilation run shows stronger influences in higher altitudes than for the OI assimilation run. It surpasses the height to which the concentrations are affected by the emission rate perturbations in the nature run.

The comparison of the SO₂ gradients shows no significant differences between the assimilation runs. This is due to the majority of the SO₂ concentrations being present in the lower atmosphere. However, SO₂ as a fraction of SO_x is emitted by strong point sources (e.g. industries) which emit in elevated layers. Thus, the corrections are strongest at approximately 500 m, which is well captured in ground and satellite observations as indicated by the gradient.

For O₃, the satellite observations introduce large gradient values in the top layers of the atmosphere. This is due to the large atmospheric O₃ concentrations in higher altitudes. The satellite observations include these concentrations in the column measurements.

In general, column density observations from a satellite influence the innovation vector of the entire model column. However, the concentration differences in the investigated OSSEs are mostly confined to the lower atmospheric layers since they are the result of the emission rate perturbations applied in the NR generation. O₃ is the only exception, since concentration differences are present in higher layers at later days. This benefits the O₃ optimization. For the emitted species, satellite observations introduce less information to recover the emission rate perturbation than ground observations, as is shown in the analysis of Fig. 5.7.

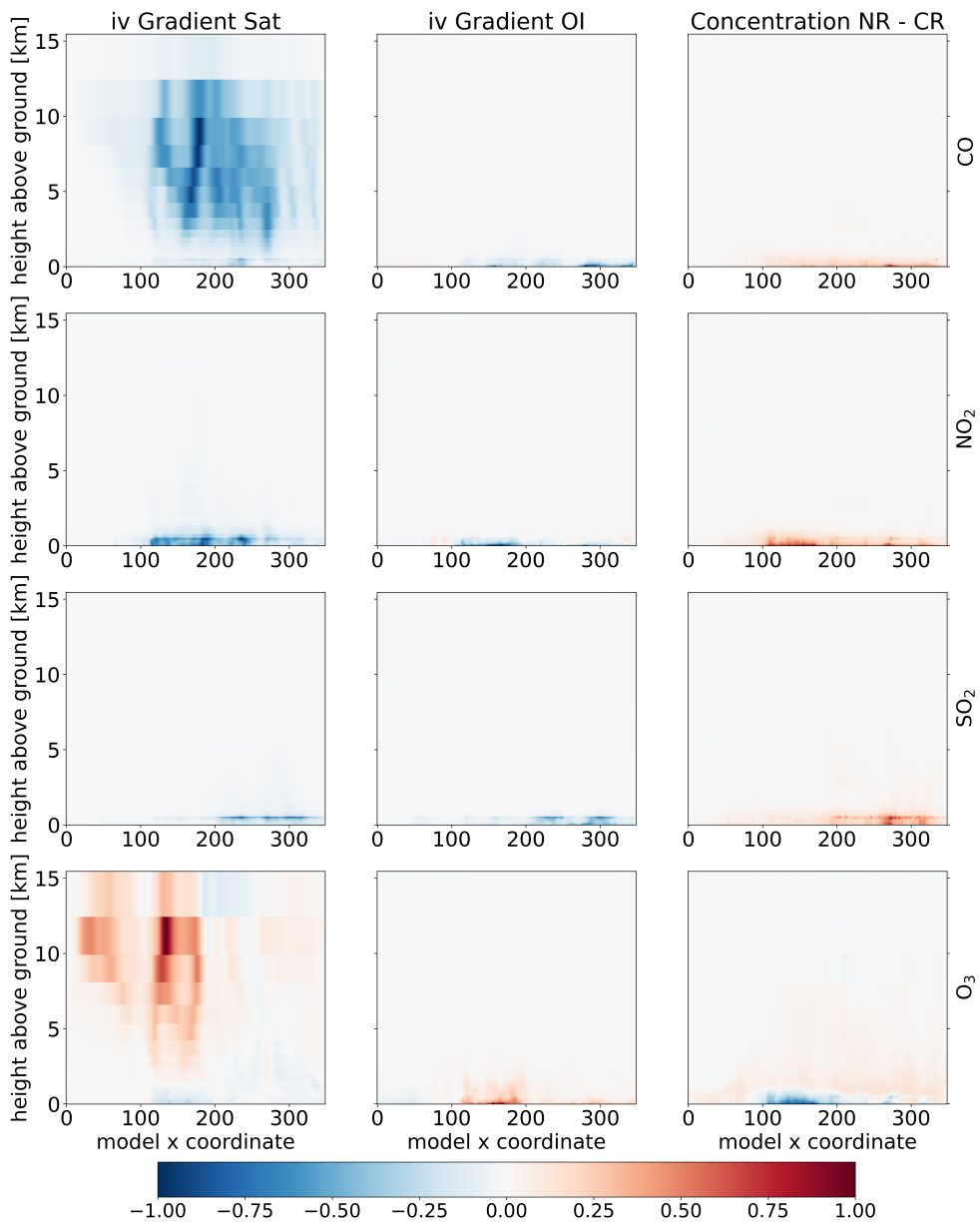


Figure 5.8. Normalized initial value gradient averaged along model y-direction for the Sat (left) and the OI+grnd assimilation run (center) and the normalized and along y-direction averaged concentrations difference between the NR and CR (right). The values are shown for 10 March 2016 and for the species CO, NO₂, SO₂, and O₃ in descending order. The y-axis shows the height above the ground in [km] and the x-axis the model x-coordinate.

6. Summary and conclusions

In this thesis, the influence of the observation network design on the representativity of the model analysis and the optimization of emission rates in the context of 4D-var was investigated. A method for the sub-sampling of existing ground observations to generate a representative model analysis was developed and compared to an operational sampling. This was done to improve the usage of the available observation data in an operational setting. Here, the representativity was defined as the difference between the RMSE of the assimilated and the validation data. A low difference, or a high representativity, indicates a lower spatial variation in the RMSE of the model analysis. An algorithm, based on k-means soft constraint clustering, was shown to generate clusters of observation stations, which distinguish atmospheric chemical properties in the observation time series. The clusters are independent of the used CTM and result purely from the observation data. The resulting split into assimilation and validation data sets improved the representativity of the model analysis compared to an operational data splitting especially with respect to the spatial coverage.

To evaluate the influence of the observation network on the optimization of emission rates, an OSSE, with a known emission rate perturbation, was established. This allowed to quantify the quality of the observation network based on the emission rate corrections directly. This is not possible in a real-world application, since there the correct emission rates are unknown. It was demonstrated that the generated nature run exhibits sufficient difference to the control run to avoid the identical twin problem commonly encountered in OSSEs. These differences quantify the sensitivity of the chemical species to the introduced emission perturbations.

However, since the model used for the generation of the nature run differed only by a perturbation of the emission rates from the model used for the control run and the assimilation runs, inaccuracies due to unknown or unrepresented physical or chemical processes between the model and the real atmosphere, were not considered. Thus, while the identical twin problem was generally avoided, the results of the benchmark OSSEs are still prone to be overly optimistic. Because similar analyses have not been presented in the literature before, this setup was chosen to focus on the emission rate corrections depending on the observation data, while excluding all other influencing factors. This allowed for a detailed assessment of the models performance in an idealized and intentionally unrealistic setting. Furthermore, the perturbation of the emission rates in the nature run by a constant factor

of 2 is not a realistic scenario, where the emission rate corrections can vary in space and time. However, introducing variations in the emission rate perturbations may cause adjacent perturbations to compensate each other and limit their effect on the concentration in the nature run. The duration of the OSSE period was initially set to two weeks of data assimilation and reduced to one week in the progression of this thesis. This was done to reduce the compute time necessary for the evaluation. Ideally, the experiments would be extended to the entire year. However, this is not feasible due to the computational demand of the 4D-var methodology. Future studies with different simulation periods are required to ensure the general validity of the results.

The observations are extracted from the nature run without the additional consideration of measurement errors. Application of these, in the form of Gaussian perturbations, is shown to reduce the effectiveness of the emission rate optimization leading to the need for longer assimilation periods. However, the results indicate that the assimilation without considering measurement errors achieve comparable results after additional days of assimilation. This simplification to ignore measurement errors was made to avoid the computational cost of the error application and to study the observation configuration in a highly idealized setting. Measurement errors introduce uncertainties in the data assimilation process and accelerate the adverse effect of high density observation networks on the condition number. Thus, the detrimental effect of a large number of observations on the optimization is expected to occur at lower observation densities. The model cannot reproduce the spatial heterogeneity of the Gaussian perturbations. This is suspected to shift the optimal observation density to lower values.

Within the OSSE framework, a benchmark experiment with a regular observation distribution was performed to analyse the influence of varying observation network densities on the emission rate optimization of the EURAD-IM. The benchmark runs provide reference data for future evaluations of observation configurations and different OSSE setups. They can furthermore provide insight on the regional error estimations in operational setups. The primary result of the benchmark experiments is the existence of an optimal observation density and the detrimental effect of higher densities. This indicates the necessity of thinning observational data if observations with high spatial coverage (such as satellites) are available. The benchmark OSSE results for an operational configuration, in form of the KSC configuration, showcases the need for the supplementation with additional observation data. Only for the optimization of O_3 , the KSC configuration shows a better performance than the benchmark experiments. Possibly, this could be a result of an O_3 -targeted placement of the observations in the operational setting. Further, the results for the KSC configuration suggest that the analysis is oversensitive to O_3 observations leading to a low performance in correcting other species but O_3 .

With the purpose of developing a targeted observation network, a surrogate model of the gradient of the 4D-var cost function specific for emission rates was proposed. From

this, the observation impact of a specific observation was defined, measuring the impact individual observations on the emission rate optimization. It combines an adjoint sensitivity field with the emission inventory data, which is then scaled by the background RMSE of the observation. The latter determines the discrepancy of the observation to the model state. In a real world application, this discrepancy does not have to correlate to the magnitude of the emission rates. Inaccuracies in the boundary data of trace gases can have strong influences on the atmospheric background concentrations for long-lived species such as CO. However, this effect is not relevant in the presented experiments since the differences in concentrations solely result from the emission rate perturbations.

By combining the findings of the benchmark experiment and the observation impact, an algorithm for the generation of a targeted observation configuration was developed and the resulting configuration was tested within the established OSSE framework. The resulting observation configuration outperforms the benchmark configurations in the emission rate and concentration optimization. When similar RMSE reductions are achieved, the proposed configuration reached them after fewer assimilation days. This indicates that it can better capture temporal variations in the emission rates in a realistic setup.

To examine the supplementation of operational ground based observation networks by other instruments, the OSSE is extended to include satellite observations. The satellite observations are extracted from the nature run without considering instrument specifics and limitations, which is done to avoid restrictions such as temporal variability and to examine the general effect of additional satellite observations under best, yet unrealistic, conditions.

It was shown that additional satellite observations are less beneficial for the optimization of emission rates and concentrations than additional ground observations. Furthermore, for CO and NO₂ (NO_x) the satellite observations are detrimental for the optimization and reduce the accuracy. This effect results from the additional information obtained by observing the entire model column, while the primary effects of emission rate perturbations are confined to lower atmospheric layers. Only the optimization of O₃ benefits from the satellite observations. However, due to the O₃ layer hindering satellite based O₃ observations in addition to measurement uncertainties, this increase in performance is not expected in an real application.

By comparing the KSC configuration to the newly generated configurations, it is shown that its relatively good performance in the concentration optimization is not due to accurately capturing the emission rate perturbations but rather due to the initial value correction, which compensates the inaccuracies resulting from inaccurate emission rates. As in an operational setting the true emission rates are unknown, increasing the accuracy of their optimization is, therefore, more beneficial for air quality assessment than the accuracy of the concentration optimization.

The studies performed in this thesis initiate the assessment and development of observation configurations for emission rate optimization. Future studies are required to investigate

the effect of the restrictions in the OSSE on the results, especially with respect to the incorporation of a second CTM for the evaluation to include deviations resulting from differences in the model realization. Furthermore, an investigation of an increase in the model resolution is expected to improve the conclusions of this study for the operational placement of observations due to the incorporation of localized regional effects. The initial studies in this thesis are undertaken with a $15\text{ km} \times 15\text{ km}$ horizontal resolution. Higher resolutions are available which would allow for a more accurate representation of the adjoint sensitivity field including topographical effects. However, the detrimental effect of high observation densities may occur sooner than in the performed benchmark tests due to the greater number of possible observation sites. This would increase the condition number for lower observation densities.

Satellites are able to provide measurements with a high spatial coverage and are increasingly available. Thus, they are used to close observational gaps in operational observation networks. However, as shown in this work, their usage in emission optimization, as for example in Gaubert et al. (2020), needs careful examination for their possible degradation of the optimization. Therefore, other measurement instruments may be more beneficial for the emission rate optimization, such as drone observation capable of measuring in the planetary boundary layer (Erraji et al., 2024), although they have a low spatial coverage. Their inclusion in the developed OSSE framework can provide further insight into an optimal design of observation networks for emission rate optimization.

The number of observations chosen for the satellite observation and observation impact configuration is not necessarily the optimal number. Finding an optimal number of instrument specific observations will aid in the development of future improved observation configurations.

General acknowledgments

This work was partially performed as part of the Helmholtz School for Data Science in Life, Earth and Energy (HDS-LEE) and received funding from the Helmholtz Association of German Research Centres.

The author gratefully acknowledge the Earth System Modelling Project (ESM) for funding this work by providing computing time on the ESM partition of the supercomputer JUWELS (Jülich Supercomputing Centre, 2021) at the Jülich Supercomputing Centre (JSC).

A. Appendix: Supplement material to the benchmark experiments

A.1. Influence of Gaussian observation error on the assimilation run

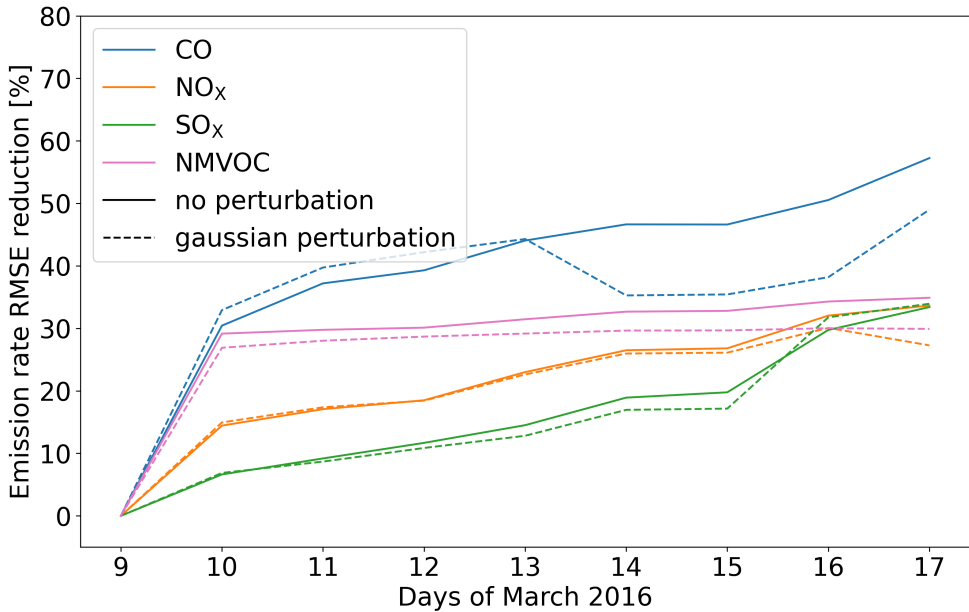


Figure A1. Evolution of the emission rate RMSE reductions in [%] of the maximum RMSE for the regular observation grid configuration with grid parameter 10 during the assimilation window from 10 to 17 March 2016. The solid lines shows the result using unperturbed observations extracted from the nature run. The dashed lines show the result for the analysis using observations with additional Gaussian perturbation to reflect observation errors. Results are shown for CO (blue), NO_x (orange), SO_x (green), and NMVOC (pink). The RMSE fraction for 09 March 2016 is shown for completeness and the y-axis is limited to values less than 80 %.

The application of the measurement errors, in the form of random Gaussian perturbations,

significantly increases the computational cost of generating the observations. To investigate whether measurement errors are necessary to ensure the realism of the artificial observations, two assimilation runs are generated. Both ARs utilise the same set of observations (grid parameter 10) extracted from the nature run, with a Gaussian perturbation being applied to the observed concentration of one AR and the others being left unperturbed (i.e., the observations are identical to the nature run concentrations). Figure A1 shows the evolution of the emission rate RMSE reduction for CO, NO_x, SO_x, and NMVOC emissions for the grid parameter 10 configuration with (dashed lines) and without (solid lines) Gaussian perturbed observations. The assimilation period is restricted to the 10 to 17 March 2016. The RMSEs are normalized by the maximum RMSEs for each species (cf. Tab. 4.5).

The evaluation of the error evolution shows that both yield similar analysis accuracies. The difference in the final RMSE reductions is largest for CO, for which the ARs achieved an RMSE reduction of 57 % without Gaussian errors and 49 % with their inclusion. For NO_x the results are 34 % and 27 %, for SO_x 33 % and 34 %, and for NMVOC 35 % and 30 %. A reduction in the effectiveness of the emission rate optimization by assimilating perturbed observations is expected, since the perturbation introduces uncertainties in the data assimilation process. However, the results indicate that the final plateau of the emission rate RMSE reduction (cf. Sect. 4.3) is expected to be reached after additional days of assimilation.

Due to this effect and the increased computational cost in applying the perturbation, in the following investigation the observations are directly extracted from the nature run without the application of Gaussian perturbations.

A.2. Comparison of the emission rate RMSE for grid parameter 10

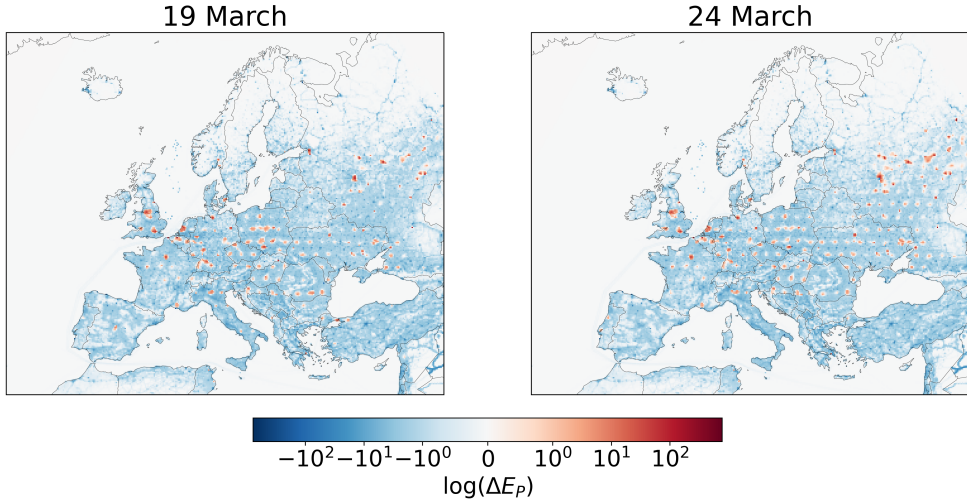


Figure A2. Logarithmic plot of the CO emission rate correction errors of the grid parameter 10 assimilation run for 19 and 24 March 2016. Blue signifies an underestimation of the nature run emission rates while red indicates an overestimation.

Figure A2 shows the CO emission rate correction errors of the grid parameter 10 configuration AR for 19 (left) and 24 (right) March 2016. Negative values, in blue, signify that the corrected emission rate is lower than the emission rates from the NR. Positive values, in red, show an overestimation. In both plots, the locations of the observations often coincide with overestimations of the emission rates. Especially in central to eastern Europe the pattern of the regular grid observation configuration is clearly noticeable by the overestimation of the emission rates. The emission rate RMSE reduction of the entire domain for 19 March is 59 % and for 24 March, 46 % (cf. Fig. 4.5). In this period the emission rate RMSE reduction is degraded by the additional days of data assimilation. This is due to an increase in the overestimation of the emission rates by the model, especially in the region around Moscow.

A.3. Regional distribution of the emission rate correction errors

In the following, emission rate correction errors for four different observation configurations (KSC, and grid parameters 10, 3, and 1) for 17 March 2016 are compared for CO (Figure A3), NO_x (Fig. A4), SO_x (Fig. A5), and NMVOC (Fig. A6).

Figure A3 for the CO emission rate reduction is explicitly discussed here but the general remarks are valid for all species. When comparing the observation configurations, the results for the lower observation density configurations show that the overestimations of the emission rates often coincide with observation locations. Especially in central to eastern Europe, the observation grid of the grid parameter 10 configuration is replicated by overestimations. This is due to the localization of the emission rate correction to the area of an observation. Thus, overestimation of the emission rates at these locations compensate for weakly corrected emission rates at remote locations. Generally, this leads to improved concentration values but lower accuracy in the emission rate correction. In the grid parameter 3 configuration, the grid structure of the observation configuration is not replicated by overestimations. This is due to the increased density of the observations but also due to the resulting improved coverage of emission sources. The more spatially accurate emission rate corrections reduce the amount of overestimation in the results. The results for grid parameter 1 show a general underestimation of the emission rates. Due to the extreme number of observations, the algorithm is unable to minimize the cost function and sufficiently optimize the emission rates, although the overestimation seen in the other experiments is mostly suppressed. The KSC configuration shows a focus of the emission rate corrections to the area of central Europe. This is due to the focus of the existing observation network to that area (cf. Fig. 4.3). However, the generally low number of observations leads to overestimation of the emission rates, comparable with the experiment using the grid parameter 10 observation configuration.

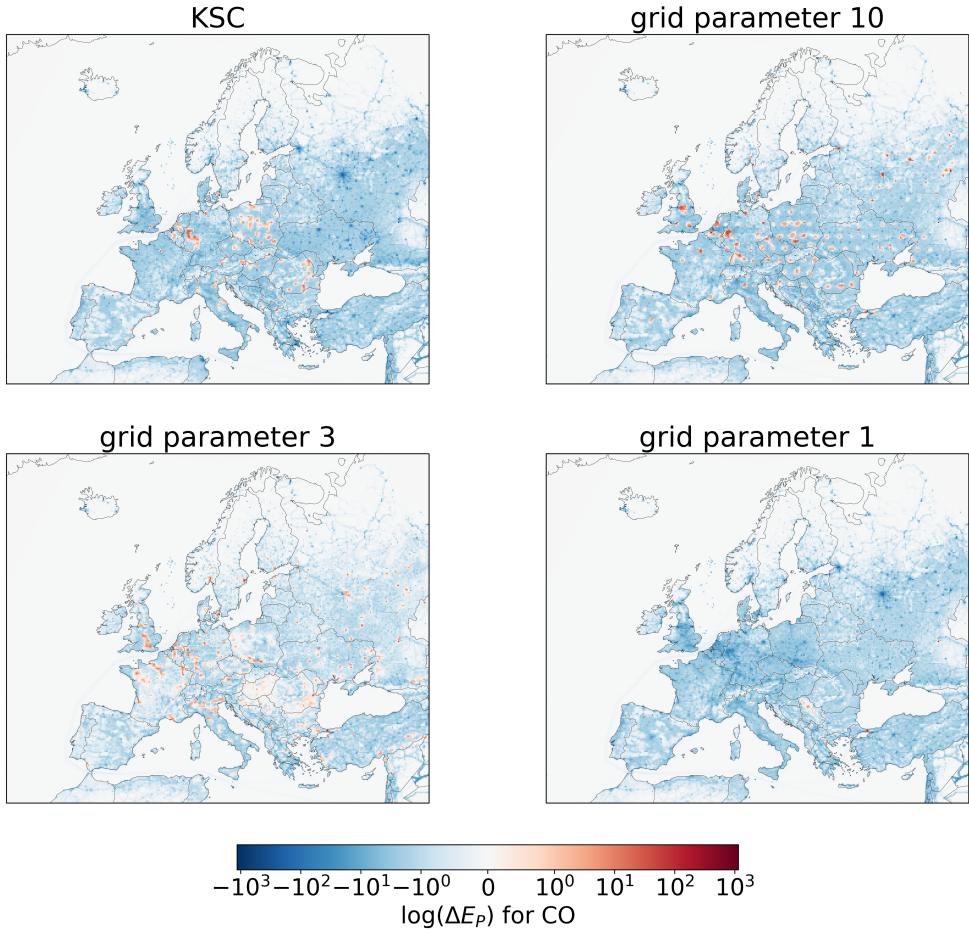


Figure A3. Logarithmic plot of the emission rate correction errors of CO for 17 March 2016. Shown are the results for the KSC (top left), grid parameter 10 (top right), grid parameter 3 (bottom left), and grid parameter 1 (bottom right) configuration.

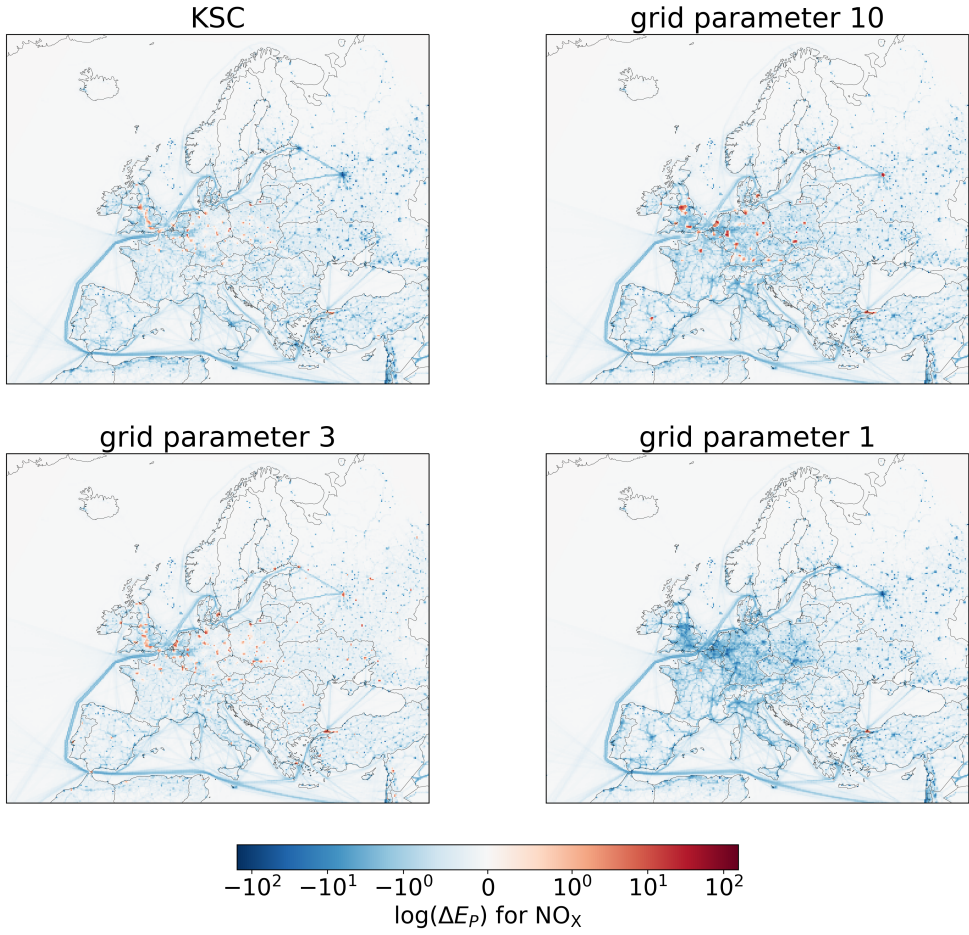


Figure A4. Logarithmic plot of the emission rate correction errors of NO_x for 17 March 2016. Shown are the results for the KSC (top left), grid parameter 10 (top right), grid parameter 3 (bottom left), and grid parameter 1 (bottom right) configuration.

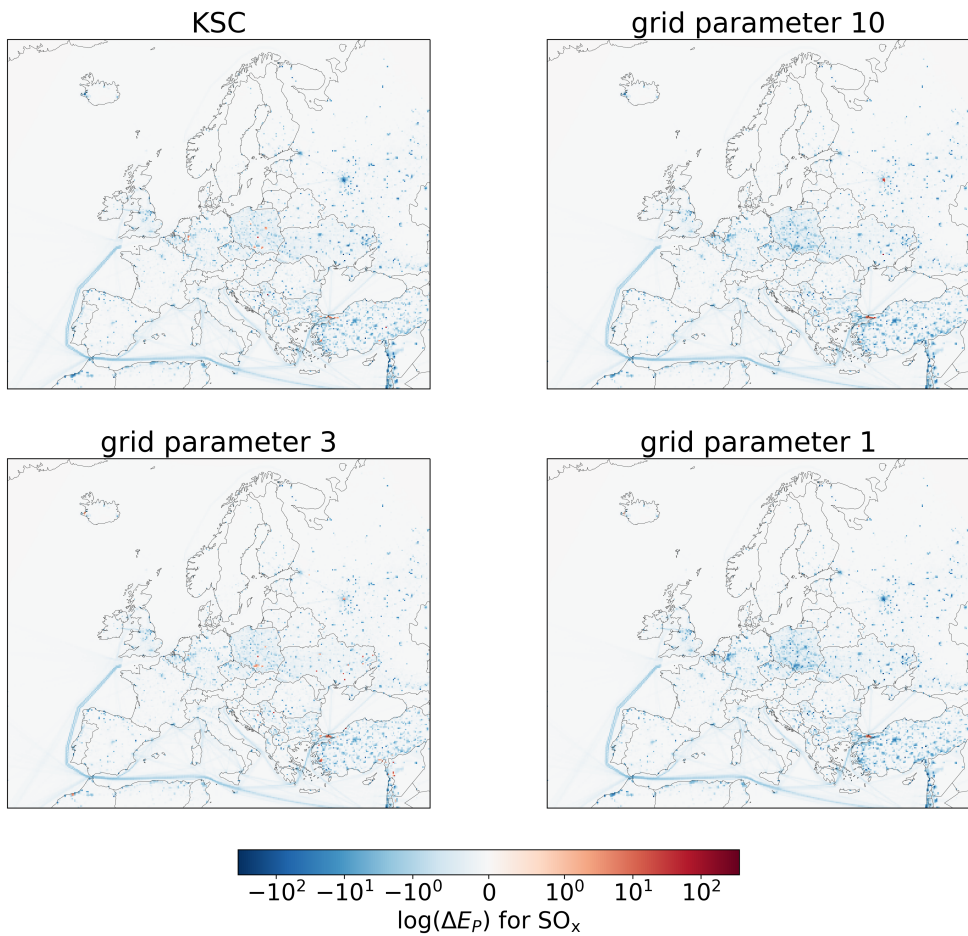


Figure A5. Logarithmic plot of the emission rate correction errors of SO_x for 17 March 2016. Shown are the results for the KSC (top left), grid parameter 10 (top right), grid parameter 3 (bottom left), and grid parameter 1 (bottom right) configuration.

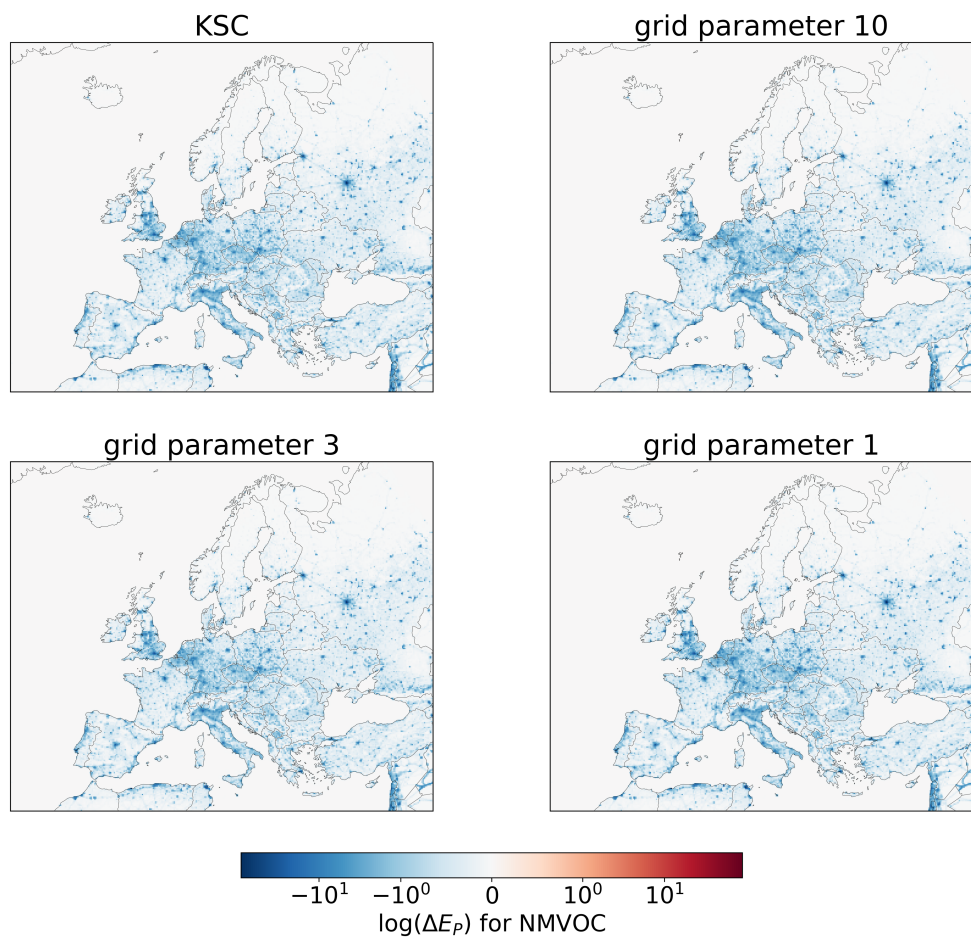


Figure A6. Logarithmic plot of the emission rate correction errors of NMVOC for 17 March 2016. Shown are the results for the KSC (top left), grid parameter 10 (top right), grid parameter 3 (bottom left), and grid parameter 1 (bottom right) configuration.

B. Appendix: Supplement material to the observation impact experiment

B.1. Comparison of the concentration RMSE evolution

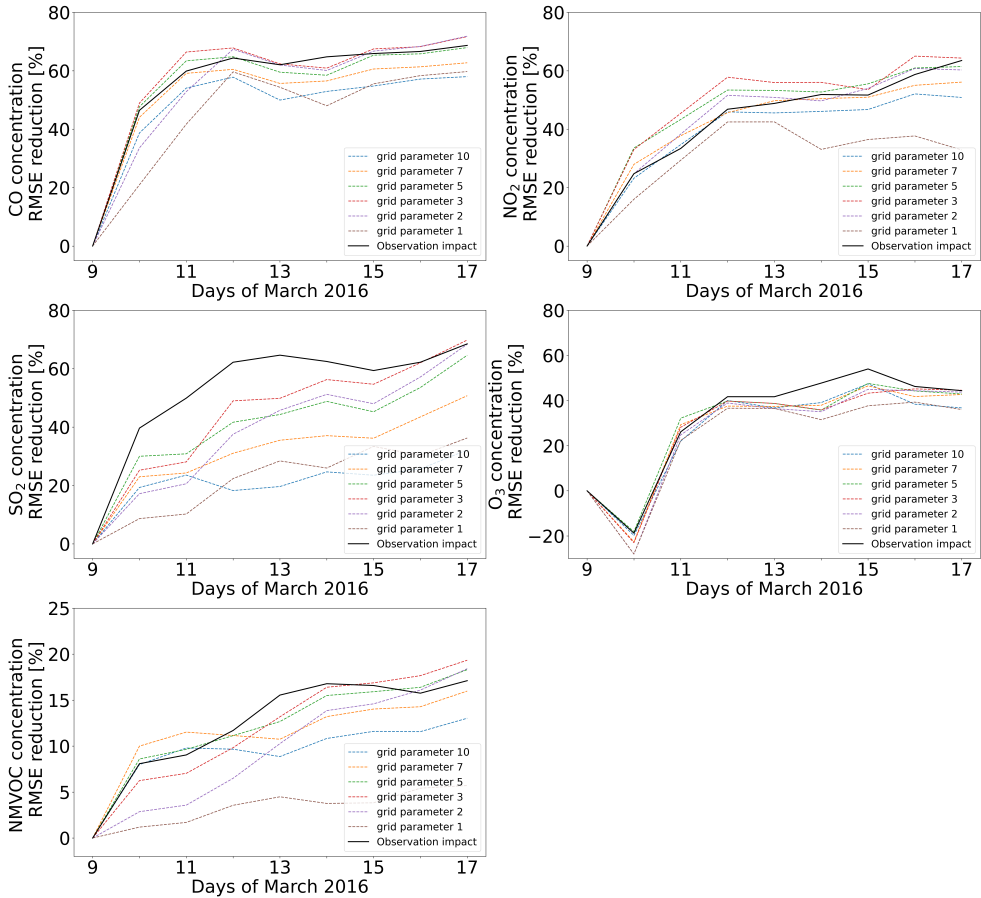


Figure B1. Evolution of the concentration RMSE in [%] relative to the maximum from 9 to 17 March 2016 for the observation impact (solid black line) and the regular grid benchmark assimilation runs (dashed lines). The figure shows the values for CO (top left), NO₂ (top right), SO₂ (center left), O₃ (center right), and NMVOC (bottom left). The y-axes are limited to values between 0 % and 80 %, except for O₃, where the lower limit is extended to -20 % and NMVOC, where the upper limit is set to 25 %.

B.2. Configurations of the observation networks in Sect. 5.4

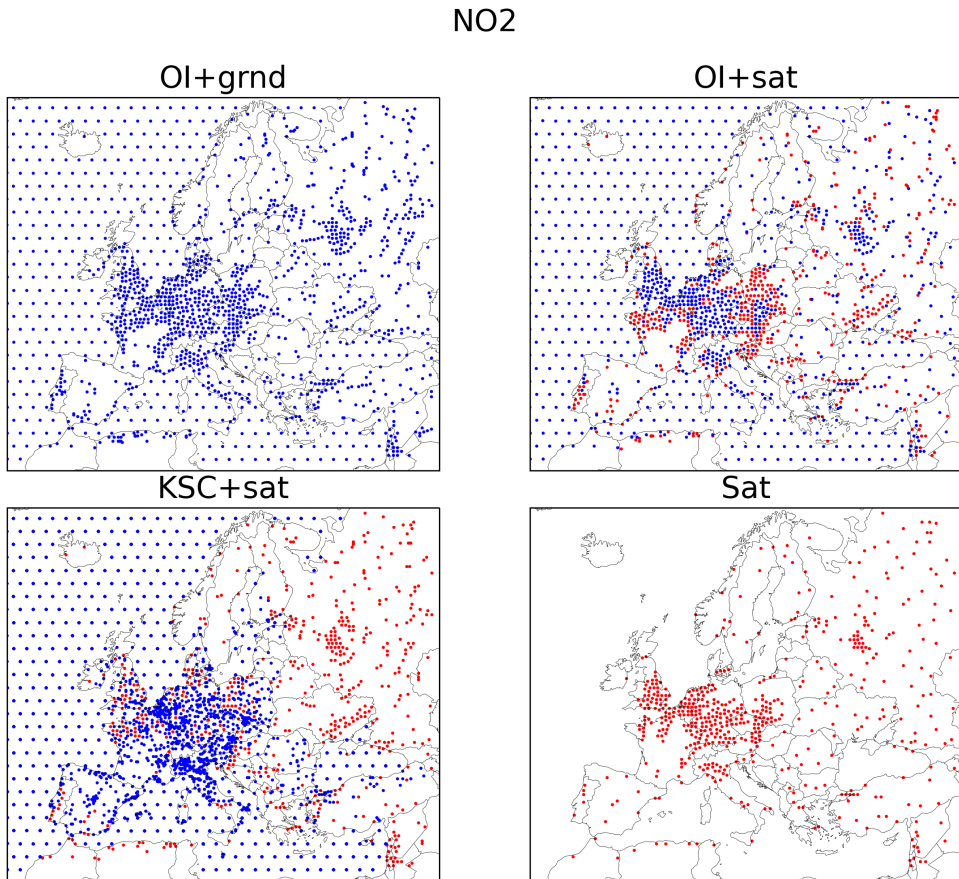


Figure B2. Configuration of the NO₂ observation networks considered in Sect. 5.4 for 10 March 2016. Each dot represents an observation site with the ground observations in blue and satellite observations in red.

SO₂

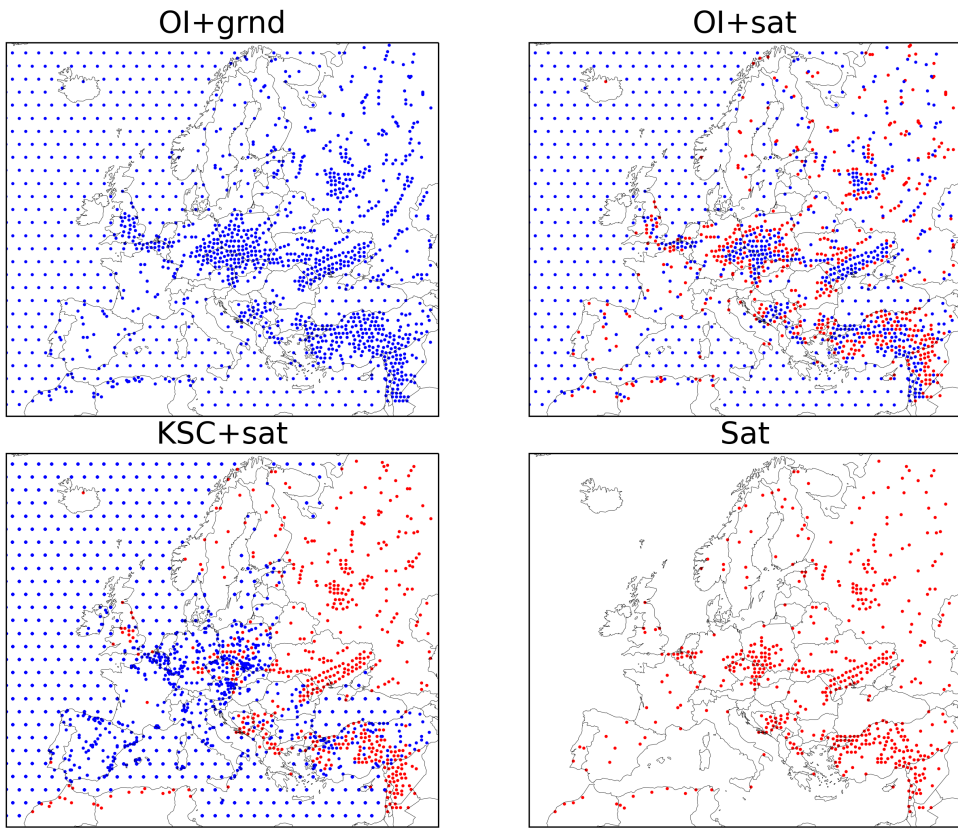


Figure B3. Configuration of the SO₂ observation networks considered in Sect. 5.4 for 10 March 2016. Each dot represents an observation site with the ground observations in blue and satellite observations in red.

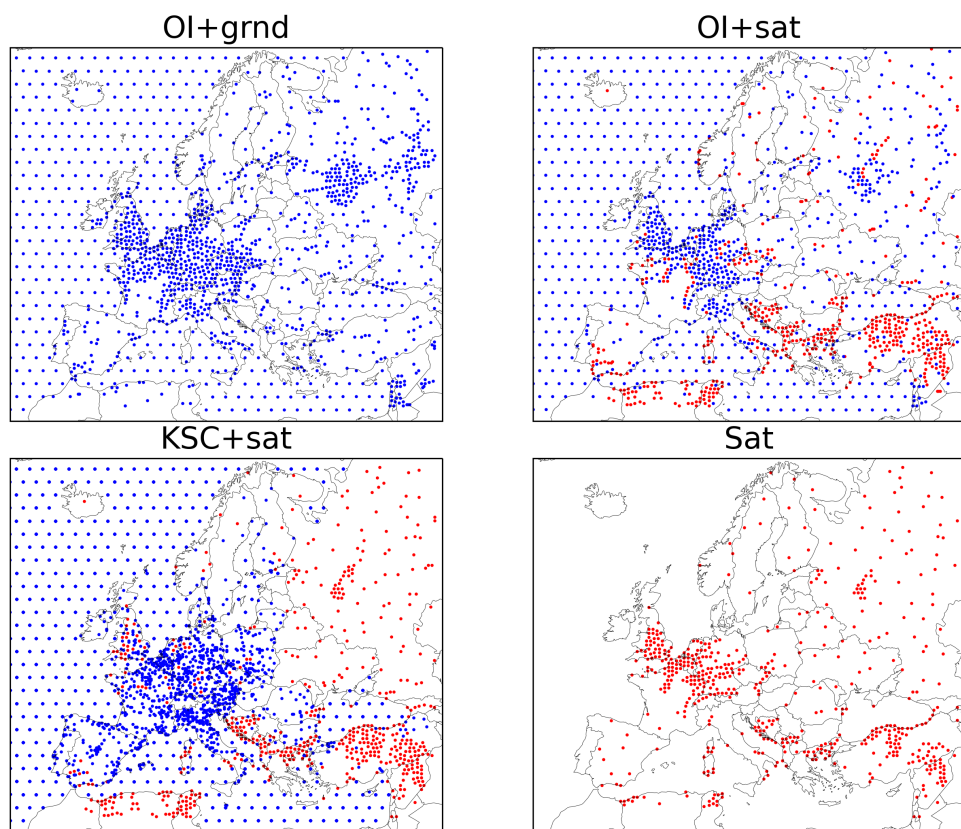


Figure B4. Configuration of the O₃ observation networks considered in Sect. 5.4 for 10 March 2016. Each dot represents an observation site with the ground observations in blue and satellite observations in red.

B.3. Initial value gradient and concentration difference for 17 March 2016

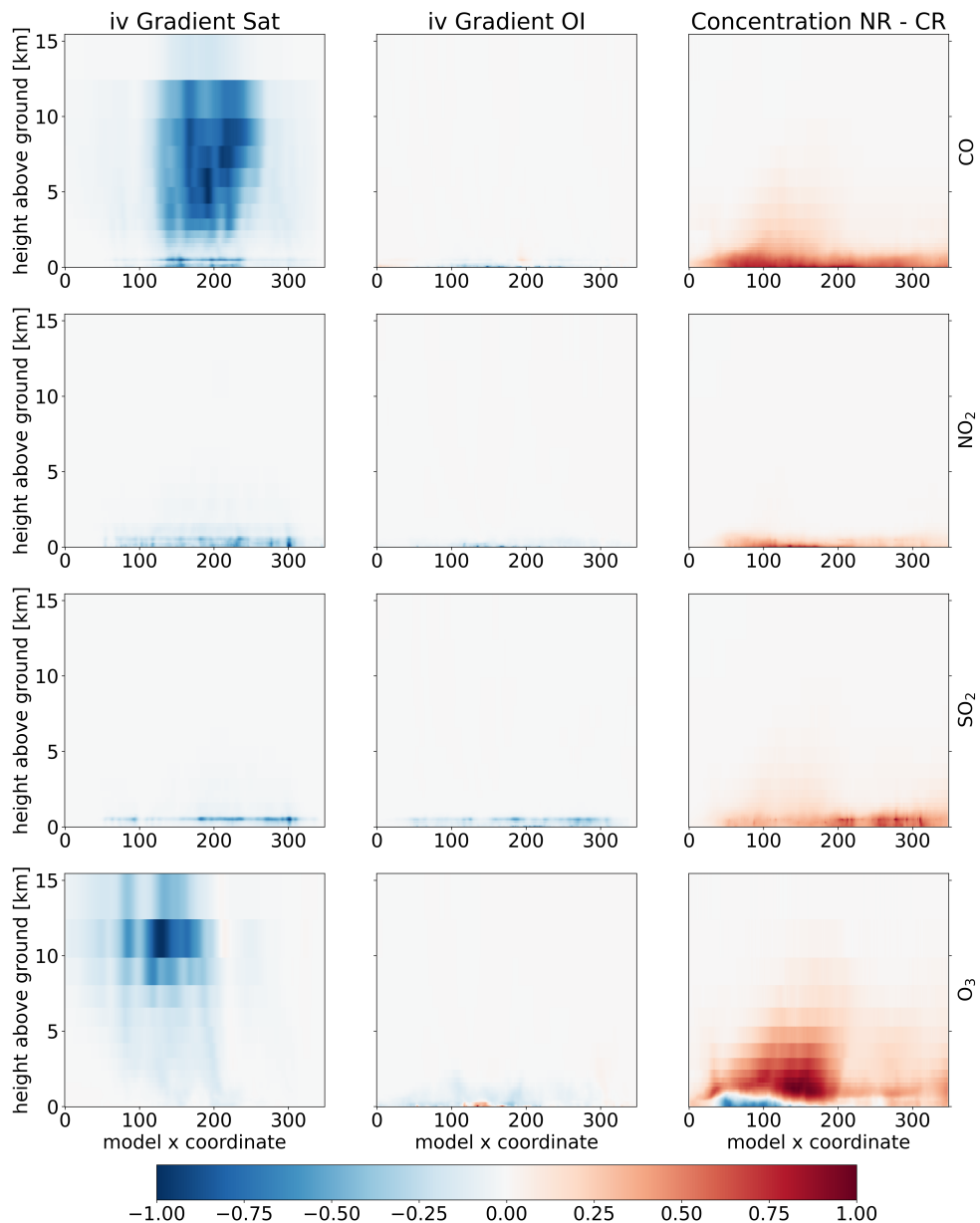


Figure B5. Normalized initial value gradient averaged along model y-direction for the Sat (left) and the OI+grnd assimilation run (center) and the normalized and along y-direction averaged concentrations difference between the NR and CR (right). The values are shown for 17 March 2016 and for the species CO, NO₂, SO₂, and O₃ in descending order. The y-axis shows the height above the ground in [km] and the x-axis the model x-coordinate.

Bibliography

- Ackermann, I. J., Hass, H., Memmesheimer, M., Ebel, A., Binkowski, F. S., and Shankar, U.: Modal aerosol dynamics model for Europe: development and first applications, *Atmos. Environ.*, 32, 2981–2999, [https://doi.org/10.1016/S1352-2310\(98\)00006-5](https://doi.org/10.1016/S1352-2310(98)00006-5), 1998.
- Alvarado, M. J., McVey, A. E., Hegarty, J. D., Cross, E. S., Hasenkopf, C. A., Lynch, R., Kennelly, E. J., Onasch, T. B., Awe, Y., Sanchez-Triana, E., and Kleiman, G.: Evaluating the use of satellite observations to supplement ground-level air quality data in selected cities in low- and middle-income countries, *Atmos. Environ.*, 218, 117016, <https://doi.org/10.1016/j.atmosenv.2019.117016>, 2019.
- Amann, M., Klimont, Z., and Wagner, F.: Regional and Global Emissions of Air Pollutants: Recent Trends and Future Scenarios, *Annu. Rev. Env. Resour.*, 38, 31–55, <https://doi.org/10.1146/annurev-environ-052912-173303>, 2013.
- Arnold, C. P. and Dey, C. H.: Observing-Systems Simulation Experiments: Past, Present, and Future, *Bull. Amer. Meteorol. Soc.*, 67, 687 – 695, [https://doi.org/10.1175/1520-0477\(1986\)067<0687:0SSEPP>2.0.CO;2](https://doi.org/10.1175/1520-0477(1986)067<0687:0SSEPP>2.0.CO;2), 1986.
- Atkinson, R.: Atmospheric chemistry of VOCs and NO_x, *Atmos. Environ.*, 34, 2063–2101, [https://doi.org/10.1016/S1352-2310\(99\)00460-4](https://doi.org/10.1016/S1352-2310(99)00460-4), 2000.
- Bouttier, F. and Courtier, P.: Data assimilation concepts and methods March 1999, Meteorological training course lecture series. ECMWF, 718, 59, URL https://twister.caps.o.u.edu/0BAN2019/Assim_concepts.pdf, last access: 16 January 2026, 2002.
- Bouttier, F. and Kelly, G.: Observing-system experiments in the ECMWF 4D-Var data assimilation system, *Quart. J. Roy. Meteorol. Soc.*, 127, 1469–1488, <https://doi.org/10.1002/qj.49712757419>, 2001.
- Buehner, M., Houtekamer, P. L., Charette, C., Mitchell, H. L., and He, B.: Intercomparison of Variational Data Assimilation and the Ensemble Kalman Filter for Global Deterministic NWP Part I: Description and Single-Observation Experiments, *Mon. Weather Rev.*, 138, 1550 – 1566, <https://doi.org/10.1175/2009MWR3157.1>, 2010a.
- Buehner, M., Houtekamer, P. L., Charette, C., Mitchell, H. L., and He, B.: Intercomparison of Variational Data Assimilation and the Ensemble Kalman Filter for Global Deterministic

-
- NWP. Part II: One-Month Experiments with Real Observations, *Mon. Weather Rev.*, 138, 1567 – 1586, <https://doi.org/10.1175/2009MWR3158.1>, 2010b.
- Cardinali, C.: Monitoring the observation impact on the short-range forecast, *Quart. J. Roy. Meteorol. Soc.*, 135, 239–250, <https://doi.org/10.1002/qj.366>, 2009.
- Chen, X., You, W., Duan, R., Gu, M., Zang, Z., and Luo, J.: Optimizing air quality forecasting with OSSE: Assessing the assimilation of aerosol species data via multi-resolution observation network design in the WRF-Chem and 3DVAR frameworks, *Atmos. Res.*, 325, 108 213, <https://doi.org/10.1016/j.atmosres.2025.108213>, 2025.
- Claeyman, M., Attié, J.-L., Peuch, V.-H., El Amraoui, L., Lahoz, W. A., Josse, B., Joly, M., Barré, J., Ricaud, P., Massart, S., Piacentini, A., von Clarmann, T., Höpfner, M., Orphal, J., Flaud, J.-M., and Edwards, D. P.: A thermal infrared instrument onboard a geostationary platform for CO and O₃ measurements in the lowermost troposphere: Observing System Simulation Experiments (OSSE), *Atmos. Meas. Tech.*, 4, 1637–1661, <https://doi.org/10.5194/amt-4-1637-2011>, 2011.
- Colette, A., Collin, G., Besson, F., Blot, E., Guidard, V., Meleux, F., Royer, A., Petiot, V., Miller, C., Fermond, O., Jeant, A., Adani, M., Arteta, J., Benedictow, A., Bergström, R., Bowdalo, D., Brandt, J., Briganti, G., Carvalho, A. C., Christensen, J. H., Couvidat, F., D’Elia, I., D’Isidoro, M., Denier van der Gon, H., Descombes, G., Di Tomaso, E., Douros, J., Escribano, J., Eskes, H., Fagerli, H., Fatahi, Y., Flemming, J., Frieze, E., Frohn, L., Gauss, M., Geels, C., Guarnieri, G., Guevara, M., Guion, A., Guth, J., Hänninen, R., Hansen, K., Im, U., Janssen, R., Jeoffrion, M., Joly, M., Jones, L., Jorba, O., Kadantsev, E., Kahnert, M., Kaminski, J. W., Kouznetsov, R., Kranenburg, R., Kuenen, J., Lange, A. C., Langner, J., Lannuque, V., Macchia, F., Manders, A., Mircea, M., Nyiri, A., Olid, M., Pérez García-Pando, C., Palamarchuk, Y., Piersanti, A., Raux, B., Razinger, M., Robertson, L., Segers, A., Schaap, M., Siljamo, P., Simpson, D., Sofiev, M., Stangel, A., Struzewska, J., Tena, C., Timmermans, R., Tsikerdekis, T., Tsyro, S., Tyuryakov, S., Ung, A., Uppstu, A., Valdebenito, A., van Velthoven, P., Vitali, L., Ye, Z., Peuch, V.-H., and Rouil, L.: Copernicus Atmosphere Monitoring Service – Regional Air Quality Production System v1.0, *Geosci. Model Dev.*, 18, 6835–6883, <https://doi.org/10.5194/gmd-18-6835-2025>, 2025.
- Courtier, P., Thépaut, J.-N., and Hollingsworth, A.: A strategy for operational implementation of 4D-Var, using an incremental approach, *Quart. J. Roy. Meteorol. Soc.*, 120, 1367–1387, <https://doi.org/10.1002/qj.49712051912>, 1994.
- Crippa, M., Solazzo, E., Huang, G., Guizzardi, D., Koffi, E., Muntean, M., Schieberle, C., Friedrich, R., and Janssens-Maenhout, G.: High resolution temporal profiles in the Emissions Database for Global Atmospheric Research, *Sci. Data*, 7, <https://doi.org/10.1038/s41597-020-0462-2>, 2020.

-
- De Souza Fernandes Duarte, E., Franke, P., Lange, A. C., Friese, E., Juliano da Silva Lopes, E., João da Silva, J., Souza dos Reis, J., Landulfo, E., Santos e Silva, C. M., Elbern, H., and Hoelzemann, J. J.: Evaluation of atmospheric aerosols in the metropolitan area of São Paulo simulated by the regional EURAD-IM model on high-resolution, *Atmospheric Pollution Research*, 12, 451–469, <https://doi.org/10.1016/j.apr.2020.12.006>, 2021.
- Deroubaix, A., Hoelzemann, J. J., Ynoue, R. Y., Toledo de Almeida Albuquerque, T., Alves, R. C., de Fatima Andrade, M., Andreão, W. L., Bouarar, I., de Souza Fernandes Duarte, E., Elbern, H., Franke, P., Lange, A. C., Lichtig, P., Lugon, L., Martins, L. D., de Arruda Moreira, G., Pedruzzi, R., Rosario, N., and Brasseur, G.: Intercomparison of Air Quality Models in a Megacity: Toward an Operational Ensemble Forecasting System for São Paulo, *Geophys. Res. Atmos.*, 129, e2022JD038179, <https://doi.org/10.1029/2022JD038179>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2022JD038179>, e2022JD038179 2022JD038179, 2024.
- ECMWF: IFS Documentation CY49R1 - Part II: Data Assimilation, in: IFS Documentation CY49R1, IFS Documentation, chap. 2, ECMWF, <https://doi.org/10.21957/105cb1333c>, 2024.
- Edwards, D. P., Arellano Jr., A. F., and Deeter, M. N.: A satellite observation system simulation experiment for carbon monoxide in the lowermost troposphere, *Geophys. Res. Atmos.*, 114, <https://doi.org/10.1029/2008JD011375>, 2009.
- Elbern, H., Schmidt, H., and Ebel, A.: Variational data assimilation for tropospheric chemistry modeling, *Geophys. Res. Atmos.*, 102, 15 967–15 985, <https://doi.org/10.1029/97JD01213>, 1997.
- Elbern, H., Strunk, A., Schmidt, H., and Talagrand, O.: Emission rate and chemical state estimation by 4-dimensional variational inversion, *Atmos. Chem. Phys.*, 7, 3749–3769, <https://doi.org/10.5194/acp-7-3749-2007>, 2007.
- Elguindi, N., Granier, C., Stavrakou, T., Darras, S., Bauwens, M., Cao, H., Chen, C., Denier van der Gon, H. A. C., Dubovik, O., Fu, T. M., Henze, D. K., Jiang, Z., Keita, S., Kuenen, J. J. P., Kurokawa, J., Lioussé, C., Miyazaki, K., Müller, J.-F., Qu, Z., Solmon, F., and Zheng, B.: Intercomparison of Magnitudes and Trends in Anthropogenic Surface Emissions From Bottom-Up Inventories, Top-Down Estimates, and Emission Scenarios, *Earth's Future*, 8, e2020EF001 520, <https://doi.org/https://doi.org/10.1029/2020EF001520>, 2020.
- Erraji, H., Franke, P., Lampert, A., Schuldt, T., Tillmann, R., Wahner, A., and Lange, A. C.: The potential of drone observations to improve air quality predictions by 4D-Var, *Atmos. Chem. Phys.*, 24, 13 913–13 934, <https://doi.org/10.5194/acp-24-13913-2024>, 2024.

-
- ESA: ESA Sentinel-4 - Sentinel Online, <https://sentinels.copernicus.eu/web/sentinel/missions/sentinel-4>, last access: 18 October 2025, 2025a.
- ESA: ESA Sentinel-5P - Sentinel Online, <https://sentinels.copernicus.eu/web/sentinel/copernicus/sentinel-5p>, last access: 18 October 2025, 2025b.
- European Environment Agency: Harm to human health from air pollution in Europe: burden of disease status, 2024, <https://doi.org/10.2800/3950756>, 2024.
- Evensen, G.: The ensemble Kalman filter for combined state and parameter estimation, *IEEE Control Syst. Mag.*, 29, 83–104, <https://doi.org/10.1109/MCS.2009.932223>, 2009.
- Evensen, G., Vossepoel, F. C., and van Leeuwen, P. J.: Strong-Constraint 4DVar, in: *Data Assimilation Fundamentals: A Unified Formulation of the State and Parameter Estimation Problem*, pp. 35–47, Springer International Publishing, Cham, https://doi.org/10.1007/978-3-030-96709-3_4, 2022.
- Fisher, M.: Estimation of entropy reduction and degrees of freedom for signal for large variational analysis systems, ECMWF, URL <https://www.ecmwf.int/sites/default/files/elibrary/2003/9402-estimation-entropy-reduction-and-degrees-freedom-signal-large-variational-analysis-systems.pdf>, last access: 16 January 2026, 2003.
- Franke, P., Lange, A. C., and Elbern, H.: Particle-filter-based volcanic ash emission inversion applied to a hypothetical sub-Plinian Eyjafjallajökull eruption using the Ensemble for Stochastic Integration of Atmospheric Simulations (ESIAS-chem) version 1.0, *Geosci. Model Dev.*, 15, 1037–1060, <https://doi.org/10.5194/gmd-15-1037-2022>, 2022.
- Franke, P., Lange, A. C., Steffens, B., Pozzer, A., Wahner, A., and Kiendler-Scharr, A.: European air quality in view of the WHO 2021 guideline levels: Effect of emission reductions on air pollution exposure, *Elementa: Science of the Anthropocene*, 12, 00127, <https://doi.org/10.1525/elementa.2023.00127>, 2024.
- Gama, C., Ribeiro, I., Lange, A., Vogel, A., Ascenso, A., Seixas, V., Elbern, H., Borrego, C., Friese, E., and Monteiro, A.: Performance assessment of CHIMERE and EURAD-IM’ dust modules, *Atmospheric Pollution Research*, 10, 1336–1346, <https://doi.org/10.1016/j.apr.2019.03.005>, 2019.
- Gaubert, B., Emmons, L. K., Raeder, K., Tilmes, S., Miyazaki, K., Arellano Jr., A. F., Elguindi, N., Granier, C., Tang, W., Barré, J., Worden, H. M., Buchholz, R. R., Edwards, D. P., Franke, P., Anderson, J. L., Saunio, M., Schroeder, J., Woo, J.-H., Simpson, I. J., Blake, D. R., Meinardi, S., Wennberg, P. O., Crounse, J., Teng, A., Kim, M., Dickerson, R. R., He, H., Ren, X., Pusede, S. E., and Diskin, G. S.: Correcting model biases of CO in East Asia: impact on

-
- oxidant distributions during KORUS-AQ, *Atmos. Chem. Phys.*, 20, 14 617–14 647, <https://doi.org/10.5194/acp-20-14617-2020>, 2020.
- Geiger, H., Barnes, I., Bejan, I., Benter, T., and Spittler, M.: The tropospheric degradation of isoprene: an updated module for the regional atmospheric chemistry mechanism, *Atmos. Environ.*, 37, 1503–1519, [https://doi.org/10.1016/S1352-2310\(02\)01047-6](https://doi.org/10.1016/S1352-2310(02)01047-6), 2003.
- Goris, N. and Elbern, H.: Singular vector-based targeted observations of chemical constituents: description and first application of the EURAD-IM-SVA v1.0, *Geosci. Model Dev.*, 8, 3929–3945, <https://doi.org/10.5194/gmd-8-3929-2015>, 2015.
- Haben, S., Lawless, A., and Nichols, N.: Conditioning and preconditioning of the variational data assimilation problem, *Comput. Fluids*, 46, 252–256, <https://doi.org/10.1016/j.compfluid.2010.11.025>, 2011.
- Hermanns, A., Lange, A. C., Kowalski, J., Fuchs, H., and Franke, P.: Data clustering to optimise the representativity of observational data in air quality data assimilation: a case study with EURAD-IM (version 5.9.1 DA), *Geosci. Model Dev.*, 18, 9417–9432, <https://doi.org/10.5194/gmd-18-9417-2025>, 2025.
- Hersbach, H., Cobb, A., Kaandrop, M., Poli, P., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., and Thépaut, J.-N.: ERA5 hourly time-series data on single levels from 1940 to present, Copernicus Climate Change Service (C3S) Climate Data Store (CDS), 2025.
- Houtekamer, P. and Mitchell, H. L.: Practical ensemble data assimilation, in: *Seminar on Recent developments in data assimilation for atmosphere and ocean*, vol. 8, pp. 193–210, URL <https://www.ecmwf.int/sites/default/files/elibrary/2003/10044-practical-ensemble-data-assimilation.pdf>, last access: 16 January 2026, 2003.
- Hu, Y., Li, Y., Ma, X., Liang, Y., You, W., Pan, X., and Zang, Z.: The optimization of SO₂ emissions by the 4DVAR and EnKF methods and its application in WRF-Chem, *Sci. Total Environ.*, 888, 163 796, <https://doi.org/10.1016/j.scitotenv.2023.163796>, 2023.
- Inness, A., Ades, M., Agustí-Panareda, A., Barré, J., Benedictow, A., Blechschmidt, A.-M., Dominguez, J. J., Engelen, R., Eskes, H., Flemming, J., Huijnen, V., Jones, L., Kipling, Z., Massart, S., Parrington, M., Peuch, V.-H., Razinger, M., Remy, S., Schulz, M., and Suttie, M.: The CAMS reanalysis of atmospheric composition, *Atmos. Chem. Phys.*, 19, 3515–3556, <https://doi.org/10.5194/acp-19-3515-2019>, 2019.
- Joly, M. and Peuch, V.-H.: Objective classification of air quality monitoring sites over Europe, *Atmos. Environ.*, 47, 111–123, <https://doi.org/10.1016/j.atmosenv.2011.11.025>, 2012.

-
- Jülich Supercomputing Centre: JUWELS Cluster and Booster: Exascale Pathfinder with Modular Supercomputing Architecture at Juelich Supercomputing Centre, *Journal of large-scale research facilities*, 7, <https://doi.org/10.17815/jlsrf-7-183>, 2021.
- Kalnay, E., Li, H., Miyoshi, T., Yang, S.-C., and Ballabrera-Poy, J.: 4-D-Var or ensemble Kalman filter?, *Tellus A*, 59, 758–773, <https://doi.org/10.1111/j.1600-0870.2007.00261.x>, 2007.
- Khalil, M. and Rasmussen, R.: The global cycle of carbon monoxide: Trends and mass balance, *Chemosphere*, 20, 227–242, [https://doi.org/10.1016/0045-6535\(90\)90098-E](https://doi.org/10.1016/0045-6535(90)90098-E), 1990.
- Krupa, S. V. and Manning, W. J.: Atmospheric ozone: Formation and effects on vegetation, *Environ. Pollut.*, 50, 101–137, [https://doi.org/10.1016/0269-7491\(88\)90187-X](https://doi.org/10.1016/0269-7491(88)90187-X), 1988.
- Kuenen, J., Dellaert, S., Visschedijk, A., Jalkanen, J.-P., Super, I., and Denier van der Gon, H.: CAMS-REG-v4: a state-of-the-art high-resolution European emission inventory for air quality modelling, *Earth Syst. Sci. Data*, 14, 491–515, <https://doi.org/10.5194/essd-14-491-2022>, 2022.
- Kurokawa, J., Yumimoto, K., Uno, I., and Ohara, T.: Adjoint inverse modeling of NO_x emissions over eastern China using satellite observations of NO₂ vertical column densities, *Atmos. Environ.*, 43, 1878–1887, <https://doi.org/10.1016/j.atmosenv.2008.12.030>, 2009.
- Lahoz, W., Khattatov, B., and Menard, R., eds.: *Data Assimilation - Making Sense of Observations*, Springer Berlin, Heidelberg, 1 edn., <https://doi.org/10.1007/978-3-540-74703-1>, 2010.
- Lahoz, W. A. and Schneider, P.: Data assimilation: making sense of Earth Observation, *Front. Environ. Sci.*, Volume 2 - 2014, <https://doi.org/10.3389/fenvs.2014.00016>, 2014.
- Lange, A. C., Franke, P., Backes, P., Elbern, H., et al.: Immissionsseitige Bewertung der Luftschadstoff-Emissionen einzelner Quellen und Anpassung der nationalen Emissionsdaten zur Beurteilung der Luftqualität, Umweltbundesamt (UBA), 149/2023, <https://doi.org/10.60810/openumwelt-6348>, 2023.
- Langland, R. H. and Baker, N. L.: Estimation of observation impact using the NRL atmospheric variational data assimilation adjoint system, *Tellus A*, 56, 189–201, <https://doi.org/10.3402/tellusa.v56i3.14413>, 2004.
- Lee, C., Martin, R. V., van Donkelaar, A., Lee, H., Dickerson, R. R., Hains, J. C., Krotkov, N., Richter, A., Vinnikov, K., and Schwab, J. J.: SO₂ emissions and lifetimes: Estimates

-
- from inverse modeling using in situ and global, space-based (SCIAMACHY and OMI) observations, *Geophys. Res. Atmos.*, 116, <https://doi.org/10.1029/2010JD014758>, 2011.
- Lee, K. K., Spath, N., Miller, M. R., Mills, N. L., and Shah, A. S.: Short-term exposure to carbon monoxide and myocardial infarction: A systematic review and meta-analysis, *Environ. Int.*, 143, 105 901, <https://doi.org/10.1016/j.envint.2020.105901>, 2020.
- Liu, D. C. and Nocedal, J.: On the limited memory BFGS method for large scale optimization, *Math. Program.*, 45, 503–528, <https://doi.org/10.1007/BF01589116>, 1989.
- Liu, F., Beirle, S., Zhang, Q., Dörner, S., He, K., and Wagner, T.: NO_x lifetimes and emissions of cities and power plants in polluted background estimated by satellite observations, *Atmos. Chem. Phys.*, 16, 5283–5298, <https://doi.org/10.5194/acp-16-5283-2016>, 2016.
- Liu, X., Mizzi, A. P., Anderson, J. L., Fung, I. Y., and Cohen, R. C.: Assimilation of satellite NO₂ observations at high spatial resolution using OSSEs, *Atmos. Chem. Phys.*, 17, 7067–7081, <https://doi.org/10.5194/acp-17-7067-2017>, 2017.
- Liu, Z.-Q. and Rabier, F.: The interaction between model resolution, observation resolution and observation density in data assimilation: A one-dimensional study, *Quart. J. Roy. Meteorol. Soc.*, 128, 1367–1386, <https://doi.org/10.1256/003590002320373337>, 2002.
- Lorenc, A. C.: Analysis methods for numerical weather prediction, *Quart. J. Roy. Meteorol. Soc.*, 112, 1177–1194, <https://doi.org/10.1002/qj.49711247414>, 1986.
- Marécal, V., Peuch, V.-H., Andersson, C., Andersson, S., Arteta, J., Beekmann, M., Benedictow, A., Bergström, R., Bessagnet, B., Cansado, A., Chéroux, F., Colette, A., Coman, A., Curier, R. L., Denier van der Gon, H. A. C., Drouin, A., Elbern, H., Emili, E., Engelen, R. J., Eskes, H. J., Foret, G., Friese, E., Gauss, M., Giannaros, C., Guth, J., Joly, M., Jau-mouillé, E., Josse, B., Kadygrov, N., Kaiser, J. W., Krajsek, K., Kuenen, J., Kumar, U., Liora, N., Lopez, E., Malherbe, L., Martinez, I., Melas, D., Meleux, F., Menut, L., Moinat, P., Morales, T., Parmentier, J., Piacentini, A., Plu, M., Poupkou, A., Queguiner, S., Robertson, L., Rouil, L., Schaap, M., Segers, A., Sofiev, M., Tarasson, L., Thomas, M., Timmermans, R., Valdebenito, A., van Velthoven, P., van Versendaal, R., Vira, J., and Ung, A.: A regional air quality forecasting system over Europe: the MACC-II daily ensemble production, *Geosci. Model Dev.*, 8, 2777–2813, <https://doi.org/10.5194/gmd-8-2777-2015>, 2015.
- Masutani, M., Schlatter, T. W., Errico, R. M., Stoffelen, A., Andersson, E., Lahoz, W., Woollen, J. S., Emmitt, G. D., Riishøjgaard, L.-P., and Lord, S. J.: Observing System Simulation Experiments, in: *Data Assimilation: Making Sense of Observations*, edited by Lahoz, W.,

-
- Khattatov, B., and Menard, R., pp. 647–679, Springer Berlin Heidelberg, Berlin, Heidelberg, https://doi.org/10.1007/978-3-540-74703-1_24, 2010.
- McRae, G. J., Goodin, W. R., and Seinfeld, J. H.: Numerical solution of the atmospheric diffusion equation for chemically reacting flows, *J. Comput. Phys.*, 45, 1–42, [https://doi.org/10.1016/0021-9991\(82\)90101-2](https://doi.org/10.1016/0021-9991(82)90101-2), 1982.
- Ochotta, T., Gebhardt, C., Saupe, D., and Wergen, W.: Adaptive thinning of atmospheric observations in data assimilation with vector quantization and filtering methods, *Quart. J. Roy. Meteorol. Soc.*, 131, 3427–3437, <https://doi.org/10.1256/qj.05.94>, 2005.
- Orellano, P., Reynoso, J., and Quaranta, N.: Short-term exposure to sulphur dioxide (SO₂) and all-cause and respiratory mortality: A systematic review and meta-analysis, *Environ. Int.*, 150, 106434, <https://doi.org/10.1016/j.envint.2021.106434>, 2021.
- Petroff, A. and Zhang, L.: Development and validation of a size-resolved particle dry deposition scheme for application in aerosol transport models, *Geosci. Model Dev.*, 3, 753–769, <https://doi.org/10.5194/gmd-3-753-2010>, 2010.
- Pleim, J. E. and Chang, J. S.: A non-local closure model for vertical mixing in the convective boundary layer, *Atmos. Environ. Part A. General Topics*, 26, 965–981, [https://doi.org/10.1016/0960-1686\(92\)90028-J](https://doi.org/10.1016/0960-1686(92)90028-J), 1992.
- Quesada-Ruiz, S., Attié, J.-L., Lahoz, W. A., Abida, R., Ricaud, P., El Amraoui, L., Zbinden, R., Piacentini, A., Joly, M., Eskes, H., Segers, A., Curier, L., de Haan, J., Kujanpää, J., Oude Nijhuis, A. C. P., Tamminen, J., Timmermans, R., and Veefkind, P.: Benefit of ozone observations from Sentinel-5P and future Sentinel-4 missions on tropospheric composition, *Atmos. Meas. Tech.*, 13, 131–152, <https://doi.org/10.5194/amt-13-131-2020>, 2020.
- Rabier, F. and Liu, Z.: Variational data assimilation: theory and overview, in: *Proc. ECMWF Seminar on Recent Developments in Data Assimilation for Atmosphere and Ocean*, Reading, UK, pp. 29–43, URL <https://www.ecmwf.int/sites/default/files/elibrary/2003/11805-variational-data-assimiltion-theory-and-overview.pdf>, last access: 16 January 2026, 2003.
- Roselle, S. J. and Binkowski, F. S.: Cloud dynamics and chemistry, in: *Science algorithms of the EPA Models-3 Community multiscale air quality (CMAQ) modeling system*, edited by Byun, D. W. and Ching, J. K. S., chap. 11, Environmental Protection Agency: EPA/600/R99/030, URL https://cmasceneter.org/cmaq/science_documentation/pdf/ch11.pdf, last access: 16 January 2026, 1999.

-
- Schindlbacher, S., Matthews, B., and Ullrich, B.: Uncertainties and recalculations of emission inventories submitted under CLRTAP, https://www.ceip.at/fileadmin/inhalte/ceip/00_pdf_other/2021/uncertainties_and_recalculations_of_emission_inventories_submitted_under_clrtap.pdf, last access: 28 Dezember 2025, 2021.
- Seinfeld, J. H. and Pandis, S. N.: Atmospheric Chemistry and Physics From Air Pollution to Climate Change, New York Academy of Sciences Series, John Wiley & Sons, Incorporated, third edn., ISBN 9781119221166, 2016.
- Skamarock, W., Klemp, J., Dudhia, J., Gill, D., Barker, D., Duda, M., Huang, X.-Y., Wang, W., and Powers, J.: A Description of the Advanced Research WRF Version 3, NCAR Technical Note NCAR/TN-475+STR. June 2008. Mesoscale and Microscale Meteorology Division. National Center for Atmospheric Research. Boulder, <https://doi.org/10.13140/RG.2.1.2310.6645>, 2008.
- Solazzo, E., Crippa, M., Guizzardi, D., Muntean, M., Choulga, M., and Janssens-Maenhout, G.: Uncertainties in the Emissions Database for Global Atmospheric Research (EDGAR) emission inventory of greenhouse gases, *Atmos. Chem. Phys.*, 21, 5655–5683, <https://doi.org/10.5194/acp-21-5655-2021>, 2021.
- Stockwell, W. R., Kirchner, F., Kuhn, M., and Seefeld, S.: A new mechanism for regional atmospheric chemistry modeling, *Geophys. Res. Atmos.*, 102, 25 847–25 879, <https://doi.org/10.1029/97JD00849>, 1997.
- Talagrand, O.: Assimilation of Observations, an Introduction (gtSpecial IssueData Assimilation in Meteorology and Oceanography: Theory and Practice), *J. Meteorol. Soc. Jpn.*, 75, 191–209, <https://doi.org/10.2151/jmsj1965.75.1B.191>, 1997.
- Tillmann, R., Gkatzelis, G. I., Rohrer, F., Winter, B., Wesolek, C., Schuldt, T., Lange, A. C., Franke, P., Friese, E., Decker, M., Wegener, R., Hundt, M., Aseev, O., and Kiendler-Scharr, A.: Air quality observations onboard commercial and targeted Zeppelin flights in Germany – a platform for high-resolution trace-gas and aerosol measurements within the planetary boundary layer, *Atmos. Meas. Tech.*, 15, 3827–3842, <https://doi.org/10.5194/amt-15-3827-2022>, 2022.
- Timmermans, R., Lahoz, W., Attié, J.-L., Peuch, V.-H., Curier, R., Edwards, D., Eskes, H., and Builtjes, P.: Observing System Simulation Experiments for air quality, *Atmos. Environ.*, 115, 199–213, <https://doi.org/10.1016/j.atmosenv.2015.05.032>, 2015.
- Timmermans, R., Segers, A., Curier, L., Abida, R., Attié, J.-L., El Amraoui, L., Eskes, H., de Haan, J., Kujanpää, J., Lahoz, W., Oude Nijhuis, A., Quesada-Ruiz, S., Ricaud, P., Veefkind, P., and Schaap, M.: Impact of synthetic space-borne NO₂ observations from the

-
- Sentinel-4 and Sentinel-5P missions on tropospheric NO₂ analyses, *Atmos. Chem. Phys.*, 19, 12 811–12 833, <https://doi.org/10.5194/acp-19-12811-2019>, 2019.
- Trémolet, Y. and Fisher, M.: Weak constraint 4D-Var, *ECMWF Newsletter*, pp. 12–16, <https://doi.org/10.21957/21fybybt>, 2010.
- Umweltbundesamt Deutschland, ed.: Submission under the United Nations Framework Convention on Climate Change and the Kyoto Protocol 2019, Umweltbundesamt, Dessau-Roßlau, Germany, <https://doi.org/10.60810/openumwelt-2978>, 2019.
- Walcek, C. J.: Minor flux adjustment near mixing ratio extremes for simplified yet highly accurate monotonic calculation of tracer advection, *Geophys. Res. Atmos.*, 105, 9335–9348, <https://doi.org/10.1029/1999JD901142>, 2000.
- Wang, C., An, X., Hou, Q., Sun, Z., Li, Y., and Li, J.: Development of four-dimensional variational assimilation system based on the GRAPES–CUACE adjoint model (GRAPES–CUACE-4D-Var V1.0) and its application in emission inversion, *Geosci. Model Dev.*, 14, 337–350, <https://doi.org/10.5194/gmd-14-337-2021>, 2021.
- Wang, C., An, X., Zhao, D., Sun, Z., Jiang, L., Li, J., and Hou, Q.: Development of GRAPES–CUACE adjoint model version 2.0 and its application in sensitivity analysis of ozone pollution in north China, *Sci. Total Environ.*, 826, 153 879, <https://doi.org/10.1016/j.scitotenv.2022.153879>, 2022.
- World Health Organization: Review of evidence on health aspects of air pollution: REVI-HAAP project: technical report, World Health Organization, URL <https://www.who.int/europe/publications/i/item/WHO-EURO-2013-4101-43860-61757>, last accessed: 16 January 2026, 2013.
- World Health Organization: WHO global air quality guidelines: particulate matter (PM_{2.5} and PM₁₀), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide, World Health Organization, URL <https://www.who.int/publications/i/item/9789240034228>, last accessed: 16 January 2026, 2021.
- Yumimoto, K.: Impacts of geostationary satellite measurements on CO forecasting: An observing system simulation experiment with GEOS-Chem/LETKF data assimilation system, *Atmos. Environ.*, 74, 123–133, <https://doi.org/10.1016/j.atmosenv.2013.03.032>, 2013.
- Zhang, L., Brook, J. R., and Vet, R.: A revised parameterization for gaseous dry deposition in air-quality models, *Atmos. Chem. Phys.*, 3, 2067–2082, <https://doi.org/10.5194/acp-3-2067-2003>, 2003.

-
- Zhang, L., Constantinescu, E., Sandu, A., Tang, Y., Chai, T., Carmichael, G., Byun, D., and Olaguer, E.: An adjoint sensitivity analysis and 4D-Var data assimilation study of Texas air quality, *Atmos. Environ.*, 42, 5787–5804, <https://doi.org/10.1016/j.atmosenv.2008.03.048>, 2008.
- Zhang, L., Jacob, D. J., Liu, X., Logan, J. A., Chance, K., Eldering, A., and Bojkov, B. R.: Intercomparison methods for satellite measurements of atmospheric composition: application to tropospheric ozone from TES and OMI, *Atmos. Chem. Phys.*, 10, 4725–4739, <https://doi.org/10.5194/acp-10-4725-2010>, 2010.
- Zhang, X., Jones, D. B. A., Keller, M., Walker, T. W., Jiang, Z., Henze, D. K., Worden, H. M., Bourassa, A. E., Degenstein, D. A., and Rochon, Y. J.: Quantifying Emissions of CO and NO_x Using Observations From MOPITT, OMI, TES, and OSIRIS, *Geophys. Res. Atmos.*, 124, 1170–1193, <https://doi.org/10.1029/2018JD028670>, 2019.
- Zhang, Y., Jacob, D. J., Maasakkers, J. D., Sulprizio, M. P., Sheng, J.-X., Gautam, R., and Worden, J.: Monitoring global tropospheric OH concentrations using satellite observations of atmospheric methane, *Atmos. Chem. Phys.*, 18, 15 959–15 973, <https://doi.org/10.5194/acp-18-15959-2018>, 2018.

Personal acknowledgments

First of all, I would like to particularly thank Philipp Franke for his outstanding supervision and support. He always found time to discuss ideas, and to provide me with his insight and encouragement. I am grateful for the many hours we spend going over my sheets of talking points; this was invaluable to me.

I would like to thank Anne Caroline Lange. She was available to provide additional angles to ideas and to share her expertise whenever my arguments lead to a seemingly dead end. I have not only profited from her scientific knowledge and personal support but also our interesting discussions about the fermentation of food.

I am grateful to Prof. Hendrik Fuchs for his supervision, his support, and for always being approachable. His attention to detail and him providing his non-modeling perspective were invaluable.

I want to thank Prof. Julia Kowalski for taking an interest in my work, our fruitful discussions, and for providing her scientific insight.

Thank you to Prof. Anke Nölscher for accepting to act as referee and upholding the homely atmosphere of the ICE-3.

I want to express my thanks to my colleagues and the members of the modellers group who have made my stay at the Forschungszentrum Jülich a very enjoyable one. In addition to the professional discussions, our lunch meetings and accompanying conversations have always been a highlight of my days.

I am deeply grateful to my parents Willi and Gabi for their love, their encouragement, and their endless support, which enabled me to pursue this incredible journey. Many thanks to my brother Christopher for always believing in me.

Thanks to my friends who provided me with much needed distractions and support whenever I needed it. Special thanks to Andy for being someone I can always rely on, and for making our holidays a highlight.

I would like to express my deepest gratitude to Eva for always motivating and believing in me, and for always being there for me whenever I needed her.

Last but certainly not least, I am deeply grateful to Natascha for her unwavering support, understanding, and rock-solid trust in me.

Band / Volume 714

Zinc-Based Catalysts for the Electrochemical CO₂ Reduction

I. Stamatelos (2026), vii, 136 pp

ISBN: 978-3-95806-932-9 (Print)

ISBN: 978-3-95806-933-6 (E-Book)

Band / Volume 715

**Charakterisierung innovativer keramischer Materialien für
Elektrolysezellen**

L. P. Wehner (2026), xiv, 110 pp

ISBN: 978-3-95806-936-7 (Print)

ISBN: 978-3-95806-937-4 (E-Book)

Band / Volume 716

**Process Optimization and Scale-Up for Ba(Zr,Ce,Y)O_{3-δ}-Based
Proton-Conducting Electrolysis Half-Cells**

L.-A. Schäfer (2026), iv, 141 pp

ISBN: 978-3-95806-940-4 (Print)

ISBN: 978-3-95806-941-1 (E-Book)

Band / Volume 717

Kostenoptimale Energieausfälle in erneuerbaren Energiesystemen

D. Franzmann (2026), xix, 235 pp

ISBN: 978-3-95806-942-8 (Print)

ISBN: 978-3-95806-943-5 (E-Book)

Band / Volume 718

**Semiclassical Thermodynamics of Metal-Solution Interfaces: A Density-
Potential Functional Theory Study**

Z. Zhang (2026), 143 pp

ISBN: 978-3-95806-946-6 (Print)

ISBN: 978-3-95806-947-3 (E-Book)

Band / Volume 719

**Development and Characterization of the Airborne Chemical Ionization
Mass Spectrometer FunMass-C for Atmospheric Trace Gas Measurements**

S. Alber (2026), xli, 265 pp

ISBN: 978-3-95806-948-0 (Print)

ISBN: 978-3-95806-949-7 (E-Book)

Band / Volume 720

Material design and stability of All-Solid-State Lithium batteries

F. H. Al-Jaljouli (2026), xv, 174 pp

ISBN: 978-3-95806-950-3 (Print)

ISBN: 978-3-95806-951-0 (E-Book)

Band / Volume 721

Towards improved representation of the hydrological cycle by assimilating multi-source remotely sensed soil moisture data in terrestrial system models

H. Zhao (2026), XVI, 139 pp

ISBN: 978-3-95806-952-7 (Print)

ISBN: 978-3-95806-953-4 (E-Book)

Band / Volume 722

Combining Lattice Boltzmann Simulation and Synchrotron Imaging to Study Bubble Dynamics in Electrodes of Vanadium Redox Flow Batteries

K. Duan (2026), xxiii, 116 pp

ISBN: 978-3-95806-954-1 (Print)

ISBN: 978-3-95806-955-8 (E-Book)

Band / Volume 723

Geophysical characterization of landscapes along a climate and vegetation gradient in the Chilean coastal cordillera

I. Dal Bo (2026), 138 pp

ISBN: 978-3-95806-956-5 (Print)

ISBN: 978-3-95806-957-2 (E-Book)

Band / Volume 724

Model informed assessment of observation impact for emission rate optimization using four-dimensional variational data assimilation

A. Hermanns (2026), xvi, 109 Seiten

ISBN: 978-3-95806-958-9 (Print)

ISBN: 978-3-95806-959-6 (E-Book)

Weitere **Schriften des Verlags im Forschungszentrum Jülich** unter
<http://www.wzb1.fz-juelich.de/verlagextern1/index.asp>

Energie & Umwelt / Energy & Environment
Band / Volume 724
ISBN 978-3-95806-959-6