

Atmospheric Methane Data Assimilation in the CMAQ  
Air Quality Model

by

Seyyedsina Voshtani

A thesis submitted to the Faculty of Graduate and Postdoctoral  
Affairs in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Environmental Engineering

Carleton University  
Ottawa, Ontario

© 2022, Seyyedsina Voshtani

## **Abstract**

Atmospheric methane is a potent greenhouse gas (GHG) and the second-largest contributor to anthropogenic climate forcing. After stabilizing in the early 2000s, the global methane concentration has sharply risen since 2007, mainly due to human-related activities. Curbing the rise of methane concentrations entails identifying and reducing methane emissions, which may otherwise significantly impact climate and air quality. Due to their near-continuous global coverage, satellite observations of methane are often combined with chemical transport models (CTMs) to improve model concentrations and emissions estimates.

Previous methane studies are still faced with significant gaps and challenges such that considerable discrepancies among their results have been reported consistently. On the estimation side, most studies assumed that the model is perfect and characterization of uncertainties is already optimal. Obtaining information on methane uncertainties using conventional approaches requires extensive computational resources compared to model integration. Furthermore, there is a lack of independent and objective evaluation of those estimated uncertainties.

The first thesis objective is to develop a novel cost-efficient data assimilation framework capable of estimating error statistics using a CTM. This method is referred to as parametric variance Kalman filter (PvKF), which relies on continuous formulation of error covariance propagation without making the perfect model assumption. We test the validity of our assumptions and the performance of the PvKF assimilation using simulated GOSAT observations.

Our next goal is to conduct near-optimal assimilation to represent the *true* methane field. Cross-validation offers an objective manner to characterize the success of the method. We extend that method to the satellite observations and multiple covariance parameter estimations. Using estimated error statistics and GOSAT observations, we found that the quality of the analysis substantially depends on the optimality of those error covariances.

Lastly, we evaluate the use of PvKF assimilation in a source inversion context in comparison with a traditional 4D-Var inversion. Using Observing System Simulation Experiments (OSSEs), we verify the ability of our new inversion framework to recover a distribution of known emissions. Our results indicate that both the analysis field and its error covariance exert a tangible influence in lowering the bias and variance of the recovered emissions.

## Preface

This thesis includes the following copyrighted articles. The published manuscripts have been reproduced in Chapter 5 and Chapter 6 of this thesis with the permission of the co-authors and the publisher.

- **Voshtani, S.;** Ménard, R.; Walker, T.W.; Hakami, A. Assimilation of GOSAT Methane in the Hemispheric CMAQ; Part I: Design of the Assimilation System. *Remote Sens.* 2022, 14, 371. <https://doi.org/10.3390/rs14020371>
- **Voshtani, S.;** Ménard, R.; Walker, T.W.; Hakami, A. Assimilation of GOSAT Methane in the Hemispheric CMAQ; Part II: Results Using Optimal Error Statistics. *Remote Sens.* 2022, 14, 375. <https://doi.org/10.3390/rs14020375>

The following non-copyrighted manuscript has been reproduced in Chapter 7 of this thesis with permission from the co-authors:

- **Voshtani, S.;** Ménard, R.; Walker, T.W.; Hakami, A. Use of Assimilation Analysis in 4D-Var Source Inversion: Observing System Simulation Experiments (OSSEs) with GOSAT Methane and Hemispheric CMAQ. *Remote Sens.* 2022. (in review)

All materials should be cited as stated above.

Sina Voshtani declares that he is the author of the entire thesis and has played the leading role in conducting the original research described in Chapters 5, 6, and 7. He is the main contributor and responsible for designing the method, conducting the research, analyzing the data, and interpreting the results and findings of this thesis.

## **Acknowledgements**

The completion of this thesis would not have been possible without help from a number of people and organizations. First and foremost, I sincerely appreciate my three inspiring advisors. I thank Dr. Amir Hakami for sharing his enthusiasm for air quality and atmospheric modelling with me, for his continuous support and help, and for giving me much freedom in my research work and study. I am thankful to Dr. Thomas Walker for sharing his knowledge of remote sensing and atmospheric inversion with me, for the insightful discussions and feedback, and for his positive attitude and encouragement. I am grateful to Dr. Richard Ménard for teaching me much of what I know about data assimilation and filtering, for his wisdom, support, and guidance, and for his promptness in responding to all my questions. I greatly appreciate all the time each of you spent reviewing my drafts and going through the presentations.

I would like to also thank my thesis committee, Dr. Avelino Arrelano, Dr. Majid Mohammadian, Dr. Robin Chhabra, and Dr. Burak Gunay, for their valuable questions, comments, and their time. Many thanks to all my friends, colleagues, and researchers in the Carleton Atmospheric Modeling Group, the Air Quality Research Division (AQRD) at Environment Climate Change Canada (ECCC), and the Georgia Institute of Technology for their openness, support, and responses.

I would like to greatly acknowledge all atmospheric modelling community and data providers and laboratories, including the CMAQ scientific community and the CMAQ Adjoint team for making their model publicly available, the Netherlands Institute for Space Research (SRON) and Karlsruhe Institute for Technology (KIT) for providing GOSAT satellite XCH<sub>4</sub> retrievals products through the ESA GHG-CCI initiative, the NOAA Global

Monitoring Laboratory for providing GLOBALVIEWplus CH<sub>4</sub> ObsPack data products and in situ and aircraft measurements, the TCCON team for providing ground-based measurements, the HIPPO aircraft data providers, and Global Hawk Pacific Mission (GloPac) for providing aircraft measurements. I also appreciate the free use of methane emissions data from the Emissions Database for Global Atmospheric Research (EDGAR) and the data providers of Wetland Methane Emissions and Uncertainty (WetCHARTs).

My research work was made possible by funding from the Ontario Trillium Scholarship that I received at Carleton University and financial support from ECCC through the Research Affiliate Program (RAP). I also greatly thank the computational resources of Compute Canada, where the simulations for this research are partially performed.

Lastly, I would like to express special thanks to my family and beloved ones, whose patience, support, and understanding always kept me motivated to accomplish this research.

## Table of Contents

<b>Chapter 1: Introduction .....</b>	<b>1</b>
1.1 Research Objectives .....	12
1.2 Contribution and Originality of the Thesis .....	16
1.3 Structure of the Thesis .....	17
<b>Chapter 2: Background .....</b>	<b>19</b>
2.1 Methane Balance in the Atmosphere .....	19
2.1.1 Global Trend of Atmospheric Methane .....	19
2.1.2 Methane Budget: Major Sources and Sinks .....	22
2.2 Modelling of Atmospheric Methane .....	26
2.2.1 Process-based Emissions Estimation (Bottom-up) .....	27
2.2.2 Use of Observations to Estimate Emissions (Top-down) .....	29
2.3 Limitations and Challenges in Methane Inversion .....	32
2.4 Use of Data Assimilation for Methane .....	37
2.5 Background of Assimilation Methods .....	39
2.6 Potential Use of Parametric Filtering for Methane Assimilation .....	42
<b>Chapter 3: Data and Research Tools .....</b>	<b>49</b>
3.1 Observations .....	50
3.1.1 GOSAT Satellite Observations .....	50
3.1.2 Ground Network and Aircraft Observations .....	51
3.1.2.1 TCCON .....	51
3.1.2.2 HIPPO-3 .....	52
3.1.2.3 UCATS-GloPac .....	53
3.1.2.4 NOAA-ObsPack v3.0 .....	53
3.2 Inputs and Model .....	54
3.2.1 Methane Emissions .....	54
3.2.2 CMAQ and its Input Processors .....	56
3.2.3 Adjoint of CMAQ .....	59
3.3 Error Covariance Modelling .....	62
3.3.1 Observation and Background Error Covariance .....	62
3.4 Covariance Parameter Estimation .....	65
<b>Chapter 4: Data Assimilation and Inverse Modelling Methods .....</b>	<b>66</b>
4.1 4-Dimensional Variational (4D-Var) .....	72
4.2 Ensemble Kalman Filter (EnKF) .....	76
4.3 Parametric Variance Kalman Filter (PvKF) .....	80
4.3.1 Continuum Representation of Covariances: A Simple Dynamical Model .....	81
4.3.2 Solutions by Operator Splitting .....	83
4.3.3 Solutions by Method of Characteristics .....	85
4.3.4 PvKF Algorithm .....	86
4.4 General Form of Parametric Kalman Filter (PKF) .....	88

<b>Chapter 5: Assimilation of GOSAT Methane in the Hemispheric CMAQ; Part I:</b>	
<b>Design of the Assimilation System.....</b>	<b>90</b>
5.1 Introduction .....	90
5.2 Model and Observation Operator .....	96
5.2.1 Modifications of CMAQ to Handle Methane Transport and Emissions.....	96
5.2.2 GOSAT Observation Operator for Data Assimilation .....	97
5.3 Data Assimilation System .....	100
5.3.1 Background of the Assimilation Scheme .....	100
5.3.2 Forecast Step .....	101
5.3.3 Analysis Step.....	103
5.3.4 Analysis Step with 3D Observation Operator Using Averaging Kernels .....	106
5.3.5 An Overview of the Assimilation Algorithm .....	109
5.4 System Setup .....	112
5.4.1 Initial Conditions.....	112
5.4.2 Observation Bias Correction .....	116
5.4.3 Construction of Spatial Correlation Functions on the H-CMAQ Grid .....	119
5.4.4 Observation, Model and Initial Error Covariance Modelling .....	122
5.5 Verification of the Basic Properties of the Assimilation System .....	124
5.5.1 One-Observation Experiment.....	124
5.5.2 Timing (Computational Efficiency) .....	132
5.6 Summary and Conclusions.....	135
<b>Chapter 6: Assimilation of GOSAT Methane in the Hemispheric CMAQ; Part II:</b>	
<b>Results Using Optimal Error Statistics.....</b>	<b>139</b>
6.1 Introduction .....	139
6.2 Background on the Theory of Covariance Parameter Estimation .....	144
6.3 Estimation of Correlation Lengths and Observation Error Variance .....	148
6.4 Estimation of Model Error and Initial Error Variance Using Innovation Variance Consistency.....	156
6.5 Evaluation against Independent Observations.....	160
6.6 Characteristics of Analysis and Error Variances .....	171
6.7 Conclusions and Summary .....	179
<b>Chapter 7: Use of Assimilation Analysis in 4D-Var Source Inversion: Observing System Simulation Experiments (OSSEs) with GOSAT Methane and Hemispheric CMAQ.....</b>	<b>183</b>
7.1 Introduction .....	184
7.2 Background.....	189
7.2.1 Satellite Observations.....	190
7.2.2 Chemical Transport Model and Methane Emissions .....	191
7.3 Methodology.....	194
7.3.1 PvKF Assimilation .....	194
7.3.2 Source Inversion Procedure (4D-Var).....	197
7.3.3 Using PvKF Assimilation Analysis in 4D-Var Inversion: the Formulation.....	199
7.3.4 Numerical Aspects of Matrix Inversion .....	202

7.4	Description of the OSSE Experiments .....	205
7.4.1	Perturbation Tests.....	206
7.4.2	Experimenting with Different Cost Functions .....	208
7.5	Results and Discussions .....	210
7.5.1	Base case Uniform Perturbation.....	210
7.5.2	Perturbation of Different Sectors .....	215
7.5.3	More Realistic Perturbations .....	223
7.5.4	Overview of the Different Experiments .....	225
7.6	Additional Implications for the Proposed Source Estimation .....	228
7.6.1	Implications for the Forecast Model Error .....	228
7.6.2	Computational Timing of Different Inversions.....	233
7.7	Summary and Conclusions .....	236
<b>Chapter 8: Conclusions and Future Work .....</b>		<b>242</b>
<b>Bibliography .....</b>		<b>251</b>
<b>Appendices.....</b>		<b>283</b>

## List of Figures

Figure 1.1. Globally-averaged, monthly mean atmospheric methane abundance and its annualized growth rate based on marine surface data .....	3
Figure 1.2. Schematic view of chemical data assimilation (DA) and inverse modelling (IM) using atmospheric CTMs.....	5
Figure 2.1. The global trend of atmospheric methane over the past two millennia; and the global monthly mean methane concentrations between 1983 to 2021 .....	21
Figure 2.2. Bottom-up and top-down estimation for each source and sink category of the global methane budget between 2008 and 2017 .....	26
Figure 3.1. Anthropogenic methane emissions inventory compiled by USEPA.....	56
Figure 3.2. CMAQ Input Processors and CCTM .....	58
Figure 5.1. Flowchart of the analysis and forecast steps of the PvKF assimilation .....	111
Figure 5.2. Evolution of the analysis error variance on Day 0, Day 4, Day 8, Day 12, Day 16, and Day 20. ....	112
Figure 5.3. Difference between methane surface observations and model concentration for the initial guess, rescaled initial to April 2010 but before bias correction, and after bias correction.....	114
Figure 5.4. Linear regression fit of the difference between GOSAT observation and H-CMAQ simulated XCH <sub>4</sub> as a function of latitude, air mass factor (AMF), solar zenith angle, $\theta_s$ , and satellite viewing zenith angle, $\theta_v$ .....	117
Figure 5.5. Difference between observations and model before bias correction and after bias correction with respect to latitude over a month with the number of observations N = 59,031.....	119
Figure 5.6. One-observation experiment near the surface at higher latitude showing the analysis increment and error variance reduction at Day 0 and Day 3 .....	126
Figure 5.7. One-observation experiment at about 600 hPa and lower latitude showing the analysis increment and error variance reduction at Day 0 and Day 3.....	127
Figure 5.8. Vertical distribution of analysis increment and error variance reduction on Day 0 and Day 3 along the cross-section starts from (28° N, 120° W) to (15° N, 80° W). 129	
Figure 5.9. Analysis increment and error variance reduction for different vertical correlation length scales and pressure weights; and when the pressure weight is uniform; along with GOSAT column averaging kernel .....	131
Figure 5.10. The relative computational time of the assimilation analysis, model forecast, and entire assimilation process with respect to the number of observations ...	134
Figure 6.1. Spatial and temporal distribution of GOSAT observations used in cross-validation and how they are separated into three sets after thinning to generate active and passive observations.....	152

Figure 6.2. Estimation of the horizontal and vertical length scale using cross-validation cost function of passive and active observations .....	154
Figure 6.3. Estimation of observation error variance parameter using a cross-validation cost function of passive and active observations. ....	155
Figure 6.4. Estimation of the horizontal and vertical correlation length and observation error covariance parameter using independent TCCON observations with the cross-validation cost function.....	156
Figure 6.5. Normalized innovation variance consistency diagnostic for low, proper, and high model error parameter value .....	159
Figure 6.6. Normalized innovation variance consistency diagnostic for a low, proper, and high initial error parameter value. ....	160
Figure 6.7. Methane measurement data, including TCCON, HIPPO-3, UCATS/GloPac and NOAA/Obstack used for the evaluation of the assimilation system in April 2010.	164
Figure 6.8. Comparison of model forecast, analysis with optimal parameters, and (c) analysis with nonoptimal parameters against independent TCCON observations.....	166
Figure 6.9. Comparison between the model forecast and analysis with GOSAT assimilation, sampled at four types of independent measurements, including TCCON, HIPPO-3, UCATS/GloPac and NOAA/Obstack. ....	168
Figure 6.10. Differences between (a) model and GOSAT, (b) analysis and GOSAT, and (c) analysis and model in the observation space in April 2010. ....	173
Figure 6.11. Monthly analysis, analysis increment, and error variance reduction at the surface and at 550 hPa. ....	175
Figure 6.12. Comparison of the time series of the analysis and the model against TCCON observations.....	177
Figure 6.13. Comparison of the model forecast, analysis with optimal parameters, and analysis with nonoptimal parameters against TCCON observations.....	179
Figure 7.1. Sketch of the assimilation window followed by an inversion window, used for estimation of methane emissions scaling factor, $x$ . ....	200
Figure 7.2. Matrix inversion of one-month non-diagonal error covariance using data selection procedure. ....	205
Figure 7.3. Flowchart of the OSSE framework for optimizing methane emissions.....	207
Figure 7.4. Prior/posterior – true emissions using +50% uniform perturbation in the total prior emissions, shown for four variations (Type 0-3) of the inversion cost function; and statistical comparison of prior and posterior emissions against true .....	214
Figure 7.5. Prior/posterior – true emissions using +50% sectoral perturbation in agriculture, energy, waste, and wetland emissions, shown for four variations (Type 0-3) of the inversion cost function.....	220
Figure 7.6. Statistical comparison of prior and posterior emissions against true emissions in scatter plots for agriculture, energy, waste, and wetland sector .....	222

Figure 7.7. Prior/posterior – true emissions using $\pm 25\text{-}50\%$ variable perturbation in the total prior emissions, shown for four variations (Type 0-3) of the inversion cost function; and statistical comparison of prior and posterior emissions against true.....	225
Figure 7.8. Diffusion effect of model transport error; and schematic form and weight of the model error through PvKF formulation compared to error due to diffusion effect ..	232
Figure 7.9. Comparison between the value of the four types of cost functions against the number of iterations; and comparison of the computational cost of adding only error statistic to the cost function with the same model forecast field .....	236
Figure C.1. Polar stereographic projection geometry .....	298
Figure C.2. (a) $\text{GOSAT}^{\text{bias(AMF)}}$ – CMAQ with air mass factor bias correction, (b) $\text{GOSAT}^{\text{bias}(\theta_s)}$ – CMAQ with solar zenith angle bias correction.. .....	300
Figure C.3. Hemispheric spatial distribution of the relative methane concentration loss due to chemical reactions.....	304
Figure D.1. Comparison of the model forecast, analysis with optimized parameters, and analysis with non-optimized parameters against NOAA/ObsPack observations. ....	305
Figure D.2. Comparison of the model forecast, analysis with optimized parameters, and analysis with non-optimized parameters against UCATS/GLloPac observations.....	306
Figure D.3. Comparison of the model forecast, analysis with optimized parameters, and analysis with non-optimized parameters against HIPPO-3 observations. ....	306
Figure D.4. Comparison of non-optimal analysis due to only correlation lengths larger than the optimal value, optimal analysis with optimal correlation lengths, and non-optimal analysis due to only correlation lengths smaller than the optimal value. ....	307
Figure D.5. Comparison of non-optimal analysis due to only observation error smaller than the optimal value, optimal analysis with optimal observation error, and non-optimal analysis due to only observation error larger than the optimal value. ....	307
Figure D.6. Comparison of non-optimal analysis due to only model error smaller than the optimal value, optimal analysis with optimal model error, and non-optimal analysis due to only model error larger than the optimal value.....	308
Figure D.7. Comparison of the model forecast, analysis with optimal parameters, and analysis with non-optimal parameters against TCCON observations at Lamont (36.60°N, 97.48°W). .....	308
Figure D.8. Comparison of the model forecast, analysis with optimal parameters, and analysis with non-optimal parameters against TCCON observations at Bremen (53.10°N, 8.85°E) .....	309
Figure E.1. Prior/posterior – true emissions using $\pm 25\text{-}50\%$ variable perturbation in the total prior emissions (type II), shown for four variations (Type 0-3) of the inversion cost function; and statistical comparison of prior and posterior emissions against true .....	310
Figure E.2. Determination of the regularization parameter using two methods; (a) traditional method of minimization of the sum of normalized cost function and (b) L-curve method .....	312

Figure E.3. Model error effect of neglecting propagation of error correlation by diffusion on the computational cost of inversion ..... 314  
Figure E.4. Normalized difference of concentration field due to the effect of model diffusion scheme ..... 315

## List of Tables

Table 6.1. Comparison of the error variance parameters along with the cross-validation cost function generated with passive observations at different stages of iterative optimization. ....	153
Table 6.2. Comparison between mean bias ( $MB$ ), standard deviation $\sigma$ , coefficient of determination $R^2$ , and the linear regression line, using all measurements, including TCCON, HIPPO-3, UCATS/GloPac, and NOAA/ObsPack in April 2010.....	171
Table 7.1. Daily methane emissions in four main sectors and their subsets. Anthropogenic emissions are based on EDGAR v6, and natural wetland emissions are from WetCHARTs v3.0 with the full ensemble mean.....	193
Table 7.2. Cost functions for different formulations of 4D-Var inversion.....	209
Table 7.3. The normalized mean bias (NMB), the normalized mean error (NME), and Pearson's correlation coefficient ( $R$ ) for each emissions perturbation case and inversion cost function.....	227
Table C.1. Algorithm of Parametric variance Kalman filter (PvKF) assimilation .....	299
Table C.2. Evaluating the multiple (iterative) regression algorithm based on Mean Square Error ( $MSE$ ) and the correlation coefficient $r$ . Latitude, $\theta_s$ , and AMF represent the order of parameters in the multiple regression algorithm, respectively.....	301
Table C.3. Evaluating the multiple (iterative) regression algorithm based on Mean Square Error ( $MSE$ ) and the correlation coefficient $r$ . $\theta_s$ , latitude, and AMF represent the order of parameters in the multiple regression algorithm, respectively.....	302
Table C.4. Evaluating the multiple (iterative) regression algorithm based on Mean Square Error ( $MSE$ ) and the correlation coefficient $r$ . AMF, $\theta_s$ , and latitude represent the order of parameters in the multiple regression algorithm, respectively.....	303
Table E.1. The normalized mean bias (NMB), the normalized mean error (NME), and Pearson's correlation coefficient ( $R$ ) for $\pm 25$ -50% variable perturbation of sectors and for each inversion cost function.....	310

## List of Appendices

Appendix A. Parameter Estimation Methods of Covariance (Chapter 3).....	283
$\chi^2$ Diagnostic .....	283
Hollingsworth–Lönnberg .....	284
Maximum Likelihood .....	284
Desroziers Diagnostics.....	285
Cross-Validation.....	286
Appendix B. Kalman Filtering Assimilation and its Variants (Chapter 4).....	289
Kalman Filter (KF) .....	289
Extended Kalman Filter (EKF) .....	291
RTS Extended Kalman Smoother (RTS-EKS) .....	293
Appendix C1. Polar Stereographic Projection (Chapter 5).....	297
Appendix C2. PvKF Data Assimilation Algorithm (Chapter 5).....	299
Appendix C3. Regression Tests for GOSAT Bias (Chapter 5).....	300
Appendix C4. Methane Chemical Reaction Effect (Chapter 5).....	304
Appendix D. Evaluation Against Independent Observations (Chapter 6).....	305
Appendix E1. OSSE Non-uniform Perturbation (type II) (Chapter 7) .....	310
Appendix E2. Determination of Regularization Parameter $\gamma$ (Chapter 7) .....	312
Appendix E3. Diffusion Effect of Modelling Transport Error (Chapter 7).....	314

## List of Acronyms and Abbreviations

3D-Var	Three-dimensional variational data assimilation/inversion
4D-Var	Four-dimensional variational data assimilation/inversion
ADE	Atmospheric Diffusion Equation
AMF	Air Mass Factor
BU	Bottom-Up (Inventory)
CCTM	CMAQ Chemical Transport Model
CEDS	Community Emissions Data System
CMAQ	Community Multiscale Air Quality Model
CTM	Chemical Transport Model
CV	Cross Validation
DA	Data Assimilation
DDM	Decoupled Direct Method
ECCC	Environment and Climate Change Canada
EDGAR	Emission Database for Global Atmospheric Research
EKF	Extended Kalman Filter
EnKF	Ensemble Kalman Filter
EPA	Environmental Protection Agency
FAO	Food and Agriculture Organization
FOAR	First-Order-Auto-Regressive
GAINS	Greenhouse gas and Air pollutant Interactions and Synergies
GCM	General Circulation Model
GFED	Global Fire Emissions Database
GloPac	Global Hawk Pacific
GLWD	Global Lakes and Wetlands Database
GOSAT	Greenhouse Gases Observing Satellite
HIPPO	HIAPER Pole-to-Pole Observations
IASI	Infrared Atmospheric Sounding Interferometer
IPCC	Intergovernmental Panel on Climate Change
JAXA	Japanese Space Agency
KF	Kalman Filter
KIT	Karlsruhe Institute for Technology

KS	Kalman Smoother
L-BFGS	Limited-memory Broyden-Fletcher-Goldfarb-Shanno
MAP	Maximum A-Posteriori
MCIP	Meteorology Chemistry Interface Processor
MLE	Maximum Likelihood Estimation
MV	Minimum Variance
NASA	National Aeronautics and Space Administration
NOAA	National Oceanic and Atmospheric Administration
OI	Optimal Interpolation
OSSE	Observing System Simulation Experiment
PKF	Parametric Kalman Filter
PvKF	Parametric Variance Kalman Filter
RTS	Rauch-Tung-Striebel
SCIAMACHY	Scanning Imaging Absorption Spectrometer for Atmospheric Cartography
SMOKE	Sparse Matrix Operator and Kernel Emissions
SOAR	Second-Order-Auto- Regressive
SRON	Netherlands Institute for Space Research
SWIR	Shortwave Infrared
TANSO-FTS	Thermal And Near-infrared Sensor for carbon Observation Fourier Transport Spectrometer
TCCON	Total Carbon Column Observing Network
TD	Top-Down (Inventory)
TES	Tropospheric Emission Spectrometer
TIR	Thermal Infrared
TLM	Tangent Linear Model
UCATS	UAS Chromatograph for Atmospheric Trace Species
UNCCC	United Nations Framework Convention on Climate Change
VMR	Volume Mixing Ratio
WMO	World Meteorological Organization
WRF	Weather Research and Forecasting model

## List of Symbols

Note that the integrated thesis is a collection of articles in which some symbols may differ for certain variables.

$\alpha$	[-]	Vector of error covariance parameters
$\beta$	[-]	Element of the error covariance matrix
$\gamma$	[-]	Regularization parameter
$\Gamma$	[-]	Innovation covariance matrix
$\varepsilon^a$	[-]	Analysis error
$\varepsilon^b$	[-]	Background error or Prior error
$\varepsilon^i$	[-]	Initial error
$\varepsilon^o$	[-]	Observation error
$\varepsilon^m$	[-]	Measurement error
$\varepsilon^q$	[-]	Model error
$\xi$	[-]	A particle in the air
$\theta_s$	[°]	Satellite solar zenith angle
$\theta_v$	[°]	Satellite viewing zenith angle
$\lambda$	[ppb] <sup>1</sup>	Adjoint variable
$\rho$	[kg m <sup>-3</sup> ]	Air density
$\sigma$	[-]	Standard deviation or error <sup>2</sup>
$\Sigma$	[-]	Diagonal matrix of error standard deviation
$\chi^2$	[-]	Chi-square (diagnostic)
$\varphi$	[ppb]	Adjoint forcing term
$\Phi$	[m]	The trajectory of a particle in the air
<b>A</b>	[-]	Average kernel or Analysis error covariance matrix
<b>B</b>	[-]	Background or prior error covariance matrix
$c$	[ppb]	Pollutant concentration
$C$	[-]	Correlation function
<b>C</b>	[-]	Correlation matrix
<b>d</b>	[ppb]	Innovation vector of observation – model concentration

<sup>1</sup> Parts per billion (ppb) is a unit of mixing ratio or concentration used for trace gases (1 ppb = 1×10<sup>-9</sup> mol/mol).

<sup>2</sup> Errors and error covariances are all normalized so that the units are cancelled out (i.e., [ppb/ppb] ≡ [ppb<sup>2</sup>/ppb<sup>2</sup>] ≡ [-]).

$D$	[m]	Chordal distance between two points
$e, E$	[kg s <sup>-1</sup> ]	Emission rate
$g$	[ppb <sup>2</sup> ]	Local cost function
$H$	[-]	Observation operator function or Nonlinear observation operator
$H'$	[-]	Linearized observation operator
$H$	[-]	Linear observation operator
$I$	[-]	Identity matrix
$J$	[ppb <sup>2</sup> ]	Cost function
$K$	[-]	Kalman gain for concentration assimilation
$L_c$	[m] or [Pa]	Correlation length scale (horizontal or vertical)
$L$	[-]	Propagator of diffusion
$m$	[-]	Number of observations
$M$	[-]	Model operator function or Nonlinear model
$M'$	[-]	Linearized model
$M$	[-]	Linear model
$n$	[-]	Number of state elements
$p$	[-] or [Pa]	Probability density function or pressure
$P$	[-]	Covariance function
$P^a$	[-]	Analysis error covariance matrix
$P^f$	[-]	Forecast error covariance matrix
$q$	[-]	Model error variance
$Q$	[-]	Model error covariance matrix
$R$	[-]	Observation error covariance matrix
$t$	[s]	Time or timestep
$V$	[-]	Variance function
$\omega$	[-]	Vector of pressure layer weights
$\mathbf{x}$	[ppb] or [kg s <sup>-1</sup> ]	State vector of concentrations or emissions
$\mathbf{x}_A$	[ppb] or [kg s <sup>-1</sup> ]	Prior estimate of concentrations or emissions
$\mathbf{X}^a, \mathbf{x}^a$	[ppb]	Analysis of concentration
$\mathbf{X}^f, \mathbf{x}^f$	[ppb]	Model forecast concentration
$\mathbf{Y}^o, \mathbf{y}^o$	[ppb]	Observations

## Chapter 1: Introduction

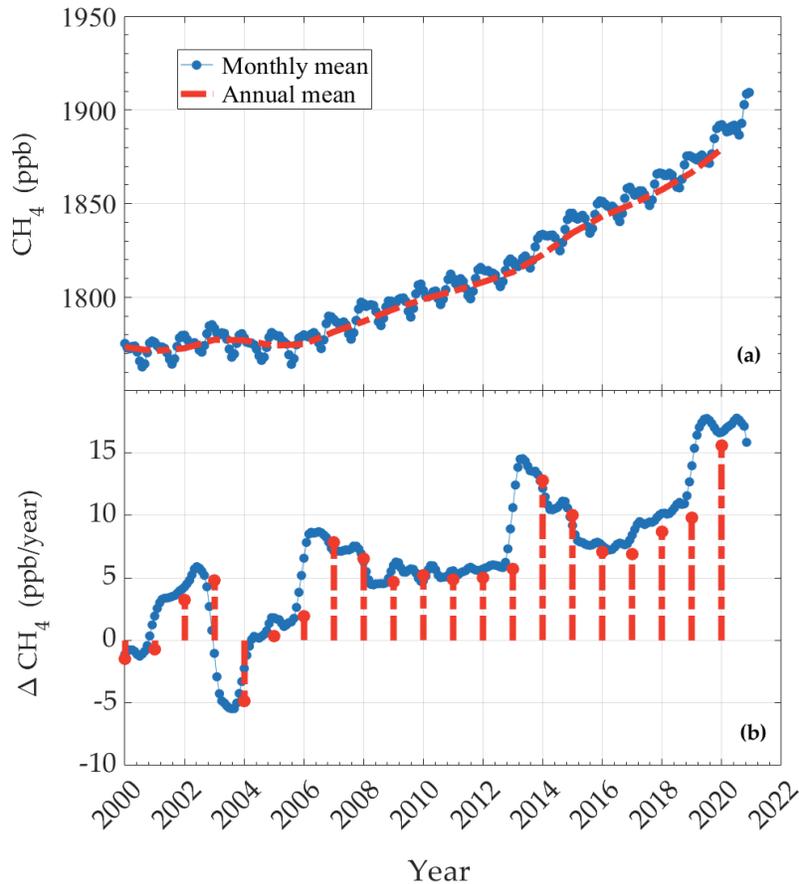
Methane ( $\text{CH}_4$ ) is a critical atmospheric component from both climate and air quality points of view (Staniaszek et al. 2022). It is the second-greatest contributor to the anthropogenic radiative climate forcing after  $\text{CO}_2$  (Myhre et al. 2013), with a shorter atmospheric lifetime and 28-32 times greater global warming potential (based on a 100-year time scale) than  $\text{CO}_2$  (Prather et al. 2012; Etminan et al. 2016). These properties have turned methane into an ideal candidate for slowing near-term climate change (Dlugokencky et al. 2011; Shindell et al. 2012; Nisbet et al. 2020). Furthermore, anthropogenic methane emissions are identified as the largest source of atmospheric methane (150% of all-natural sources), while its abatement cost is relatively low (Ganesan et al. 2019; Nisbet et al. 2020). Thus, methane emissions reduction strategies are considered a rapid, yet cost-effective route for climate mitigation (National Academies of Sciences 2018; Fletcher and Schaefer 2019).

In addition, methane is a chemically active species oxidized by photochemical reactions that mainly involve the hydroxyl radical (OH). Methane oxidation not only produces significant greenhouse gases (GHG) (e.g.,  $\text{CO}_2$ ), which directly contribute to global warming themselves, but it indirectly affects the oxidation of other pollutants such as stratospheric ozone, mainly through the production of water vapour (Brasseur and Jacob 2017). Ozone near the surface is considered as a criteria pollutant with adverse human health and ecosystem impacts. Variation of methane as the precursor of ozone through its reactions with OH can play a key role in altering local ozone abundances, thus exerting a significant air quality impact, particularly in populated areas (Fiore et al. 2002; Forster et

al. 2007a). Therefore, besides the climate gain, reducing methane emissions gives the added value of improving the air quality system.

The global methane concentration shows substantial growth in its climatological trend. Its globally averaged mixing ratio has increased from approximately 710 ppb in the preindustrial period of the 18<sup>th</sup> century to its highest level in November 2021 (1909.3 ppb) (Dlugokencky 2022). Despite that, large fluctuations in the growth rate of CH<sub>4</sub> have been recorded from year to year. Figure 1.1 shows that during the past two decades, the global methane concentrations suddenly began rising after a period of no growth between 2000 to 2007, also known as the stabilization period. Although various researchers have attempted to explain the drivers of the stabilization and the renewed growth period, it still remains a debatable research topic after one decade (Turner et al. 2019). An increase in fossil fuel (Rice et al. 2016) and a decline in biomass burning (Worden et al. 2017) and OH concentrations (Turner et al. 2017) have been argued to be the leading causes of this variation. Besides that, methane acceleration after 2007 was not well-predicted in the future climate scenarios compliant with the Paris Agreement target (constraining global climate warming below 2 °C according to the Intergovernmental Panel on Climate Change (IPCC 2006)); hence, the continuous and accelerating rise that is being observed may drift the atmosphere away from 2 °C scenario pathways and challenge the efforts to meet the Paris Agreement goals (Nisbet et al. 2022). All these considered, methane emissions reductions have been shifted to a priority for greenhouse gas mitigation strategies. However, effective methane mitigation not only requires reliable and comprehensive knowledge of its sources, but needs a clear understanding of its interactions and evolutions in the atmosphere (e.g., the oxidation with OH).

Methane emissions are generally derived from inventories using “bottom-up” methods that estimate emissions by allocating an emissions factor to a source activity data (e.g., applying the mass of methane emitted per unit of gas consumed to the volume of gas consumed per year).



**Figure 1.1. a) Globally-averaged, monthly mean atmospheric methane abundance and (b) its annualized growth rate based on marine surface data (Dlugokencky et al. 1994) from The Global Monitoring Division of NOAA’s Earth System Research Laboratory. Data are used from the website: [https://gml.noaa.gov/ccgg/trends\\_ch4/](https://gml.noaa.gov/ccgg/trends_ch4/) (Dlugokencky 2022)**

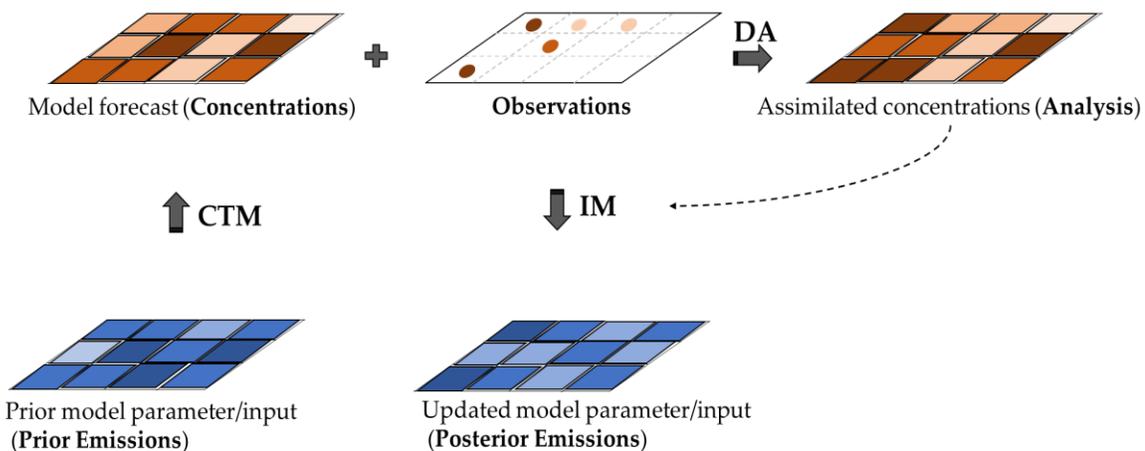
Accordingly, individual countries report their annual anthropogenic methane emissions to The United Nations Framework Convention on Climate Change (UNFCCC 2020) based on the IPCC guideline (IPCC 2013). However, these emissions may contain significant

uncertainties mainly due to either inaccurate emission factors or missing sources. Nevertheless, the bottom-up estimates can be verified and improved by “top-down” approaches that use information from atmospheric methane observations. Although the top-down method of deriving an emissions inventory can alleviate the issues of the bottom-up approach, it has its own challenges. Its performance depends on the observation characteristics (i.e., observation quality, density, errors, etc.) and an atmospheric model that links the emissions to the atmospheric concentrations. A precise and comprehensive representation in time and space of atmospheric methane concentrations not only provides valuable information on the abundance and distribution of methane with the aim of climate and air quality mitigation, but informs the top-down method to help better constrain methane emissions, particularly in a highly-resolved regional domain by providing estimates of the initial and boundary conditions.

A representation of global methane concentrations is typically obtained using chemical transport models (CTMs), driven by its inputs or parameters such as emissions, meteorology, etc. Although the model representation (also called the model forecast here) provides a comprehensive and discretized spatiotemporal estimate across the domain, it can be biased due to either an inaccurate input, such as emissions (Maasakkers et al. 2021) or an inadequate modelling capability to simulate methane in the atmosphere (Stanevich et al. 2021). Observations of methane concentrations, on the other hand, are often sparse but more precise and reliable than the model forecast. In recent years, methane satellite observations have been utilized extensively in atmospheric chemistry (Jacob et al. 2016; Buchwitz et al. 2017; Parker et al. 2020; Lu et al. 2022; Jacob et al. 2022), mainly due to their global coverage and reasonable precision and frequency. These observations can be

used along with the CTM to maintain consistent, yet accurate methane concentrations or to put a proper constraint on model inputs/parameters, which might be impossible to find otherwise (i.e., non-observable parameters such as emissions). A proper constraint and proper values of covariance parameters both correspond to the solution of optimal estimation.

An approach combining the observations with CTMs to improve the model forecast is generally known as chemical data assimilation (DA), while using observations solely for the purpose of improving the model inputs/parameters are often called inverse modelling (IM) (Asch et al. 2016). Figure 1.2 shows the schematic view of data assimilation and inverse modelling technique with CTMs. A DA system, in its simplest form, can be recognized by analogy with an intelligent inter-extrapolation scheme aiming at the most accurate and realistic (i.e., *true*) representation of a chemical species in the atmosphere (i.e., methane concentrations).



**Figure 1.2. Schematic view of chemical data assimilation (DA) and inverse modelling (IM) using atmospheric CTMs.**

On the other hand, a simple IM system can be envisioned as a complex regression model that applies the needed corrections to the atmospheric model parameters (e.g., methane

emissions scaling factors). The updated concentrations drawn from an assimilation system are often called the *analysis*, while updated emissions resulting from inverse modelling are commonly known as *posterior* emissions. Data assimilation and inverse modelling can be performed individually to improve the model performance or together in a coupled parameter-state estimation system (Elbern et al. 2007; Bocquet et al. 2015). Note that inverse modelling techniques are an essential part of top-down methods for developing methane emissions inventories.

Most of the past research in methane emissions inverse modelling has focused on a global scale problem with low spatial and no (or limited) temporal resolution. This can result in degraded source attribution, particularly for the source sectors with large spatial overlaps (Turner and Jacob 2015; Turner et al. 2015) and for high-emitting and fugitive sources such as those from industry (e.g., oil and gas), which can be highly time-variable (Zavala-Araiza et al. 2017; McNorton et al. 2022). Furthermore, significant discrepancies have been reported between the estimated emissions of different studies. Those discrepancies exist not only between top-down and bottom-up estimation, but among top-down estimations themselves, even those that used similar data sets (Ganesan et al. 2019; Miller et al. 2019). It implies that inverse modelling for methane emissions is still faced with significant challenges.

These challenges mainly arise from unaccounted uncertainties in methane sources and sinks as well as the errors in CTMs (e.g., transport error) and in observations (e.g., representativeness, interpolation, and numerical errors). Some studies have attempted to tackle this by resolving part of those errors and uncertainties prior to their inversion, so that they can better meet the perfect model/operator assumptions made in their system (Turner

et al. 2016; Maasakkers et al. 2019; Wang et al. 2019; Janardanan et al. 2020; Zhang et al. 2021). However, addressing all of those errors and uncertainties is not a trivial task (particularly errors in CTMs and observations); thus, it may require significant advances in model development or enormous computational resources to perform sensitivity analyses. Another practical alternative is to account for those errors, whether they are known or not, as part of the solution to an estimation problem (Yu et al. 2018; Stanevich et al. 2020; 2021). In other words, estimation of the modelling error (or transport error), whether in an assimilation or an inversion system, is the desired approach. However, estimating those error variances along with an atmospheric assimilation or inversion system is generally a tedious task entailing extensive computational cost.

An extension of the challenges presented to global methane emissions estimation is exemplified by scaling the problem to higher spatial resolutions. Inverse modelling in a limited domain and/or at high resolution requires precise knowledge of the initial and boundary conditions—likewise, for their uncertainties. Although the impact of the initial conditions diminishes over time by using longer integration (i.e., model spin-up), the bias and uncertainties in the boundary conditions will not disappear, and thus can exert a significant impact on emissions estimation. In fact, the contribution of emissions to the model concentrations is significantly smaller than the inflow of the background methane concentrations from the lateral boundary (Berchet et al. 2013; Wecht et al. 2014). Therefore, estimating the emissions becomes severely difficult over time once the domain is influenced everywhere by the boundary inflow. From an estimation point of view, the effect of the inflow will dominate the signal required to constrain the emissions at the

surface; and depending on the extent of the regional domain, the time for this to occur can vary from several days to a couple of weeks.

There are two main approaches to address the issues with the boundary conditions. These include (i) estimating the boundary conditions simultaneously with emissions in an inverse modelling framework (i.e., there are one estimation problem and an augmented state vector includes boundary conditions, besides emissions) and (ii) performing data assimilation of concentrations in the larger domain and likely with coarser resolutions, then obtaining the (coarse-scale) analysis to maintain boundary conditions for the limited domain (i.e., there are two estimation problems, yet not augmented, with their own state vector: concentration state vector in the larger domain and emissions state vector in the smaller domain). The latter (ii) can outperform the former (i) since we can distinguish the error propagations between emissions and boundary conditions, and more importantly, provide dynamically consistent and smooth boundary conditions to the regional domain. This may not be the case in the former (i) approach (Wecht et al. 2014; Jiang et al. 2015; Turner et al. 2015). All these facts considered, conducting data assimilation on a larger domain (global or hemispheric domain for methane) can reduce the uncertainties on the model state (i.e., concentration) to a level comparable to the sensitivity of the output concentrations relative to the emissions, hereafter referred to as emissions signal (i.e., changes in the output concentrations with respect to a unit of change in the emissions). This is necessary to accurately estimate emissions and their errors in the limited domain. Therefore, the former (i) approach is likely to be contaminated by large errors that affect the emissions estimations, contrary to the latter approach (ii).

Since the key role of data assimilation is either to constrain the initial/boundary conditions or to improve the representation of methane over the whole domain in space and time, we will begin this research by focusing on constructing a reliable and efficient data assimilation framework. This not only aims to provide an assimilation system for estimating methane state concentrations but to complement the past studies that are often conducted solely for methane source inversion. In addition, Massart et al. (2014) show that the methane assimilation system (based on 4D-Var) can be used to evaluate a new observation network. Data assimilation is useful not only to estimate the concentrations but also can help estimate the error statistics (including the analysis error covariances).

Formulating a data assimilation system capable of estimating the errors has its own challenges. Common data assimilation systems applicable to methane are based either on a variational (e.g., 4D-Var) or ensemble approach (e.g., EnKF) or sometimes a hybrid version of those two approaches (e.g., En4DVAR). All these methods not only require tens of model integrations to compute the analysis and its error covariances, but they may need a particular modification to obtain the desired results. For example, the assimilation may lead to degraded results without proper inflation of error variances in the EnKF method (Menard et al. 2021) or without appropriate preconditioning of the cost function in the 4D-Var method (Bousserez et al. 2015; Skachko et al. 2016). Hence, there still exists a lack of a cost-efficient and comprehensive data assimilation system for atmospheric methane or any long-lived species.

Without regard to the specifics of the assimilation algorithm, the performance of data assimilation of long-lived species is quite sensitive to accurate input error covariances (Daley 1992a; 1992d; Menard and Deshaies-Jacques 2018a). That includes the observation

and model errors as well as the correlation length scales of the background error covariances. Although the assimilation schemes listed above are derived from an optimal estimation theory (i.e., optimal assimilation) that assumes the analysis is optimal by nature, it still may not represent the real atmosphere (i.e., realistic estimation or the *true* analysis) unless the input error covariances are close to the truth (Menard 2016). Estimating the true analysis and its true error covariances is nontrivial, and very few studies attempted to tackle this for their particular problem (Menard and Deshaies-Jacques 2018b). Note that true and realistic assimilation are equivalent, and both correspond to a system that is not only optimal but also represents the real atmosphere.

As one main objective of this thesis, we attempt to develop a new data assimilation system that can provide us with a true analysis of methane concentrations and their realistic uncertainties. Given that these estimates are obtained with the goal of constraining emissions using observations, we also aim to achieve a robust and efficient inversion system further in this thesis. To accomplish these goals, we require a clear and sufficient understanding of the capabilities of different inverse modelling schemes. An inverse modelling system based on a variational method (e.g., adjoint inversion) reveals fine-scale corrections at the model's native resolution. Although it is considered a low-cost inversion scheme for the given resolution (assuming the adjoint model is already derived), it does not provide a posterior error covariance of emissions unless at the cost of additional computations (Bousserez and Henze 2018; Yu et al. 2021). It also suffers from frozen error statistics (i.e., error covariances are not being updated during assimilation) as well as a smoothing error once the number of observations is significantly smaller than the size of

the emission vector, both of which may degrade the inversion results (Wecht et al. 2014; Turner and Jacob 2015).

Inversion can also be constructed from a Bayesian inversion that explicitly computes the gain matrix (i.e., based on Kalman filtering and also referred to as an analytical inversion (Brasseur and Jacob 2017)) and systematically updates posterior errors along with the emissions, but it is computationally prohibitive in a model native resolution of a large dimension. Although aggregating the emissions at the grid level can facilitate the performance of this type of inversion, it usually results in a sub-optimal estimation (with respect to the model native resolution) with a considerable aggregation error (Turner and Jacob 2015; Bousserez et al. 2016). Overall, more criteria need to be addressed once the estimation focuses on the model parameters through an inversion system rather than only the model state. To achieve an appropriate inversion based on each of those methods, we need to develop a controlled environment with a known solution. Observing System Simulation Experiments (OSSEs) are a standard approach in atmospheric inverse modelling and data assimilation that provide a basis for conducting a variety of system assessments. OSSEs give us insights to design a reliable and efficient inversion system for constraining methane emissions.

Altogether, the proposed research addresses how to use methane observations in a cost-effective way to improve atmospheric CTMs (e.g., Community Multiscale Air Quality (CMAQ)) in their prediction of atmospheric methane concentrations along with deducing realistic statistics, and eventually, to improve on the current standard inversion approach (e.g., 4D-Var inversion) and revise the underlying assumptions in a controlled environment

in order to constrain emissions with higher spatial representativeness. More specifically, this thesis provides insights into addressing the following questions:

- How might one design a low-cost yet powerful data assimilation system with CMAQ for estimating atmospheric methane and its uncertainties?
- What can satellite observations, particularly GOSAT, tell us about methane within the air quality CMAQ model?
- How can we determine realistic uncertainties in that new methane data assimilation framework?
- What is the role of estimating input error characteristics, including observation and modelling error, in advancing the performance of a data assimilation system?
- How can optimal assimilation analyses, along with their uncertainties, help better resolve methane emissions through an inverse modelling system?

## **1.1 Research Objectives**

In this thesis, we use various observation types, including satellite, surface, and aircraft observations, to assess an assimilation process with CTMs governing the distribution of atmospheric methane. This assimilation system uses satellite observations from GOSAT in combination with the hemispheric air quality model (hemispheric CMAQ) driven by meteorology from the Weather Research and Forecasting (WRF) model and anthropogenic and natural emission inventories, mainly from Emission Database for Global Atmospheric Research (EDGAR v6) and WetCHARTs v3.0. The Sparse Matrix Operator Kernel Emissions (SMOKE) emissions processing model deal with both emission

inventories to provide hourly gridded emissions for hemispheric CMAQ. The observations used for evaluation consist of in situ and aircraft observations from GLOBALVIEWplus CH4-ObsPack v3.0 compiled by National Oceanic and Atmospheric Administration (NOAA), total column observations from Total Carbon Column Observing Network (TCCON), and aircraft observations from HIAPER Pole-to-Pole Observations (HIPPO-3) and Global Hawk Pacific (GloPac) missions. The thesis then pursues several objectives that can be described in three phases as follows.

In Phase I, we present the development of a data assimilation system, as a first of its kind for atmospheric methane, that is capable of estimating error statistics. The assimilation system can be used as a stand-alone model to improve the methane forecast in CMAQ (at native resolution and the current time of the model) or to complement a regional methane source inversion by providing consistent initial and boundary conditions. One main characteristic of this assimilation system is to alleviate the high computational cost of other assimilation approaches (e.g., 4D-Var and EnKF) with the same error estimation capabilities. The new scheme is based on a parametric Kalman filter (PKF) data assimilation that was recently introduced to CTMs (Pannekoucke et al. 2016). This method fundamentally relies on a continuous formulation of error covariances propagation using the dynamics of the same model, which is solved by the method of characteristics (Cohn 1993). Given a (near-) linearity assumption between methane concentrations and their emissions (Jacob et al. 2016), we design a simpler version of the PKF assimilation method, relying on the advection of error variance and a stationary correlation model. We call this scheme parametric variance Kalman filter (PvKF) assimilation. Using the advection of error variance not only makes the assimilation significantly cheaper in cost (i.e., only two

model integrations), but avoids the loss of error variance—a common phenomenon in standard Kalman filtering (Menard et al. 2021; Pannekoucke et al. 2021; Gilpin et al. 2022). We verify the performance of the PvKF assimilation and the validity of the assumptions using synthetic observation experiments. The system setup preparation and assumptions prior to the assimilation are also demonstrated in this phase. That consists of how (i) methane initial and boundary conditions and its emissions are configured in hemispheric CMAQ, (ii) bias in GOSAT observations is addressed, (iii) spatial correlation functions are designed, and (iv) error covariance in the forecast model, observations, and initials are modelled. Chapter 5 presents the development of the novel assimilation system and delineates the major findings of Phase I.

Phase II of this thesis examines the PvKF assimilation system, developed in Phase I, with actual GOSAT observations over one month in April 2010. The primary objective of this phase is to frame an (near) optimal assimilation system that represents the *true* analysis of atmospheric methane. The true analysis is not only optimized, but its error statistics also reflect the true uncertainties. This may not be the case, even though the analysis error covariances from an assimilation scheme (e.g., derived from a minimum variance estimation theory) are assumed to be optimal by its nature. Obtaining the true error covariances is nontrivial as it depends on estimating the true Kalman gain (Menard 2016). However, an alternative formulation exists based on the cross-validation technique. The theoretical development and the application of this approach with one covariance parameter and in situ observations have been shown previously (Menard and Deshaies-Jacques 2018a; 2018b). We adopt this approach and extend it to the estimation using satellite (i.e., GOSAT) observations and for multiple error covariance parameters. We

optimize the horizontal and vertical correlation length scales and observation error covariance parameters using the cross-validation technique. Another diagnostic based on normalized innovation variance matching is demonstrated to provide an estimate of the model and initial error covariance parameters. The analysis obtained from the estimated covariance parameters is considered the optimal analysis representing the truth. The performance of optimal analysis is verified by comparing it, the free-running model, and a non-optimal analysis against independent surface and aircraft observations. Eventually, we examine the result of the optimal assimilation in addressing the spatial structure of the bias and uncertainty across the spatial domain as well as its temporal consistency.

Phase III is concerned with creating an improved 4D-Var inverse modelling system to constrain methane emissions in the Northern Hemisphere. Since the PvKF assimilation system provides us with both an optimal state (i.e., analysis) and its uncertainties, it is worth exploring their influence on emissions estimates when it is linked to a source inversion system. We recall that the contribution of emissions to the state on a short time scale over a regional domain is weaker than the contribution from initial and boundary conditions. This necessitates a more accurate and unbiased estimate of the state before and throughout the inversion. Furthermore, many inversion systems suffer from limitations that either affect the quality of inversion result or degrade their computational efficiency. Avoiding modelling errors (i.e., perfect model assumptions) and providing inadequate information on the background initial and error statistics in the inversion system are among those restricting factors. Therefore, as the main objective in Phase III, we investigate the potential benefits of linking PvKF assimilation, developed in Phases I and II, to a 4D-Var inversion. Besides being realistically estimated, the PvKF analysis field is more capable

than the forecast model in (efficiently) providing an unbiased initial state for the inversion. In addition, propagating PvKF analysis error variance can retain the off-diagonal observational errors (i.e., error correlations in observation space) in a dynamically coherent manner. That information is often missed from a typical (4D-Var) inversion, where in the absence of realistic error statistics, a diagonal observation error covariance is assumed. Accordingly, using Observing System Simulation Experiments (OSSEs), we test the ability of our modified inversion framework to recover the true emissions. In particular, we examine the effect of the PvKF analysis field and its error covariance on the statistics of the optimized emissions and on the computational efficiency of the scheme as a whole.

## **1.2 Contribution and Originality of the Thesis**

The contribution to the original knowledge in this thesis can be summarized in the following:

- Presenting a novel low-cost chemical assimilation framework (i.e., PvKF) for the state and uncertainty estimation of long-lived species such as methane. (Chapter 5)
- Extending the application of covariance modelling with cross-validation capabilities for assimilating satellite observations such as GOSAT, aiming at optimal analysis. (Chapter 6)
- Demonstrating the significance of estimating appropriate error covariance parameters such as correlation length scales for maintaining assimilation that represents the real atmosphere. (Chapters 5 and 6)
- Quantifying the effect of optimal states and their propagated uncertainties for improving methane emissions estimation using 4D-Var inversion. (Chapter 7).

Overall, the contribution of this thesis to the community can be viewed mainly as technical developments in advancing a particular chemical data assimilation methodology. Furthermore, the accomplishments mentioned above elucidate the value of state and source estimation for long-lived species such as methane. Hence, this work tackles fundamental problems in data assimilation and atmospheric inversion, and its novelty offers an application for improving the current inventory of methane emissions.

### **1.3 Structure of the Thesis**

This thesis, in Chapter 2, provides a background of atmospheric methane global trends, the major contributors to its budget, and standard methods for estimating methane emissions, particularly those that use observations (i.e., inverse modelling). The limitations involved with inversion methods are then discussed, and the potential use of methane data assimilation, particularly the efficient PvKF assimilation, to complement the model forecast as well as the inversion system is presented. At the end of the section, a background of assimilation methods capable of estimating error statistics is provided.

Chapter 3 presents the two key components deriving an assimilation and inversion system: observations and model. A description of the GOSAT satellite observations as the main contributor to the assimilation/inversion system, along with four types of surface and aircraft measurements used for validation purposes, are presented. A summary of methane emissions inventories, the CMAQ model main processors, and a description of the adjoint of CMAQ, is then provided in this section. Lastly, as part of estimation tools, method of covariance modelling and covariance parameter estimations, including the cross-validation that is applied in this thesis, are illustrated.

Chapter 4 starts with the Bayesian estimation fundamentals involved in various assimilation/inversion methods, including those used in this thesis. It then gives an overview of the algorithms and formulations of popular data assimilation and inversion methods from both variational and Kalman filtering categories. Those methods are also compared to each other from a different aspect, but are mainly evaluated based on their capability to meet our first objective of maintaining a low-cost and reliable assimilation system.

The three main objectives of the thesis are successfully addressed in Chapters 5, 6, and 7, respectively. A conclusion and summary of the accomplishments are provided in Chapter 8, followed by suggestions for future research works.

## **Chapter 2: Background**

### **2.1 Methane Balance in the Atmosphere**

#### **2.1.1 Global Trend of Atmospheric Methane**

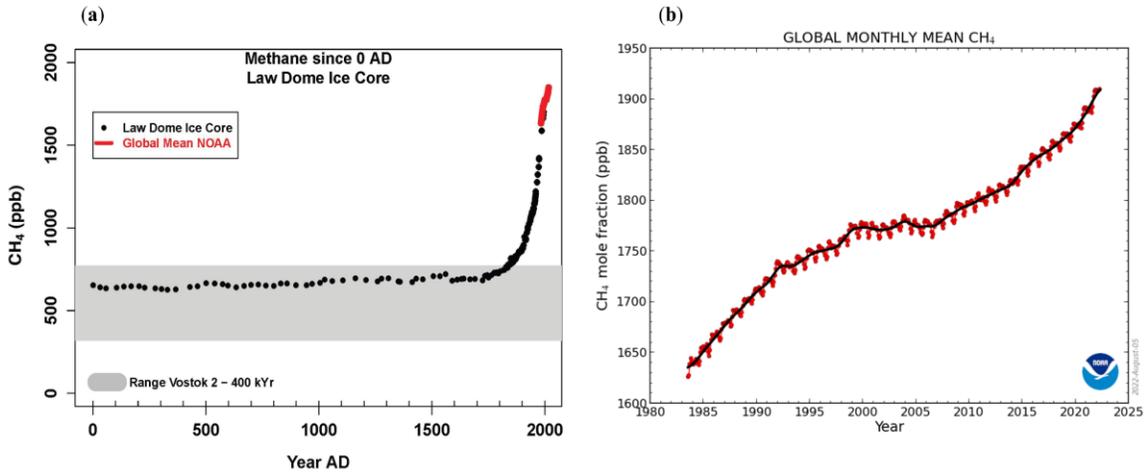
The global concentration of methane has risen continuously from the preindustrial period up to the end of the 20<sup>th</sup> century (IPCC 2013; Saunio et al. 2016a). The natural sources of methane, particularly wetland emissions, have been modelled by Arora et al. (2018), who showed a considerable increase in those emissions due to the change in temperature between 1850 and 2000 (by 30% from about 130 to 169 Tg CH<sub>4</sub> yr<sup>-1</sup>). Dean et al. (2018) also confirmed the key role of natural methane emissions in controlling climate change. However, there is no debate that the major cause of the rapid and bulk increase of atmospheric methane from preindustrial times lies in human activities (Figure 2.1), especially during the industrialization period in the 20<sup>th</sup> century when fossil fuels consumption has increased dramatically (more than 100% increase in total methane sources from 250 to 560 Tg CH<sub>4</sub> yr<sup>-1</sup>) (Etheridge et al. 1998; Ferretti et al. 2005). The first accurate global in situ measurements were made by Blake et al. (1982), Dlugokencky et al. (1994) at NOAA and Prinn and Weiss (1983) at the Advanced Global Atmospheric Gases Experiment (AGAGE) in 1983, and since then, a continued increase of methane has been monitored until 2000 (Dlugokencky 2022).

The global mean concentration of methane, following an almost steady period between 2000 to 2007 referred to as the stabilization period (Figure 2.1b), resumed rising in 2007 with an almost 7.8 ppb annual growth (Rigby et al. 2008; Dlugokencky et al. 2011). NOAA's recent analysis showed the growth rate at the highest recorded value during 2021

as 18.1 ppb, which was 15% greater than the period between 1984-2006 (Dlugokencky 2022).

Various explanations have been proposed to describe the mechanics behind the stabilization and the recent accelerating increase of atmospheric methane. Some of them have solely concentrated on the causes of the stabilization period, whereas others have taken that as an anomaly and mainly focused on explaining the resumption. Nevertheless, all studies attribute those changes to either the methane removal processes (i.e., sinks) or emissions (i.e., sources) (Heimann 2011). For example, Aydin et al. (2011) and Simpson et al. (2012) explained the slow-down as a result of a decline in oil and gas emissions, while Kai et al. (2011) and France et al. (2022) indicated that it was possibly due to a decrease in rice emissions. Rigby et al. (2017) also demonstrated that the steady period is potentially due to an increase in OH concentrations. Dlugokencky et al. (2011) and Zhang et al. (2022) suggest that methane rise occurred for the biogenic source due to a decline in the  $\delta\text{C-CH}_4$  isotope, particularly over the Tropics. Basu et al. (2022) showed that using joint assimilation of  $\delta^{13}\text{CH}_4$  and methane, about 85% of re-growth is due to microbial sources. Using an ensemble of models and relying on observations, Saunois et al. (2017; 2020) showed a significant increase in anthropogenic methane emissions after 2007. They also proposed that the growth is potentially caused by agricultural emissions, considering that the fossil fuel emissions have been estimated to be even smaller than before in many regions. However, Rice et al. (2016) estimated an increase in global fossil fuel emissions. Dalsoren et al. (2016); Rigby et al. (2017), and Turner et al. (2019) also examined methane observations and suggested that the decline in OH concentrations possibly explains the recent growth of methane. Finally, Worden et al. (2017), using CO and CH<sub>4</sub> measurements,

inferred a decline in biomass burning emissions, which reconcile the potential increase of fossil fuel throughout the recent methane growth.

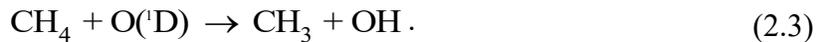
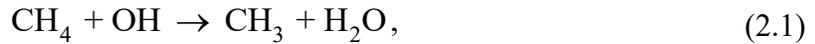


**Figure 2.1.** (a) The global trend of atmospheric methane over the past two millennia is obtained by analysis of air bubbles trapped in ice cores (black dots). Methane global concentrations have tripled since preindustrial times. The grey shaded area shows the methane range between 2,000 to 400,000 years ago. The Law Dome data are from Meure et al. (2006); The Vostok data are from Petit et al. (1999); (b) Global monthly mean methane concentrations are measured between 1983 to 2021 (red dots in both subfigures) by NOAA, and the running average in black. Figure 2.1a is adapted from (National Academies of Sciences 2018a), and Figure 2.1b is from Dlugokencky (2022).

All those non-mutually exclusive conclusions imply that the underlying process driving the recent trend of atmospheric methane is still not well known, likely due to a large uncertainty involved in modelling methane and particularly in quantifying its sources and sink process (Turner et al. 2017; Ganesan et al. 2019). Nonetheless, due to the recent advancements in space-borne technology, various new observations have been used to address those uncertainties and to better explain the recent trend of methane growth (Jacob et al. 2016; Jacob et al. 2022; Worden et al. 2022).

### 2.1.2 Methane Budget: Major Sources and Sinks

Methane (CH<sub>4</sub>) is a chemically reactive compound, and its main removal process occurs chemically in the atmosphere. Atmospheric methane is oxidized with hydroxyl radical (OH), Cl radicals, and excited atomic oxygen O(<sup>1</sup>D) in the atmosphere as follows



CH<sub>4</sub> removal by OH (Equation (2.1)) in the troposphere is the largest sink of methane, accounting for more than 90% of its total sink. The two other sink reactions of methane (Equation (2.2) and (2.3)) are minor and largely take place in the stratosphere (Brasseur and Solomon 2005). CH<sub>4</sub> is a major precursor of H<sub>2</sub>O and other greenhouse gas such as CO<sub>2</sub> in the stratosphere, which indirectly affects the climate condition. On the other hand, it is a critical source of ozone (O<sub>3</sub>) in the troposphere. In fact, in a NO<sub>x</sub> saturated environment, associated with many populated cities, CH<sub>4</sub> acts as a precursor of ozone through a chain of chemical reactions with hydroperoxyl radical (HO<sub>2</sub>), nitric oxide (NO), and the photolysis reaction of NO<sub>2</sub> (Seinfeld and Pandis 2016). Furthermore, by removing tropospheric OH, CH<sub>4</sub> negatively influences the oxidation capacity of the troposphere in removing criteria pollutants (Jacob 1999). Therefore, higher concentrations of methane not only exert a substantial climate impact, but can also be significantly harmful to air quality (Staniaszek et al. 2022). Other minor losses of methane occur by soil oxidation (i.e., soil uptake), usually in the presence of methanotrophic bacteria that consume soil's methane (Dutaur and Verchot 2007), and by photochemistry in the marine boundary layer (Sauniois et al. 2020).

The lifetime of methane depends on those removal processes, which can be driven by obtaining the ratio of methane burden to its loss rates (i.e., OH oxidation, stratospheric photochemistry, and soil uptake). In theory, a simple lifetime of a species is represented by the e-folding time for loss (e.g., exponential decay), indicating the time it takes for the abundance of the species in the absence of sources to decrease by 1/e. Methane's lifetime is estimated to be about 10 years (Prinn et al. 2005; Prather et al. 2012), implying that CH<sub>4</sub> concentrations may last more than a decade in the troposphere while their spatial distribution is influenced by OH and its spatial variability (Zhang et al. 2018). OH concentrations also vary in time (e.g., seasonally or annually), depending on humidity and ozone photochemistry through the main production pathway:



Hence, it is expected that during summer and in tropical regions, due to larger solar radiation and water vapour abundance, the rate of OH removal is higher; as a result, a greater methane chemical removal is expected (see Figure C.3 in Appendix C4). Despite those regional variations, the global burden of OH (or its mean concentrations) is relatively stable and well-known. The interannual variability of OH is estimated at about 1% and has changed only slightly over the past 150 years (Naik et al. 2013; Zhao et al. 2019; Zhao et al. 2020b). This suggests that the global methane concentration, which is well-mixed in the troposphere, behaves in a roughly linear manner (i.e., quasi-linear) over a short time scale (e.g., a few months or less) (Jacob et al. 2016).

The spatial distribution of methane, however, is also determined by two other factors: methane sources and atmospheric transport. Methane sources are divided into two

main categories: anthropogenic and natural. About 60% of total methane emissions are attributed to anthropogenic sources, which mainly include emissions from energy production (e.g., coal, oil, and gas), agriculture, landfills, and water wastes (Saunois et al. 2020). Anthropogenic emissions have shown a continuous overall increase since the preindustrial times, likely due to population and economic growth. Although it is sometimes assumed that anthropogenic emissions have an overall small temporal variability, a few source sectors have a substantial seasonal variability, such as rice cultivation, oil and gas, and agriculture (Sass et al. 2002; Vaughn et al. 2018).

Natural emissions are mainly due to anaerobic respiration that occurs in wetlands. Wetlands account for more than 85% of natural emissions and about 30% of total methane emissions globally (Kirschke et al. 2013; Saunois et al. 2020). Since natural emissions, particularly biospheric greenhouse gases, are sensitive to climate change, their feedback provides valuable information for understanding their characteristics as well as their interaction with the atmosphere and ecosystems (Heimann and Reichstein, 2008). Other natural methane emissions in the global methane budget include open fires (van der Werf et al. 2017), termites (Fung et al. 1991), seeps (Etiopie et al. 2009) and lakes (Walter et al. 2006). Although these sources are considered minor, they can still impact the methane budget, particularly when it comes to regional and local scale estimation or during a particular period of time when those emissions play a key role (Giglio et al. 2013).

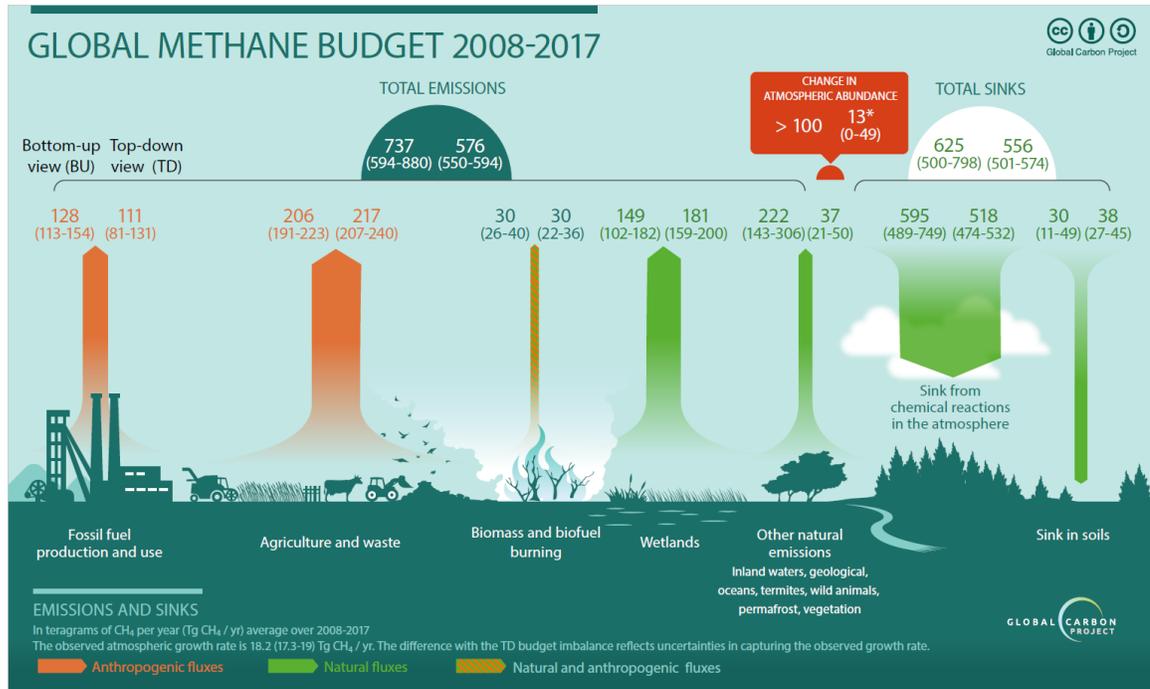
Volcanos also emit a significant amount of methane; however, they only occur occasionally, and their long-term impact will be faded away (Banda et al. 2015). Another important source of methane exists in the Arctic permafrost soil (e.g., sea sediments) and in the deep ocean as methane hydrates. Those, however, are trapped in solid sea ices and

thus have little chance to impact the atmosphere at present. Still, due to the continuous rise of Earth's temperature, melting permafrost and destabilizing methane in the oceans are expected in future. Those, if they took place, would exert a drastic climate impact by releasing a significant amount of methane into the atmosphere (Kort et al. 2012; Reichstein et al. 2013; Ciais et al. 2014). Since the dynamic of the process changes in this situation, a data assimilation system may need to be revised in its configurations. One common practice is to perform sensitivity analyses to test the performance of the DA system in response to new changes and apply appropriate modifications.

Methane emissions are generally quantified in inventories using bottom-up methods that estimate emissions by allocating an emissions factor to source activity data; for example, by applying the volume of methane released per volume of gas consumed to the volume of gas consumed per year. Accordingly, each individual country reports their annual anthropogenic methane emissions to The United Nations Framework Convention on Climate Change (UNFCCC 2020) based on the IPCC guideline (IPCC 2013). This leads to producing a globally consistent methane emissions inventory. However, these emissions may contain substantial uncertainties mainly due to inaccurate emission factors or miscounted activities. Nevertheless, the bottom-up estimates can be evaluated and improved by "top-down" constraints that use information from atmospheric methane observations. Figure 2.2 from the Global Carbon Project (Saunio et al. 2020) shows both bottom-up and top-down estimates of methane sources and sinks from both anthropogenic and natural categories with their uncertainties.

Saunio et al. (2020) found that the bottom-up inventories contain almost 30% higher global emissions than top-down estimates, although the split among different sectors

varies. Similar studies under Global Carbon Projects in the past also reported a large uncertainty in the wetland emissions as well as a double-counting of some particular sectors (Kirschke et al. 2013; Saunio et al. 2016a) in the bottom-up inventories.



**Figure 2.2. Bottom-up and top-down estimation for each source and sink category of the global methane budget between 2008 and 2017. Figure 2.2 is adapted from Saunio et al. (2020).**

## 2.2 Modelling of Atmospheric Methane

One main reason for modelling methane, similar to other atmospheric compounds, is to better understand the evolution of atmospheric methane concentrations and to quantify its global budget. Atmospheric models, specifically chemical transport models (CTMs), maintain a discretized and domain-wide consistent representation of atmospheric methane. Considering that direct measurements in space and time are rather limited within the currently available observation networks, CTMs can be useful when combined with those measurements. The use of methane observations with a CTM to indirectly infer its sources and sinks is considered inverse modelling or top-down (TD) estimation. Note that

quantifying methane sources, due to the high uncertainties involved and, more importantly, their value for GHG mitigation policy, has been at the center of attention for the research community since the 1980s. Top-down methods complement the direct process-oriented or bottom-up (BU) approach for constructing inventories. However, since the top-down approach generally relies on the bottom-up inventory as its prior estimates, it is essential to understand the mechanics of both.

### **2.2.1 Process-based Emissions Estimation (Bottom-up)**

The bottom-up inventories use knowledge of the underlying process to calculate methane emissions. They integrate statistical information on various activities for different census divisions to estimate anthropogenic emissions. Every country allocates emissions scaling factors to each type of activity and then transforms their metric to the amounts of methane emissions. Accordingly, emissions are computed based on

$$E = A \times F \times S, \quad (2.6)$$

where  $A$  is the rate of activity corresponding to the emissions process,  $F$  denotes the emissions factor showing the amounts of emissions per unit of activity, and  $S$  represents a scaling factor for any other processes that are not accounted for in  $A$  and  $F$ , such as surface or meteorological properties (Brasseur and Jacob 2017). Anthropogenic methane emissions of individual countries are then reported to (UNFCCC 2020) based on the IPCC guideline. Finally, to obtain uniformly resolved or gridded emissions, the country statistics are distributed with some appropriate spatial weighting (i.e., surrogates) maps (UNC 2017). For example, the distribution maps of agriculture and animal density are commonly used to spatially allocate livestock emissions. This information may also contain significant uncertainties, particularly in developing countries. Several widespread global BU methane

anthropogenic emissions inventories include Emissions Database for Global Atmospheric Research (EDGAR) (Janssens-Maenhout et al. 2019), the United States Environmental Protection Agency (USEPA) inventories (US-EPA 2016), the Greenhouse gas and Air pollutant Interactions and Synergies (GAINS) (Hoglund-Isaksson 2012; 2017; Gómez-Sanabria et al. 2018), the Community Emissions Data System for historical emissions (CEDS) (Hoesly et al. 2018) and the Food and Agriculture Organization (FAO) dataset emission database (Tubiello et al. 2022). Besides the difference in their assumptions and data used, these inventories do not provide the same level of details for many sectors; furthermore, they are not entirely independent as they must follow the same guideline (IPCC 2006) to classify their emissions. In addition to global bottom-up anthropogenic inventories, quite a few independent inventories are developed at the regional or country-level or for some particular sectors, generally aiming to provide detailed and better spatially and/or temporally resolved emissions (Maasakkers et al. 2016; Scarpelli et al. 2020; Scarpelli et al. 2022)

Bottom-up inventories of natural emissions can also be derived from process models. Biochemical models are typically used to compute natural methane emissions by incorporating a large amount of information from the soil, water, vegetation, meteorology and climate conditions and simulating the transport and biological processes in the soil and water. However, significant uncertainties can be involved in their simulations, particularly when it comes to a regional emissions estimation (Bohn et al. 2015). For example, Tian et al. (2016) indicate that for wetland, as the dominant natural source of methane, there is a significant difference between BU model estimation and TD estimate. The former range between 117 to 203 Tg CH<sub>4</sub> yr<sup>-1</sup>, whereas the latter are between 111 to 167 Tg CH<sub>4</sub> yr<sup>-1</sup>.

Parameters associated with climatic variability responses, model process descriptions, and driver data for wetland inundation extent are considered as primary factors causing large uncertainties in the inventories (Bloom et al. 2017; Ma et al. 2021).

Primary natural inventories of methane are provided by WetCHARTs (Bloom et al. 2017), Kaplan (2002), Pickett-Heaps et al. (2011), and Global Lakes and Wetlands Database (GLWD) (Lehner and Doll 2004) for wetland emissions; Global Fire Assimilation System (GFAS) (Kaiser and Benner 2012) and Global Fire Emissions Database (GFED) (van der Werf et al., 2017) for biomass burning and fire emissions; Fung et al. (1991) for termites emissions; Etiope et al., (2009) and Walter et al. (2006) for seeps and lakes emissions; Etiope and Milkov (2004) for mud volcanos emissions; and Lambert and Schmidt (1993) for oceanic exchanges emissions.

### **2.2.2 Use of Observations to Estimate Emissions (Top-down)**

Although bottom-up emissions inventories are comprehensive and contain a large amount of information at various levels, they have two distinct weaknesses. First, they involve a large range of estimations due to inaccurate data collection and missing information. Second, they are not directly constrained by atmospheric observations to retain a closed-form global budget. Hence, CTMs driven by bottom-up emissions usually disagree with measurements. Evaluating those disagreements may lead to errors that originated in the bottom-up estimation, emphasizing room for improving them. This can be accomplished with top-down estimations that integrate atmospheric observations with CTMs to constrain methane emissions. Using top-down constraints, although it is hardly possible to estimate each underlying emissions process individually, a combination of several sources or sinks (i.e., emissions categories) or the entire budget can be estimated.

This provides a closed-form estimation at the global scale (or across the domain) with smaller uncertainties, while a better agreement can be achieved with observations (Saunois et al. 2020; Jacob et al. 2022).

In principle, there are two methods to optimize emissions using observations or top-down constraints. In the first method, the observed surface air concentrations are used while the bottom-up information on the emissions is completely ignored, and thus emissions are implicitly obtained from the inverse method using a mass balance (Butler et al. 2004; Baray et al. 2018). The mass balance approach is typically used to estimate local emissions at higher spatial resolutions, particularly point sources such as coal mines (Varon et al. 2021). Airborne platforms have been used for a decade to constrain emissions using this approach (Bradley et al. 2011; Frankenberg et al. 2016; Hulley et al. 2016). In addition, satellite observations such as CarbonSat (Buchwitz et al. 2013) and GHGSat (Jervis et al. 2021) were also designed to estimate methane (e.g., methane plumes) at very fine spatial resolution (<100 m). The second top-down approach is more prevalent as it offers a variety of applications on different scales. It uses observations to apply correction factors to the bottom-up emissions estimates, aiming at a better fit for the observation network. This top-down approach, also used in this thesis (Chapter 7), integrates statistical optimization frameworks and is commonly known as inverse modelling. Due to its widespread application, top-down corrections are commonly referred to as inverse modelling (Brasseur and Jacob 2017).

Inverse modelling integrates different types of observations. Satellite observations, due to their higher density and spatial coverage globally (Palmer et al. 2021), have been used extensively over the past decades to infer methane emissions on different scales. Jacob

et al. (2016; 2022) reviewed the use of various satellite observations in an atmospheric inverse modelling context. Despite their sparsity, in situ measurements maintain higher precision than satellite instruments, so that they are often used for validation purposes (Liang et al. 2017). In this thesis, GOSAT satellite observations (Section 3.1.1) are used in the inversion, and four types of ground network and aircraft measurements are used for evaluation (Section 3.1.2). Other than observations, inverse methods rely on CTMs and some pieces of information from bottom-up inventory, including the magnitude of the prior emissions, their spatial distribution and temporal variability. Section 3.2 illustrates the model and prior information involved with the inversion system. The mathematical framework of an atmospheric inversion system is typically based on a Bayesian estimation theory, described in Chapter 4.

Inversion methods are divided into two categories: variational optimizations and filtering methods (for details, see Sections 4.1 and 4.2). Theoretically, both of them solve the same problem aiming at a maximum a-posteriori (MAP) solution (i.e., maximum posterior probability density), yet from different paths. In the variational method, a quadratic cost function of the error-weighted difference between observations and model is minimized numerically. This approach often requires cost function gradients with respect to emissions (or variable to be estimated), which are obtained from the adjoint of the model (see Section 3.2.3). Although it is generally considered a low-cost method of obtaining emissions estimates at the model native resolution, the error statistics are not part of the solutions unless at the cost of additional computations (Bousserez et al. 2015). The four-dimensional variational (4D-Var) method is a standard variational approach for source estimation (see Section 4.1) that has been widely applied in the estimation of the methane

budget (Bergamaschi et al. 2015; Parker et al. 2015; Houweling et al. 2017; Bergamaschi et al. 2018; Bousserez and Henze 2018; Yu et al. 2021). Chapter 7 of this thesis adopts a form of 4D-Var with a modified cost function.

In the filtering method, on the other hand, the MAP solution is obtained using probability theory, in particular, from a minimum variance estimation approach. Theoretically, one can show that the filtering method is equivalent to the estimation from the cost function minimization (Asch et al. 2016). The filtering method in the context of emissions inversion can also be viewed in two forms: (i) analytical inversion that performs a deterministic estimation for a linear Gaussian problem (Kopacz et al. 2010; Turner et al. 2015; Turner and Jacob 2015; Wang et al. 2019; Ma et al. 2021; Maasakkers et al. 2021; Lu et al. 2022; Zhang et al. 2021), and (ii) ensemble method (e.g., EnKF) that relies on stochastic estimation while approximately characterizing the optimal solution (Whitaker and Hamill 2002; Peters et al. 2005; Zupanski et al. 2007; Chatterjee et al. 2012; Feng et al. 2022). Both of these methods are applicable to high dimensions and various scales. Hence, in the context of inverse modelling, three approaches, including 4D-Var, analytical inversion, and EnKF, are widely used for methane emissions estimation in the research community. Although each method has its own strengths and weaknesses, they might also face similar limitations that are sometimes inherent in the top-down estimation. Below we summarize some of those gaps while comparing these methods.

### **2.3 Limitations and Challenges in Methane Inversion**

The behaviour of all inversion systems, including those mentioned above, can be sensitive to various factors. Model configurations such as spatial and temporal resolutions, prior, background, and observation error covariances, and size of ensembles in the case of

the EnKF system, are among those factors that influence the result of inversions (Peters et al. 2005; Babenhauserheide et al. 2015; Yu et al. 2021). Furthermore, the choice of inputs, particularly the prior estimate (i.e., first guess for optimization) of emissions and the observation network, is also quite important. Although theoretically, we may expect a unique solution regardless of the prior estimate, it has been shown that the results can still substantially depend on the choice of prior, particularly over areas where the observation constraints are limited (Bergamaschi et al. 2018; Maasakkers et al. 2021). Prior estimates are typically derived from global bottom-up inventories (e.g., EDGAR) that sometimes contain large amounts of uncertainties (Miller et al. 2013; Turner et al. 2018). Nevertheless, some recent studies have independently developed more detailed bottom-up emissions estimates at the regional or country level (Maasakkers et al. 2016; Scarpelli et al. 2020; Scarpelli et al. 2022), aiming to complement the global inventories and maintain a more accurate prior for their inverse analysis (Qu et al. 2021; Zhang et al. 2021; Worden et al. 2022).

Besides model input and configuration, other factors must be addressed before performing a top-down estimation. We know that every inversion method requires the feedback (or signal) of the emissions at the observation space. Chemical transport models (CTMs), together with the observation operator, are the two main components responsible for simulating realistic atmospheric methane in observation space. However, there are potential errors in both CTMs and observations. It is known that for methane, contrary to other gases, the emissions signal in satellite observation space is fairly weak (e.g., < 1% on a regional scale) and linear (Jacob et al. 2016; Saunio et al. 2020). In fact, as mentioned earlier (Section 2.1.2), methane is characterized by fairly linear chemistry. In addition,

satellite retrievals typically observe a total column weighted amount of concentrations, so that in the presence of a large background burden and well-mixed tropospheric methane, they show a rather weak and near-linear sensitivity to the air mass changes near the surface (due to emissions) (Jacob et al. 2016). Hence, this entails a more stringent constraint on the quality of the modelled and observed methane concentrations. In other words, to obtain a reliable estimation of emissions, we must reduce the model state uncertainties to a level comparable to and even smaller than the emissions signal. This requires a more accurate model with detailed and efficient sub-processes, which is computationally prohibitive to maintain given the currently available computational resources. Otherwise (with a less accurate model), a perfect model assumption that is frequently made in analytical inversion and 4D-Var can result in a degraded inversion performance. Note that computational cost is another important factor for inverse modelling analysis.

Locatelli et al. (2013) showed that inverse modelling running with different atmospheric models results in up to 150% discrepancy in emissions estimates, although those discrepancies may have different origins in the model and underlying assumptions. The primary factors responsible for those differences are the meteorological field, model physical parameterization, stratospheric transport, and model spatial and temporal resolutions (Stanevich et al. 2020; 2021). For example, for computational purposes, CTMs are usually driven using pre-computed meteorological fields. Although those meteorological fields are usually improved with observation assimilation, their uncertainties are not taken into account in CTMs, which results in limited predictability of the model state (Monge-Sanz et al. 2013; Polavarapu et al. 2016; Stanevich et al. 2020). In addition, if the CTMs' spatiotemporal resolution largely differs from meteorological

resolution, it may lead to a misrepresentation of key dynamical aspects of the atmosphere (Richter et al. 2014). Model physical parametrization, such as one for convective subgrid-scale, is another source of CTMs error, especially when the meteorological field uses a different parametrization (Orbe et al. 2017).

One way to address the model transport errors, including those indicated above, is to account for them during the inversion process. For example, a variant of 4D-Var, known as weak constraint 4D-Var (see Section 4.1), accounts for the model transport error. Stanevich et al. (2020; 2021) developed weak constraint 4D-Var with the GEOS-Chem model that simultaneously optimizes methane emissions and model error forcing. They showed that if the model error is not fully addressed in the global methane inversion, it can result in up to a 35% bias in monthly methane emissions estimates. However, as mentioned earlier, CTMs can also suffer from other types of error besides the model transport, such as inaccurate initial conditions, deficient chemical simulation, insufficient assumptions of the model top boundary conditions (e.g., zero flux assumption), etc., that may not be resolved through weak constraint 4D-Var.

Computational cost is another significant limitation for inverse modelling problems. For instance, although a weak constraint 4D-Var is relatively robust to address the model transport error, it requires much more computational cost than a regular 4D-Var (strong constraint 4D-Var) inversion, especially for high spatial resolutions (Stanevich et al. 2020). Analytical inversion also suffers from high computational costs. It requires an explicit construction of the Jacobian matrix (i.e., observation operator for an inversion system), for which the computations rely on either the dimension of the state or observation space. Hence, for a typical high-dimensional atmospheric inverse problem with a

significant number of observations, analytical inversion computations are insurmountable (Berchet et al. 2021). There are yet other categories of inversion methods with an intermediate adaptation of analytical inversion (i.e., sequential analytical inversion) that aims to tackle some of those limitations while maintaining a lower computational cost (Brunner et al. 2012b; Miller and Michalak 2017; Miller et al. 2019). However, despite the moderate cost (e.g., equivalent to EnKF), these inversion methods are also limited to particular linear and simple cases. In EnKF inversion also, the number of ensembles imposes a computational constraint. For a small number of ensembles (e.g.,  $< 30$ ), posterior uncertainties may not be reliable, resulting in spurious errors with unrealistic correlations that degrade the inversion results. Localization of the ensemble and error inflation are two techniques used to enhance the consistency of inversion results (Zupanski et al. 2007; Babenhauserheide et al. 2015; van der Laan-Luijkx et al. 2017).

Input error covariances (or uncertainties), including observation error covariances  $\mathbf{R}$  and prior error covariances  $\mathbf{B}$ , can also result in a degraded estimation if not adequately determined (Daley 1992a; Tandeo et al. 2020). In both analytical inversion and 4D-Var, in which the cost function depends on error covariances,  $\mathbf{B}$  and  $\mathbf{R}$  are both taken static within a single assimilation window (not updated over time and through iterations for the assimilation window). In many practical applications, they are also assumed to be diagonal and with spatially uniform error weights. Yu et al. (2021) showed that using a spatially varying emissions error and accounting for the correlation in  $\mathbf{B}$  (simple exponential decay correlation model), can improve the inversion results, particularly for estimating heterogeneous and missing sources. A conventional method in inverse modelling to make the fit closer to observations is to balance the weight between  $\mathbf{R}$  and  $\mathbf{B}$ , using a

regularization parameter (Hakami et al. 2005; Kopacz et al. 2010). It is commonly used to compensate for the missing objective information in quantifying error correlation in  $\mathbf{R}$ , which is often assumed diagonal (Lu et al. 2022); however, a single coefficient may not provide sufficient information to resolve  $\mathbf{R}$  (for details, see Chapter 7).

## 2.4 Use of Data Assimilation for Methane

Besides the above challenges inherent to the inverse estimation problems, emissions and their error statistics in the regional (rather than global/hemispheric) domain require precise knowledge of the initial and boundary conditions together with their uncertainties. Although the initial conditions' impact can fade over a long period of time with a (perfect) model spin-up, errors in the boundary conditions will persist and can significantly impact the emissions estimation, particularly for a limited regional domain (Berchet et al. 2013; Wecht et al. 2014). We recall from Section 2.3 that the emissions signals are significantly weaker than the background or lateral boundary inflow concentrations. From an estimation point of view, the effect of the initial or boundary conditions can contaminate the signal needed to constrain the emissions at the surface. There are, however, three approaches aiming to resolve this problem:

- (i) A significantly long model integration at a larger scale (e.g., global scale), considering methane has a lifetime of almost 10 years in the troposphere, diminishes the effects of the initial field and boundaries. However, for that purpose, besides the computational expenses of extended simulations, we need a perfect and unbiased model that is hard to achieve;
- (ii) Initial/boundary conditions and emissions can be estimated simultaneously (using one cost function) within an inverse modelling system. Although this approach has

been used in several studies (Wecht et al. 2014; Jiang et al. 2015), it is complex to implement, and it may not lead to a convergence of emissions, especially if the state uncertainties are still large compared to the emissions signal. Additionally, it may not result in dynamically coherent boundary conditions (Turner et al. 2015; Stanevich et al. 2021); or,

- (iii) Performing data assimilation on a larger domain and using the assimilation analysis with its error statistics over the region of interest can outperform the previous methods. It not only provides smooth and dynamically consistent initial/boundary conditions that favour emissions estimates, but is also capable of separately estimating concentrations and their error statistics (e.g., using a separate cost function).

Performing reliable assimilation that is capable of reducing the uncertainty of the concentrations to a level comparable to the emissions signal can result in a more accurate emissions estimation. However, there is a lack of independent research on methane data assimilation (as opposed to many studies in methane inverse modelling), and those few assimilation studies are limited to a particular case and method, which may have limitations in the realistic estimation of methane and error statistics (Massart et al. 2014). Assimilation methods, in principle, are similar to inversion approaches indicated above, yet the target vector (i.e., control variables) to be estimated is comprised of concentrations rather than emissions. Accordingly, these methods may face similar limitations as the inversion methods, such as high computational costs. Hence, this thesis first focuses on developing a novel cost-effective assimilation framework called PvKF (Chapter 5). There is also a lack of independent and objective evaluation of methane error statistics in the literature

(whether in assimilation or inversion context). That is discussed in Chapter 6, while a new methodology is demonstrated for the realistic estimation using satellite observations. Finally, we explore the combination of this assimilation system with a 4D-Var inversion framework in Chapter 7.

## **2.5 Background of Assimilation Methods**

Data assimilation in the atmosphere aims to estimate the state of the system by combining observations with an atmospheric model (Asch et al. 2016). In atmospheric CTMs, for example, the state might involve the concentrations of a chemical species that is usually resolved to a scale such that the state vector is sizeable, on the order of  $10^6$  to  $10^8$  (degree of freedom). Daley (1992a; 1992b; 1992d) has argued that because of the error/noise embedded in the observations network and the model forecast or both, a reliable and realistic representation of the state cannot be determined adequately without having the knowledge of its uncertainties. The size of the covariance matrix (i.e., uncertainty matrix) thus is significantly greater than the state (on the order of  $10^{12}$  to  $10^{16}$ ), making the practical application based on Kalman filtering theory almost impossible (Menard et al. 2000; Segers et al. 2005; Pannekoucke et al. 2018b), considering the current capacity of supercomputers. From the Bayesian inference point of view, it is complicated to allocate a prior distribution to each element of an input covariance matrix of this size (i.e., observation and model error covariances (Daley 1992a; Dee 1995; Tandeo et al. 2020); hence, to make the estimation of error covariances tractable, a parametric form of a covariance matrix, also referred to covariance modelling, is used in data assimilation (Menard et al. 2016; Satterfield et al. 2018). Several methods have been developed over the past three decades to overcome the limitation of estimating the error covariances within

the assimilation scheme. Ensemble Kalman filter (EnKF), 4-dimensional variational (4D-Var), and hybrid ensemble-variational (En-Var) methods are among those common approaches.

EnKF originates from a combination of the Kalman filter theory and the Monte Carlo estimation method (Evensen 1994; 2009b). The applicability of the EnKF methods in a large state space arises from the use of a limited number (e.g., dozens) of ensemble members in combination with covariance localization (Hamill et al. 2001). To approximate the error covariance matrix, localization is employed to eliminate spurious correlations at large distances and increases the rank of the sample covariance to a value comparable to the dimension of the state space (as required by the assimilation algorithm); thus, it can maintain a full rank forecast error covariance to assimilate observations (Houtekamer and Mitchell 2005). Besides the added computational expenses, the major drawback of localization is that it might neglect some physical and true long-distance correlations, which can induce physical imbalance in the analysis (Greybush et al. 2011). Ensemble methods are well-adapted for nonlinear estimation and have been used in many atmospheric problems, such as in numerical weather prediction (Buehner 2005; Houtekamer and Mitchell 1998; 2005; Houtekamer and Zhang 2016; Kurosawa and Poterjoy 2021; Lorenc 2003) and in atmospheric chemistry (Peng et al. 2015; van der Laan-Luijkx et al. 2017; Kong et al. 2019; Tenkanen et al. 2021). Although most EnKF studies for methane are restricted to emissions estimation, the formulation and the cost of an assimilation system are almost the same as an equivalent inversion system.

The applicability of the 4D-Var assimilation algorithm to large state space arises from introducing the adjoint of operators combined with the use of an initial or background

error correlation, often assumed to be homogeneous and isotropic (Courtier 1997; Courtier et al. 1994; Gauthier et al. 2007). Accordingly, a sequence of operators representing the effect of multiplication of error covariance with a state vector prevents the storage of extremely large covariance matrices. Despite the significant number of 4D-Var methane inverse modelling studies, there are few works that performed pure methane assimilations. For example, Massart et al. (2014) investigated the impact of transitioning the methane observing system from Scanning Imaging Absorption Spectrometer for Atmospheric Chartography (SCIAMACHY) to Greenhouse Gases Observing Satellite (GOSAT) and Infrared Atmospheric Sounding Interferometer (IASI) using their 4D-Var methane assimilation system. By comparing their assimilation results against independent surface and aircraft observations, they examined the global bias and uncertainties in their analysis concentrations. In another study, Errera et al. (2016) assimilated new methane profiles retrieved from Michelson Interferometer for Passive Atmospheric Sounding (MIPAS) with their 4D-Var method. They examined the quality of the analysis against independent ACE-FTS observations.

Although both EnKF and 4D-Var estimation methods are identified as robust and reliable methods that have been running operationally for a long time in various atmospheric centers, they suffer from some drawbacks that need to be considered, particularly for the assimilation of long-lived species such as methane. The first and most important downside is that they both require a relatively high computational cost, which is typically several tens of times as expensive as model integrations (Chatterjee and Michalak 2013). For example, 4D-Var assimilation entails the adjoint model. Besides that, several forward-backward iterations are required until the optimization (e.g., steepest descent)

algorithm is converged. Hence, at least tens of model integration are necessary to properly implement 4D-Var assimilation (Massart et al. 2014). The computational time of EnKF assimilation depends on the number of ensembles carried out with the model that adopts it. Typically, dozens of ensembles are expected to maintain reliable assimilation (Houtekamer and Zhang 2016); for example, Peng et al. (2015) used 48 ensembles in their CFI-CMAQ for optimization of CO<sub>2</sub> surface flux. In addition to the high computational cost and the necessary localization, the variance loss that is likely to occur in Kalman filter and ensemble Kalman filter systems degrades the assimilation performance and results in a filter divergence (Menard et al. 2021; Pannekoucke et al. 2021; Gilpin et al. 2022). To alleviate this, one requires inflation of error variance to restore the loss of variance at the grid level (Menard et al. 2021).

## **2.6 Potential Use of Parametric Filtering for Methane Assimilation**

There is another category of assimilation methods with error estimation capability, in which the error covariance dynamics of Kalman filtering, including evolutions of forecast and analysis error covariance matrices, are approximated by covariance models with a parametric formulation (Khattatov et al. 2000; Menard and Chang 2000; Pannekoucke et al. 2016). The details of the covariance modelling approach are demonstrated in Section 3.3 and Appendix A, and a short review of its background literature is presented later in this section. The parametric propagation of the covariance model relies on the time evolution of the error variance and the associated correlation length scales. This assimilation method is also identified recently as a parametric Kalman filter (PKF) in CTMs (Pannekoucke et al. 2016).

One main cause of emerging the parametric form of Kalman filtering was to avoid the challenges and drawbacks of the previous assimilation approaches (e.g., KF, EnKF, and 4D-Var), particularly the high computational cost of assimilation. The high computational cost in standard Kalman filtering occurs primarily because of the creation and evolution of the error covariance matrices in a high dimensional discretized state space problem. However, in the parametric Kalman filtering, those large matrices are created efficiently through covariance modelling (see Section 3.3) and evolved using a continuum formulation of the transport dynamics (e.g., advection scheme). The idea of the evolution of the error variance with an advection scheme emerged originally in Kalman filtering by Cohn (1993). This will be explained in detail in Section 4.3.1. He demonstrated the continuum properties of the error covariances for quasi-linear hyperbolic equations and found that the error variance can be simply predicted without having the knowledge of the error correlations. One of the first implementations of this method in a realistic atmospheric context was demonstrated by Menard et al. (2000) for stratospheric assimilation of a long-lived chemical tracer. Menard and Chang (2000) also used the same scheme with model error estimation. They showed that the residual model error growth in standard KF is unrealistic and can reach two times larger than the one obtained with the continuous variance evolution. The use of continuous covariance evolution with an appropriate correlation scheme leads to the development of simplified Kalman filtering in atmospheric models, also known as the sequential filter (Khattatov et al. 2000).

Sequential filtering was used as an efficient method in various atmospheric data assimilation schemes, particularly for estimating long-lived chemical species such as stratospheric ozone. For example, Eskes et al. (2003) applied the method for the

assimilation of GOME total-ozone and examined the analysis result over a two years period by comparing it against independent observations from TOMS and Brewer instruments. In another study, Rosevall et al. (2007) performed sequential filtering of ozone data from the ENVISAT/MIPAS and Odin/SMR satellite instruments to analyze the polar ozone depletion and compare the capabilities of those two satellite observations. Using the sequential filter, Parrington et al. (2008) assimilated ozone and CO retrievals from TES observations into GEOS-Chem and AM2-Chem models. They showed that TES measurement is sufficient to constrain the tropospheric ozone and to improve the description of free tropospheric ozone. van der A et al. (2010; 2015) also performed sequential assimilation from fourteen satellite retrieval datasets to create a global and comprehensive dataset of total ozone analysis over 30 years of assimilation. They found that the analysis statistics (i.e., observations-minus-forecast) decreased to less than 1% for the bias and about 2% for the RMS standard deviation. Zoogman et al. (2014) conducted joint assimilation of ozone and CO based on the sequential filtering method with the aim to improve the surface ozone concentrations. Using an Observing System Simulation Experiment (OSSE), they found a substantial benefit from the joint assimilation to informing U.S. ozone air quality.

Now we briefly review some techniques in the literature used for covariance parameter estimation (see Section 3.4 and Appendix A for the details and formulation of those popular methods), which is the central part of the covariance modelling used in parametric Kalman filtering. Covariance modelling is a common approach in data assimilation aiming to determine the covariance matrices by a few parameters that describe the magnitude and shape of the errors (Tandeo et al. 2020). Several techniques have been

conducted to optimally estimate the error covariance parameters, which are primarily based on the residual between the model/analysis and observations in the observation space (Menard 2016). One of the first estimation approaches is proposed by Hollingsworth and Lonnberg (1986) in the context of optimal interpolation, where a spatial autocorrelation function describes the Observation–Model (i.e., innovation) residual (Rutherford 1972) (see Appendix A for the details). A fitting process is applied to the spatially correlated background errors, which leads to estimating the error variance of observations and background along with the correlation length scales. Dee et al. (1999) developed another form of estimation method based on maximizing the likelihood of the observations (i.e., maximum likelihood) that yields an estimation based on the innovation (see Appendix A for the details). Many atmospheric studies applied this method for real data assimilation applications (Ueno et al. 2010; Liang et al. 2012; Pulido et al. 2018). Another well-known approach that is used in many data assimilation (and inverse modelling) studies is developed by Desroziers et al. (2005). The method is based on different types of innovation statistics, either from a difference between observations and analysis or between model and analysis in observation space (see Appendix A for the details and discussions).

Despite the popularity of the two past methods due to their simple adaptation to various problems, they are limited, mainly due to providing an estimation of error statistics only in observation space, so that the estimation results can be inadequate to prescribe the assimilation error statistics (Menard 2016). Other error covariance estimation approaches are based on model outputs, which surpass the limitation mentioned above. A method that characterizes an ensemble of data assimilation trajectories (Fisher 2003) and the National Meteorological Center (NMC) lagged-forecast method (Parrish and Derber 1992) are

among those. However, they have their own challenges. For instance, the ensembles method, besides the high computational demand for generating those ensembles, requires a sophisticated tuning of model error due to inflation and localization (Menard et al. 2019; Menard et al. 2021). The lagged forecast method also highly depends on a complete observational coverage; thus, a practical implementation is not achievable in extensive areas with no observations (Osinski and Bouttier 2018).

Although the statistical diagnostics mentioned above can provide a reasonably accurate estimate of the error covariance parameters, they are all derived based on the assumption that the underlying assimilation system results are optimal (Desroziers et al. 2005; Menard 2016; Waller et al. 2016a; Tandeo et al. 2020). Indeed, from the Bayes theorem perspective, the estimate of the state can be obtained by either maximizing a posteriori probability or minimizing the error variance, both of which systematically provide an optimal solution. However, this optimal solution may not represent the true analysis unless the input error covariances, including the observations and model error covariances and the correlation lengths, also represent the true errors corresponding to the optimal solution (Menard and Deshaies-Jacques 2018a).

A new error covariance estimation method was recently proposed by Menard and Deshaies-Jacques (2018a; 2018b), which does not rely on the optimality assumptions (i.e., assumptions made through optimal estimation theory) of the assimilation system. This approach revolves around cross-validation optimization that uses independent observations (also called passive observations) to diagnose the analysis error. Using the cross-validation estimation technique, they showed that the true analysis error variance can be obtained without assuming optimal assimilation. Accordingly, they found a minimum analysis error

variance corresponding to a single tunable parameter of the correlation model (i.e., horizontal correlation length scales) by minimizing the cross-validation cost function (see Appendix A for the details and equations). Thus, the analysis formed is (nearly) optimal (only because the estimation is performed with a covariance modelling than the full covariance matrix), and the error statistics obtained are close to the true error statistics. The method was then applied to the GEM-MACH model of Environment and Climate Change Canada (ECCC) to generate the hourly surface ozone and PM<sub>2.5</sub> analysis using in situ observation (Menard and Deshaies-Jacques 2018a).

This thesis proposes a novel cost-efficient framework with parametric filtering capabilities, referred to as PvKF, for assimilating methane concentrations. Since methane is characterized by almost linear chemistry, a modified and even less expensive version of parametric Kalman filtering, in which only the error variance evolution is computed explicitly while the correlation lengths are kept fixed over the assimilation window, is derived. The details of the development of the assimilation, including its algorithm, formulation, and verifications, will be explained in Chapter 5 (Voshtani et al. 2022a).

The PvKF scheme can be classified as part of a suboptimal filter due to the approximation of the covariance matrix and covariance propagation (Asch et al. 2016; Pannekoucke et al. 2021). Therefore, estimating appropriate covariance parameters within the covariance modelling can aid in achieving a more reliable and realistic analysis. Since cross-validation is a promising technique for maintaining that goal, this research extends cross-validation application to GOSAT satellite observations and multiple covariance parameters estimation. The details of the assumptions and the modifications, along with

the evaluation of the assimilation analysis against independent sources of observations, are demonstrated in Chapter 6 (Voshtani et al. 2022b).

As described earlier in this section regarding the importance of mitigating methane emissions, this research also aims to address some limitations of the source inversion method (e.g., 4D-Var inversion) used in the past and assists in advancing their capabilities in capturing methane emissions. This is achieved by linking a cost-effective and optimal assimilation system developed earlier to a standard 4D-Var source inversion and examining the effect of optimal state estimation on reproducing known emissions (using OSSEs). The details of this demonstration are provided in Chapter 7 (Voshtani et al. 2022c).

### **Chapter 3: Data and Research Tools**

In its main direction, this research aims to develop an assimilation system for predicting atmospheric methane concentrations and further perform an inversion to constrain methane emissions. Every atmospheric data assimilation and inversion system requires two categories of information: (i) an atmospheric model (e.g., CTMs) driven by model inputs and a set of governing equations that result in predicting the evolution of atmospheric constituents (i.e., forecasting); and (ii) a set of measurement data or observations that monitor the change of their concentrations. In an assimilation system, the model is used to be combined with observations for obtaining an improved methane state (concentrations) prediction, while in an inversion system, the model is mainly used to provide a linkage between model inputs/parameters (e.g., emissions) and state concentrations; thus, it aims to estimate those model inputs/parameters. Note that concentrations are the observed states while emissions are considered unobserved inputs/parameters. This research uses satellite observations, mainly GOSAT, and the CMAQ air quality model to estimate both states and inputs/parameters.

Methane observations are obtained from various measuring platforms, including ground network measurements, aircraft, ships, and satellites. Due to their greater coverage and frequency, satellite observations are used to operate assimilation/inversion, while other types of observations are mainly employed for validation purposes due to their higher precision. CMAQ is an atmospheric CTM, on which the assimilation system relies for the prediction of the methane concentrations and evolution of the error variances. It is also a key component in performing source inverse modelling through linking methane emissions to concentrations. In this section, we first briefly describe all types of methane observations

used in this research, either satellite observations that are used within the assimilation/inversion system or other types of observations such as in situ and aircraft observations used for evaluations. Secondly, we provide an overview of the tools and methods driving the CMAQ model, particularly methane emissions inventory inputting the model.

### **3.1 Observations**

#### **3.1.1 GOSAT Satellite Observations**

GOSAT was launched in January 2009 by the Japanese Space Agency (JAXA) (Kuze et al. 2009). It is in a Sun-synchronous orbit at an altitude of 666 km with a 3-day revisit time. GOSAT's primary objective is to monitor the abundance of greenhouse gas, including atmospheric methane, globally. Owing to the high near-surface sensitivity of GOSAT retrieval and acceptable spatiotemporal resolution, the assimilation of atmospheric methane and inverse modelling of its sources and sinks are desired (Butz et al. 2011; Schepers et al. 2012; Parker et al. 2015; Turner et al. 2015; Buchwitz et al. 2017).  $XCH_4$ , retrieved from GOSAT, is a column-average dry-mole fraction of methane corresponding to the methane average volume mixing ratio (VMR) of a partial column atmosphere with a particular surface and top pressure. Methane VMR is obtained by performing a retrieval algorithm on the radiance spectrum. There are two retrieval algorithms, including Full Physics (FP) and Proxy (PR); the former integrates a sophisticated radiative transfer model and solely relies on methane modelling and its corresponding errors, while the latter provides more data points but relies on an accurate  $CO_2$  model simulation and its retrieval ( $XCO_2$ ) (Schepers et al. 2012). Both algorithms were developed at the Netherlands Institute for Space Research (SRON) and Karlsruhe Institute for Technology (KIT) (Butz et al.

2011), and their products are available through the ESA GHG-CCI initiative, <https://climate.esa.int/en/projects/ghgs/> (Buchwitz et al. 2017). Both of the algorithm's products are used in this research, with the main focus on the PR method due to its higher density and higher spatial coverage (Butz et al. 2011; Schepers et al. 2012).

For assimilation/inversion purposes, in addition to the retrieval data ( $y^o \equiv \text{VMR}$ ), we need supplementary products (Bovensmann et al. 1999; Buchwitz et al. 2017). Each retrieval includes vectors of the normalized column-average kernel,  $\mathbf{A}$ , pressure levels,  $p_l$ , at which the average kernels are derived, and the corresponding vector of a priori,  $\mathbf{y}^p$ . For GOSAT, a layer-based approach described in Bergamaschi et al. (2007) is applied to compute the model partial-column value,  $y^m$ , equivalent to the retrieval,  $y^o$

$$y^m = [(\mathbf{1} - \mathbf{A})\mathbf{y}^p + \mathbf{A}\mathbf{y}^m] \boldsymbol{\omega}^T \quad (3.1)$$

where  $\mathbf{y}^m$  represents the mapped concentration of CMAQ on the pressure layers of the observations, and  $\boldsymbol{\omega}$  is the vector of pressure layers weights

### 3.1.2 Ground Network and Aircraft Observations

#### 3.1.2.1 TCCON

Total Carbon Column Observing Network (TCCON) is a ground-based Fourier-Transform Spectrometer (FTS) network that provides a time series of  $\text{CH}_4$  column-averaged abundance. It is a standard measurement against model simulations and satellite data, while being primarily used for validation purposes (Yoshida et al. 2013; Scheepmaker et al. 2015; Zhou et al. 2016; Liang et al. 2017; Wunch et al. 2019; Stanevich et al. 2021; Zhang et al. 2021). In this thesis proposal, the GGG2014 version of TCCON XCH<sub>4</sub> data from seven sites, including Park falls (Wennberg et al. 2017), Orleans (Warneke et al.

2017), Lamont (Wennberg et al. 2016), Bremen (Notholt et al. 2019), Sodankyla (Kivi et al. 2014), Izana (Blumenstock et al. 2017), and Bialystok (Deutscher et al. 2019) are used, all of which are available at <https://tccodata.org/2014>. TCCON is calibrated using aircraft profiles to maintain more than 0.5% accuracy of XCH<sub>4</sub> retrievals (Wunch et al. 2017). In order to compare it with the model or analysis, we need to convolve model concentration with TCCON average kernels and the corresponding a priori profiles, for which the procedure is described by Wunch et al. (2010) or through the website <https://tcon-wiki.caltech.edu/Main/AuxiliaryData>. Similar to the aircraft measurements, TCCON retrievals are assumed to provide an independent evaluation for assimilation results. We note that any influence of TCCON calibration on GOSAT data prior to this research will be entirely alleviated since we applied an independent bias correction on GOSAT using surface observations (see Section 5.4.2 for details).

### **3.1.2.2 HIPPO-3**

HIAPER Pole-to-Pole Observations (HIPPO-3) provides methane concentration measurements (Wofsy et al. 2011) from the surface to 14 km across the Pacific Ocean between 20 March and 20 April 2010. The measurement is conducted every second using a quantum cascade laser spectrometer (QCLS) with an accuracy of 1 ppb. Methane measurements, meteorology and flight tracking data are accessible at: [https://www.eol.ucar.edu/field\\_projects/hippo](https://www.eol.ucar.edu/field_projects/hippo). The measurement data for the hemispheric domain of CMAQ are available on 10, 13 and 15 April 2010. We interpolate the model concentration at the specific location, height, and time of the measurement to perform verification on our assimilation result. The significantly shorter time scale of the measurements (i.e., 1 second) compared to the model simulation time step (i.e., 1 hour)

causes a large variation. Thus, we exclude those HIPPO-3 data that depart by more than three standard deviations from the average measurement over one minute to make a meaningful comparison.

### **3.1.2.3 UCATS-GloPac**

Global Hawk Pacific (GloPac) is an aircraft mission operated by NASA in April 2010 primarily designed to validate monitoring satellite missions and to measure trace gases in the upper troposphere and lower stratosphere. UAS Chromatograph for Atmospheric Trace Species (UCATS) was integrated into GloPac to measure methane alongside  $\text{N}_2\text{O}$ ,  $\text{SF}_6$ . It maintains an overall precision of 0.5% for methane (Hints et al. 2021). GloPac primary and supplementary data are available at <https://espoarchive.nasa.gov/archive/browse/glopac>, where the methane measurements were recorded from 7 to 13 April 2010 with a time resolution of 140 sec. Note that we use methane measurements without filtering; thus, the assimilation results are computed at every measurement.

### **3.1.2.4 NOAA-ObsPack v3.0**

GLOBALVIEWplus ObsPack v3.0 data product (Schuldt et al. 2021) published via NOAA Global Monitoring Laboratory provides high accuracy measurements of methane concentration from a variety of sampling platforms, including surface, tower, aircraft, and shipboard measurements. Since the surface flask and tower data are used in calibrating GOSAT (see Section 5.4.2 for details), we only use ObsPack aircraft data to ensure an unbiased comparison with the analysis. Accordingly, only daily observations above 800 m from the dataset of ObsPack, available at <https://gml.noaa.gov/ccgg/obspack/>, is used to

represent the aircraft measurements, which are collected from multiple aircraft campaigns (Schuldt et al. 2021).

## **3.2 Inputs and Model**

### **3.2.1 Methane Emissions**

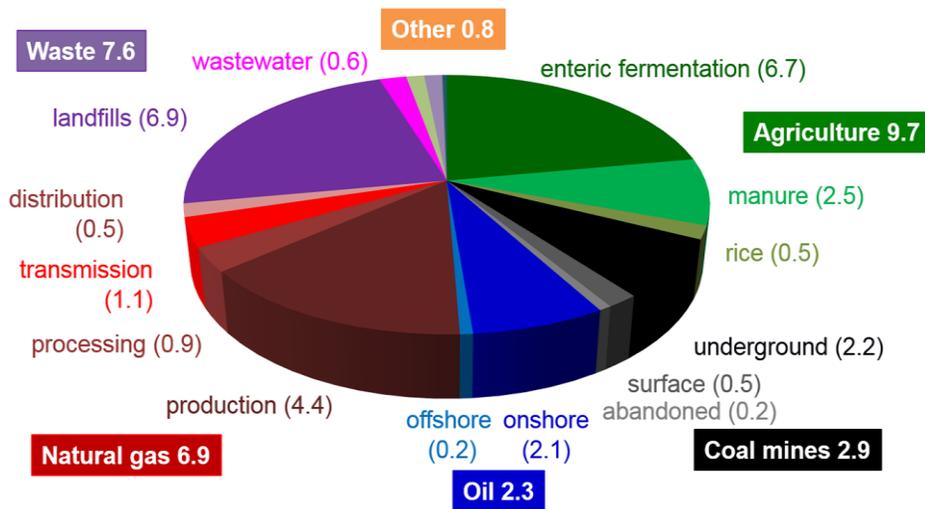
Methane emissions are generally derived from bottom-up inventories with two categories of anthropogenic (~ 60%) and natural emissions (~40%) that are used in various methane climate and atmospheric studies. Both natural and anthropogenic methane can be attributed to biogenic and non-biogenic sources. Natural methane emissions, such as wetland and termites, have a biogenic origin, whereas wildfire and natural seeps are considered non-biogenic. On the other hand, anthropogenic methane, such as livestock and landfills, has biogenic traits, whereas fossil-fuel burning and waste management are considered non-biogenic (Saunois et al. 2020). There are various bottom-up (and top-down) inventories, and most of them follow IPCC guidelines (IPCC 2006) to derive emissions in different source sectors. Saunois et al. (2016a; 2020) and Minx et al. (2021) conduct a comprehensive assessment of the available methane emission inventories along with their uncertainties using a bottom-up approach.

EDGAR methane emissions inventory is commonly used as a default prior anthropogenic source in different assimilation and inverse modelling studies. It provides  $0.1^\circ \times 0.1^\circ$  spatial and monthly/yearly temporal resolution of emission gridmaps. EDGAR v4.3.2 has been used extensively in inverse modelling (or top-down) analysis, likely due to the availability of spatially resolved emissions for almost every source sector worldwide. However, several inverse modelling analyses have recently reported an over-and underestimation of EDGAR v4.3.2, particularly from oil and gas, coal mining, and landfills

(Turner et al. 2015; Maasakkers et al. 2019; Scarpelli et al. 2020). Accordingly, in some assimilation/inversion studies, an independent highly-resolved emission inventory has been developed at a national level or for a particular uncertain source sector, which replaced EDGAR estimation for that particular region or sector (Maasakkers et al. 2016; Qu et al. 2021; Lu et al. 2022). In addition, a considerable adjustment occurred in the recent version of EDGAR (v6), with a particular focus on those inaccurate source sectors (Crippa et al. 2020; Crippa et al. 2021). This aided other studies to more comfortably rely on EDGAR as a prior emission estimate in their regional or global assimilation/inversion (Minx et al. 2021). Similarly, EDGAR v6 is used as a prior emission inventory in this research in both hemispheric assimilation and inversion setup. The anthropogenic inventory contains more than 20 subsectors that are aggregated into four main source categories (Crippa et al. 2021).

Wetlands are the primary source of natural emissions, covering 86% of the total methane emissions. Monthly wetlands emissions data from WetCHARTs v3.0 with the full ensemble mean (Bloom et al. 2017) is used in this research and mapped into the domain using a uniform temporal profile. The rest of the natural methane emissions include open fires (~7%) (van der Werf et al. 2017), termites (~6%) (Fung et al. 1991), and seeps (~1%) (Etiope et al. 2009). These emissions and those from anthropogenic sources are processed using Sparse Matrix Operator Kernel Emissions (SMOKE v4.5) (UNC 2017) to provide hourly gridded methane emissions into the CMAQ model. Three anthropogenic source categories, including agriculture, waste, and energy/industry (containing oil, natural gas, coal mines, and others), are considered in the inversion (Figure 3.1). They also represent the main categories reported to UNFCCC from EPA (US-EPA 2016). Wetland is the fourth

source category that is considered for inversion in this study (see Chapter 7 for details). The rest of the source sectors, such as “others” in the anthropogenic and open fire, seeps, and termites, are processed but will not be estimated in the inversion process, mainly due to their small weight that cannot be constrained separately from satellite observations.



**Figure 3.1. Anthropogenic methane emissions inventory compiled by EPA (US-EPA 2016). Adapted from Jacob et al. (2016)**

### 3.2.2 CMAQ and its Input Processors

The dynamical part of a chemical data assimilation and inversion system essentially relies on a CTM and/or its variants (e.g., adjoint of the CTM). Besides CTMs, most assimilation/inversion systems require knowledge about the error covariances and their evolutions, which can be maintained by properly conducting covariance modelling. This section briefly overviews the CMAQ CTM model and its driving processors.

CMAQ is a limited-area atmospheric CTM developed by the U.S. Environmental Protection Agency (Byun & Schere, 2006). It is particularly used as a regional air quality model to predict the evolution of chemical pollutants in the lower atmosphere. The regional CMAQ model is driven at the lateral boundaries by the hemispheric version of this model

(referred to as H-CMAQ) or other global CTMs. H-CMAQ is used in this research for the assimilation of methane in the Northern Hemisphere. Its domain is defined on a polar stereographic projection with 187×187 grid cells horizontally, 108 km grid spacing, and 44 vertical layers extended from the surface to the model top at 50 hPa. With the aim of a better representation of transport processes, particularly for long-lived species such as methane, the vertical structure of H-CMAQ is modified to retain finer resolution than regional CMAQ above the boundary layer (Mathur et al., 2017).

Forming an Advection-Diffusion Equation (ADE) (Equation (3.2)) in every grid cell of CMAQ leads to a set of differential equations to be solved that track the evolution of species concentrations over the domain with known initial and lateral boundary conditions.

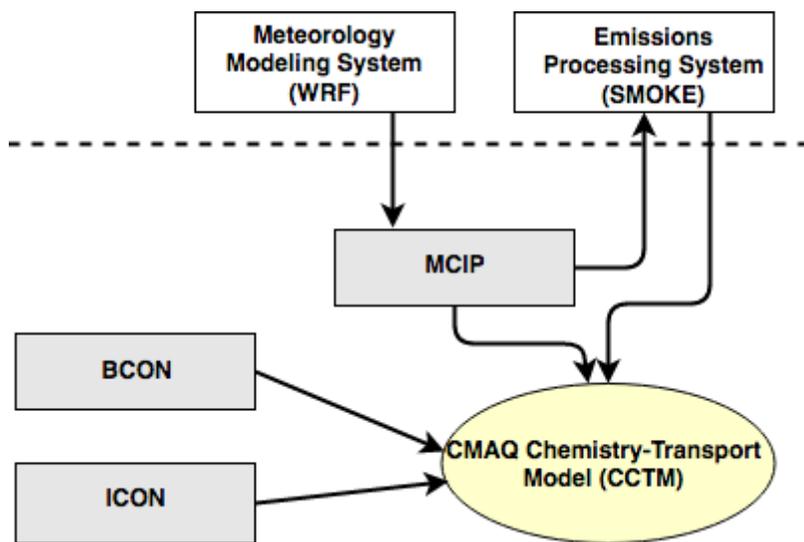
$$\frac{\partial c_i}{\partial t} + \mathbf{V} \cdot \nabla c_i - \frac{1}{\rho} \nabla \cdot (\rho \mathbf{K} \nabla c_i) + R_i = E_i, \quad (3.2)$$

where  $c_i$  is the concentration (mixing ratio) of species  $i$ ,  $\mathbf{V}$  denotes the 3D wind field,  $\mathbf{K}$  represents a diffusivity tensor,  $\rho$  is the air density,  $R$  denotes the rate of chemical reactions, and  $E$  represents the corresponding emissions rate of that species. Note that in an operator form, Equation (3.2) has the form

$$\mathbf{c}_{t+1} = \mathbf{M}(\mathbf{c}_t, \mathbf{p}), \quad (3.3)$$

where  $\mathbf{c}_0$  and  $\mathbf{c}$  are the initial and output model concentrations, respectively.  $\mathbf{p}$  includes any model parameters (e.g., emissions, initial and boundary conditions, etc.), and  $\mathbf{M}$  indicates the forward model (i.e., CMAQ), integrating all linear processes such as advection and diffusion and nonlinear processes such as chemistry.

Solving Equation (3.2) (ADEs) is performed within CCTM (CMAQ Chemistry Transport Model) processor, which interacts with other CMAQ processors as well as other systems (Figure 3.2) to obtain the necessary inputs. In general, three processors and two (independent) systems are involved: (i) ICON and (ii) BCON processors provide the CMAQ domain initial and boundary conditions of concentrations, respectively; (iii) Sparse Matrix Operator Kernel Emissions (SMOKE v4.5) (UNC 2017) is a model to process the emission inventories and generate hourly gridded emissions at the model resolution; (iv) Weather Research and Forecasting (WRF) model (Skamarock et al., 2005) creates the Meteorological inputs (e.g., wind field) that are processed in Meteorology Chemistry Interface Processor (MCIP) processors to maintain a maximum consistency with CMAQ and SMOKE.



**Figure 3.2. CMAQ Input Processors and CCTM. Adapted from USEPA CMAQ Github:**

[https://github.com/USEPA/CMAQ/blob/5.2.1/DOCS/User\\_Manual/images/Figure4-2.png](https://github.com/USEPA/CMAQ/blob/5.2.1/DOCS/User_Manual/images/Figure4-2.png)

Finally, we modified H-CMAQ to account for methane concentrations and emissions.

Methane can be configured in CMAQ either as an inert trace gas or a reactive gas-phase species oxidized by hydroxyl radical ( $\text{CH}_4 + \text{OH} \rightarrow \text{CH}_3 + \text{H}_2\text{O}$ ). We considered reactive

methane with the gas-phase model within CMAQ v5.3 and based on the CB06 chemical mechanism (CMAQ tutorials 2021).

### 3.2.3 Adjoint of CMAQ

CMAQ not only predicts the spatial and temporal distribution of a chemical species (i.e., model state concentration), but can be configured to provide sensitivities of atmospheric concentrations with respect to various model inputs. These sensitivities are useful in various applications such as sensitivity analysis, data assimilation, and inverse modelling (Hakami et al. 2006). Using a perturbation in model inputs or initials (for example, in emissions  $\delta E_i$ ), one can obtain a perturbed form of ADE as

$$\frac{\partial \delta c_i}{\partial t} + \mathbf{V} \cdot \nabla \delta c_i - \frac{1}{\rho} \nabla \cdot (\rho \mathbf{K} \nabla \delta c_i) + \mathbf{F} \delta \mathbf{c} = \delta E_i \quad (3.4)$$

where  $\mathbf{F}$  is a Jacobian of the chemistry operator. An equivalent operator form of the perturbed equation in a timestep  $[t, t+1]$  can be derived as

$$\delta \mathbf{c}_{t+1} = \mathbf{M}'_c(\mathbf{c}_t, \mathbf{p}) \delta \mathbf{c}_t + \mathbf{M}'_p(\mathbf{c}_t, \mathbf{p}) \delta \mathbf{p} \quad (3.5)$$

where  $\mathbf{M}'_c = \partial \mathbf{M} / \partial \mathbf{c}$  and  $\mathbf{M}'_p = \partial \mathbf{M} / \partial \mathbf{p}$  represent the linearized model with respect to the model state and parameter, respectively. The linearized model is also referred to as the tangent linear model (TLM), and in matrix form, is equivalent to the Jacobian of the forward model operator. We note that the chemistry effect plays a minor role in a particular case of methane modelling with a long lifetime and relatively short simulation period in our experiments. In fact, we consider that the OH concentrations are well-known and thus do not exert a tangible nonlinear effect. As a result, the gradients that construct  $\mathbf{F}$  are considered fairly small.

Considering that explicit construction of the Jacobian matrices ( $\mathbf{M}'$ ) in a large dimension is computationally prohibitive, the TLM in operator form can be used to efficiently compute the sensitivity of all model output with respect to a few inputs (source-oriented problem). However, in a receptor-oriented problem, where we seek to determine the sensitivity of a few model outputs with respect to all model inputs, one can benefit from the adjoint of TLM without explicitly constructing the forward model Jacobian (Hakami et al. 2007; Zhao et al. 2020). A brief description of how the adjoint model is used to efficiently compute those sensitivities is provided in the following.

The adjoint formulation, instead of computing sensitivities of all model outputs to one model input, considers a single scalar metric of model outputs derived from all inputs. This metric is known as an adjoint cost function, which can be any function of concentration  $g(\mathbf{c}, t, \Omega)$ , to be integrated for any particular time period  $t$  and space  $\Omega$

$$J = \int_t \int_{\Omega} g(\mathbf{c}, t, \Omega) dt, d\Omega. \quad (3.6)$$

The adjoint variable of species  $i$ , is then defined as the sensitivity of the cost function with respect to concentration at any given model time ( $\lambda_i = \partial J / \partial c_i$ ). The governing equation for the adjoint model computes the evolution of  $\lambda_i$  backward in time as

$$-\frac{\partial \lambda_i}{\partial t} = \nabla \cdot (\mathbf{V} \lambda_i) + \nabla \cdot \left( \rho \mathbf{K} \nabla \frac{\lambda_i}{\rho} \right) + F_i^T \lambda_i + \varphi_i, \quad (3.7)$$

where  $F_i^T$  is the transpose of the Jacobian of the chemistry operator in Equation (3.4), and  $\varphi_i$  is referred to as the adjoint forcing term, given as an input to the adjoint model ( $\varphi_i$  is similar to  $E_i$  for the forward model).

Variational analysis, in general, uses adjoint sensitivities (i.e., gradients of cost functions) to constrain the model parameter or model state at an earlier time. Let  $\mathbf{c}_0$  be the

initial state we want to constrain based on the general definition of the adjoint cost function in Equation (3.6). In particular,  $J$  can be taken as the difference between all observations and model outputs at any given time over the assimilation window  $[0,t]$ . For a more straightforward demonstration, the operator form of ADE as in Equation (3.5) is used. By perturbing  $\mathbf{c}_0$ , one can find the perturbed cost function as

$$\delta J = \sum_{t=0}^n \langle \nabla_{\mathbf{c}_t} J, \delta \mathbf{c}_t \rangle, \quad (3.8)$$

where  $\langle \cdot \rangle$  denotes an inner product,  $\nabla_{\mathbf{c}_t} J$  indicates the directional derivative of  $J$  with respect to  $\mathbf{c}$  at the final time  $t$ . Note that the summation is taken into account since  $J$  can be a function of time (e.g., depends on observation in  $[0,t]$ ). Thus, for a linear operator  $\mathbf{M}'$ , one can apply the following duality principle

$$\delta J = \langle \nabla_{\mathbf{c}_t} J, \mathbf{M}'_{t-1} \mathbf{M}'_{t-2} \cdots \mathbf{M}'_0 \delta \mathbf{c}_0 \rangle = \langle (\mathbf{M}'_0)^* (\mathbf{M}'_1)^* \cdots (\mathbf{M}'_{t-1})^* \nabla_{\mathbf{c}_t} J, \delta \mathbf{c}_0 \rangle, \quad (3.9)$$

where  $(\mathbf{M}')^*$  is the adjoint operator of a linear model (i.e., the linearized model operator, TLM). The successive application of the forward model can link a perturbation in the final time  $t$  to the initial time, 0. Hence, one can find the adjoint gradient at the initial time and with respect to  $\mathbf{c}_0$  as

$$\nabla_{\mathbf{c}_0} J = \sum_{t=0}^n \mathbf{M}_0^* \mathbf{M}_1^* \cdots \mathbf{M}_{t-1}^* \nabla_{\mathbf{c}_t} J. \quad (3.10)$$

These gradients can be computed in the same way for any model parameters, such as emissions. In data assimilation and inverse modelling (see Section 4.1), these gradients are used along with an optimization system to constrain model states or parameters. In this thesis, The adjoint of CMAQ is used within the modified 4D-Var inversion presented in Chapter 7.

### 3.3 Error Covariance Modelling

#### 3.3.1 Observation and Background Error Covariance

Almost every data assimilation and inversion method, besides the model inputs and variables, requires some knowledge about the error covariances (Daley 1992a). Even though most assimilation/inversion methods are derived from an optimal estimation theory (minimum variance, maximum a-posteriori, maximum likelihood, etc.), the optimal solution might not be realistic unless the input error covariances, including the observations and model error covariances and the correlation lengths, also represent the realistic ones (Menard 2016). In a high-dimensional atmospheric problem, estimating every single element of error covariances is impossible due to inadequate information to constrain those large matrices. Along the direction of variational assimilation, one can perform a stochastic (low-rank) estimation method which requires quite a few realizations or deterministic approximations that often lead to underestimating error covariances (Kopacz et al. 2010; Bousserez et al. 2015; Bousserez and Henze 2018). Another common approach to estimating the covariance matrices is to approximate them with covariance models (i.e., covariance functions) containing a few parameters. Accordingly, in data assimilation, estimating the covariance parameters turns out to be a crucial step in adequately determining error covariances (Tandeo et al. 2020). Observation error covariance  $\mathbf{R}$  and background error covariances  $\mathbf{B}$  (also forecast error covariances  $\mathbf{P}^f$ ) are the two essential matrices considered in covariance modelling.

It is often assumed that the observation errors are spatially uncorrelated. The observation error has two components: (i) the measurement error from the instrument team ( $\varepsilon^m$ ) and (ii) the representativeness error ( $\varepsilon^r$ ) from the mismatch between the subgrid-

scale represented in the observation and the model (Cohn 1997; Janjic et al. 2018). Due to the uncorrelated error assumptions, the observation error covariance takes the form of  $\mathbf{R} = (\varepsilon^o)^2 \mathbf{I}$ . A simple, but common, approach in data assimilation to estimating such a diagonal matrix is to consider a global correction factor that tunes the overall effect of the observation error. The observation error covariance to be estimated thus has the form

$$\mathbf{R}' = (f^o \varepsilon^m)^2 \mathbf{I}, \quad (3.11)$$

where,  $f^o$  is the parameter that needs to be estimated, and  $\mathbf{R}'$  represents the covariance after tuning. Note that the model error covariance and initial error covariance are also considered separately in this research for the estimation. They maintain the same form as observation error covariance and are detailed in Section 5.4.4.

Contrary to the observation error, the background errors (also forecast/analysis error over time) are often correlated, so that the corresponding background error covariance ( $\mathbf{B}$ ) can not be assumed diagonal. A background covariance matrix can be denoted as

$$\mathbf{B} = \mathbf{\Sigma} \mathbf{C} \mathbf{\Sigma}, \quad (3.12)$$

where  $\mathbf{\Sigma}$  is a diagonal matrix of standard deviations, and  $\mathbf{C}$  is the corresponding error correlation matrix. In the 3D spatial domain of methane, it is possible to account for a separability assumption between horizontal and vertical correlation (Menard et al. 2019), such that

$$\mathbf{C} = \mathbf{C}_h \otimes \mathbf{C}_v, \quad (3.13)$$

where  $\mathbf{C}_h$  and  $\mathbf{C}_v$  denote a horizontal and vertical correlation matrix. These matrices are nondiagonal and fairly large; thus, they are usually considered for the application of covariance modelling.

Almost every spatial correlation models used in data assimilation relies on an underlying uniform and isotropic grid representation, taking either an infinite or periodic domain into account. Gaspari and Cohn (1999) described the development of such a function that was used later in various studies of the same type. The detail of the construction of the spatial correlation functions used in this thesis is provided in Section 5.4.3. Here, we briefly overview those correlation functions that are applied in this research.

For every pair of model grid points, a chordal distance  $D_{ij}$  is defined that connects the position of those two in  $\mathbb{R}^3$  (surface of a sphere). We used three correlation functions, including Gaussian, First-Order-Auto-Regressive (FOAR), and Second-Order-Auto-Regressive (SOAR), for covariance modelling in this reseach. A general form of those are

$$C_G(i, j) = \exp\left(-\frac{D_{ij}^2}{2L_c^2}\right), \quad (3.14)$$

$$C_{FOAR}(i, j) = \exp\left(-\frac{D_{ij}}{L_c}\right), \quad (3.15)$$

and

$$C_{SOAR}(i, j) = \left(1 + \frac{D_{ij}}{L_c}\right) \exp\left(-\frac{D_{ij}}{L_c}\right), \quad (3.16)$$

where  $L_c$  denotes a correlation length scale (in both horizontal  $C_h$  and vertical  $C_v$  correlation functions). After constructing  $C$  on the surface of a sphere, these correlations are projected back to the model space and used to generate the error covariances. The estimation of background or forecast error covariances is thus reduced to estimating the horizontal and vertical correlation length scales as two individual parameters.

### 3.4 Covariance Parameter Estimation

Once the covariance models, particularly for  $\mathbf{R}$  and  $\mathbf{B}$  ( $\mathbf{B}$  can be treated the same as forecast error covariance,  $\mathbf{P}^f$ ), are defined, the underlying parameters of the covariance model need to be properly determined. This typically involved the estimation of background correlation length scales ( $L_h, L_v$ ) and/or background covariance parameter ( $f^b$ ) in  $\mathbf{B}$  (or similarly in  $\mathbf{P}^f$ ), and observation covariance parameter ( $f^o$ ) in  $\mathbf{R}$ . In general, any covariance parameter can be taken into a vector, called a vector of parameters  $\boldsymbol{\alpha} = \{L_h, L_v, f^o, \dots\}$  to be estimated, which determines the optimality of the analysis. Several methods are devised to estimate these parameters, most of which rely on innovation (observation – background) statistics. Innovation,  $\mathbf{d}$ , contains the information from both observations and background (or model forecast), which has a form

$$\mathbf{d} = \mathbf{y}^o - H\mathbf{x}^b, \quad (3.17)$$

where  $\mathbf{x}_t^b$  is a vector of background (similar to a short forecast  $\mathbf{x}_t^f$ ). Estimation methods of covariance parameters based on innovation include  $\chi^2$  diagnostic (Menard and Changs 2000; Menard et al. 2000), the Hollingsworth–Lönnerberg method (Hollingsworth and Lonnberg 1986), maximum likelihood (Dee et al. 1999), Desroziers diagnostic (Desroziers et al. 2005), and cross-validation (Menard and Deshaies-Jacques 2018a). Descriptions and formulations of these methods with a focus on the cross-validation technique as it is performed in this research are provided in Appendix A.

## Chapter 4: Data Assimilation and Inverse Modelling Methods

In chemical data assimilation, the variable to be estimated ( $x$ )<sup>3</sup>, also called the state variable, is the model state or concentrations ( $c$ ), whereas in atmospheric inverse modelling,  $x$  is considered as the model parameter, usually emissions ( $e$ ) (Brasseur and Jacob 2017). Both of these systems are established with a complete and clear understanding of the uncertainties associated with the model (including model inputs) and observations.

Bayes' theorem is a standard foundation in data assimilation and inverse modelling approaches, where the state variable of interest (model state or model parameter) is inferred from observations (Rodgers 2000). It has a form

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}, \quad (4.1)$$

where  $p(x)$  and  $p(y)$  indicate the probability density function (*pdf*) of the state variable  $x$  and observation  $y$ , respectively.  $p(y|x)$  is a conditional *pdf* of the observations given the true value of  $x$ . From the estimation point of view, a prior *pdf* of the state from the model simulation,  $p(x)$  is constructed initially, then by combining it with the observations information and transforming back to the model space (given  $p(y)$  and  $p(y|x)$ ), posterior (analysis) estimation of the state variable  $p(x|y)$  is obtained. The posterior estimation, which accounts for the observations, represents the improved estimation of the state variable. Note that the expressions “a-priori” and “a-posteriori” are equivalently used to refer to “prior, and “posterior” estimations in this study.

---

<sup>3</sup> Alternatively referred to as control variable

Let  $\mathbf{x} \in \mathbb{R}^n$  be a state vector in a state space of size  $n$  (e.g., model grid cells) and  $\mathbf{x}_A$  be a prior (or mean) estimate of  $\mathbf{x}$ , such that

$$\mathbf{x} = \mathbf{x}_A + \boldsymbol{\varepsilon}^b \quad (4.2)$$

where  $\boldsymbol{\varepsilon}^b$  is the prior estimate error with zero mean and covariance  $\mathbf{B} \in \mathbb{R}^{n \times n}$  attributed to  $\mathbf{x}_A$ . Assuming a Gaussian form of error ( $\boldsymbol{\varepsilon}^b \sim N(0, \mathbf{B})$ ), we can obtain  $\mathbf{x} \sim N(\mathbf{x}_A, \mathbf{B})$  by taking the expectation of  $\mathbf{x}$  ( $E[\mathbf{x}] = \mathbf{x}_A$ ) and its variance ( $E[(\mathbf{x} - \mathbf{x}_A)^2] = E[\boldsymbol{\varepsilon}^b (\boldsymbol{\varepsilon}^b)^T] = \mathbf{B}$ ) using Equation (4.2). Given these, the prior *pdf*,  $p(\mathbf{x})$ , has the form

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n \det(\mathbf{B})}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{x}_A)^T \mathbf{B}^{-1}(\mathbf{x} - \mathbf{x}_A)\right). \quad (4.3)$$

Now, considering that the observations  $\mathbf{y}^\circ$  be a vector of size  $m$  in observation space ( $\mathbf{y}^\circ \in \mathbb{R}^m$ ), which relates to the model state by an observation operator  $\mathbf{H}$ , we have the expression

$$\mathbf{y}^\circ = \mathbf{H}\mathbf{x} + \boldsymbol{\varepsilon}^\circ, \quad (4.4)$$

where  $\boldsymbol{\varepsilon}^\circ$  is the error term of observations, also assumed Gaussian ( $\boldsymbol{\varepsilon}^\circ \sim N(0, \mathbf{R})$ ) with zero mean and covariance  $\mathbf{R} \in \mathbb{R}^{m \times m}$ . According to the observation equation (Equation (4.4)), if  $\mathbf{x}$  is given, we can obtain  $p(\mathbf{y}^\circ | \mathbf{x}) = p(\boldsymbol{\varepsilon}^\circ)$  in the form

$$p(\mathbf{y}^\circ | \mathbf{x}) = \frac{1}{\sqrt{(2\pi)^m \det(\mathbf{R})}} \exp\left(-\frac{1}{2}(\mathbf{y}^\circ - \mathbf{H}\mathbf{x})^T \mathbf{R}^{-1}(\mathbf{y}^\circ - \mathbf{H}\mathbf{x})\right) \quad (4.5)$$

where  $p(\mathbf{y}^\circ | \mathbf{x})$  is the *pdf* of the observation data, also called likelihood, and  $E[\boldsymbol{\varepsilon}^\circ (\boldsymbol{\varepsilon}^\circ)^T] = \mathbf{R}$ .

Now, we aim to find the relation for the marginal likelihood,  $p(\mathbf{y}^\circ)$ . First, we take the expectation of  $\mathbf{y}^\circ$ , and we have

$$E[\mathbf{y}^\circ] = \mathbf{H}E[\mathbf{x}] = \mathbf{H}\mathbf{x}_A \quad (4.6)$$

Then, we take the expectation of the variance (covariance) of  $\mathbf{y}^\circ$ , we have

$$E\left[(\mathbf{y}^\circ - E[\mathbf{y}^\circ])(\mathbf{y}^\circ - E[\mathbf{y}^\circ])^T\right] \quad (4.7)$$

Substituting  $\mathbf{y}^\circ$  and  $E[\mathbf{y}^\circ]$  from Equation (4.4) and (4.6) into  $(\mathbf{y}^\circ - E[\mathbf{y}^\circ])$  provides

$$\mathbf{y}^\circ - E[\mathbf{y}^\circ] = \mathbf{y}^\circ - \mathbf{H}\mathbf{x}_A = \mathbf{H}\mathbf{x} - \mathbf{H}\mathbf{x}_A + \boldsymbol{\varepsilon}^\circ = \mathbf{H}(\mathbf{x} - \mathbf{x}_A) + \boldsymbol{\varepsilon}^\circ. \quad (4.8)$$

Using Equation (4.2), we can replace  $(\mathbf{x} - \mathbf{x}_A)$  into Equation (4.8) and obtain

$$\mathbf{y}^\circ - E[\mathbf{y}^\circ] = \mathbf{H}(\boldsymbol{\varepsilon}^b) + \boldsymbol{\varepsilon}^\circ, \quad (4.9)$$

and then substituting Equation (4.9) into Equation (4.7) provides the expectation of the variance of  $\mathbf{y}^\circ$  as

$$E\left[(\mathbf{y}^\circ - E[\mathbf{y}^\circ])(\mathbf{y}^\circ - E[\mathbf{y}^\circ])^T\right] = E\left[(\mathbf{H}(\boldsymbol{\varepsilon}^b) + \boldsymbol{\varepsilon}^\circ)(\mathbf{H}(\boldsymbol{\varepsilon}^b) + \boldsymbol{\varepsilon}^\circ)^T\right] \quad (4.10)$$

In order to solve Equation (4.10), we account for two estimation properties: (i)  $\boldsymbol{\varepsilon}^b$  and  $\boldsymbol{\varepsilon}^\circ$  are considered uncorrelated. We will not show the proof here, but it can be found elsewhere using different approaches, such as the one based on collocated observations (Daley and Menard 1993; Stoffelen 1998; Tangborn et al. 2002); (ii) for Gaussian distribution, if two variables are uncorrelated, then they will be shown as independent as well. Accordingly, based on these two properties, we can expand the right-hand side of Equation (4.10) as

$$\mathbf{H}E[\boldsymbol{\varepsilon}^b(\boldsymbol{\varepsilon}^b)^T]\mathbf{H}^T + \cancel{\mathbf{H}E[\boldsymbol{\varepsilon}^b(\boldsymbol{\varepsilon}^\circ)^T]} + \cancel{E[\boldsymbol{\varepsilon}^\circ(\boldsymbol{\varepsilon}^b)^T]\mathbf{H}^T} + E[\boldsymbol{\varepsilon}^\circ(\boldsymbol{\varepsilon}^\circ)^T], \quad (4.11)$$

where the second and third terms of Equation (4.11) are each equal to zero because the expectation of multiplying independent  $\boldsymbol{\varepsilon}^b$  by  $\boldsymbol{\varepsilon}^\circ$  (or vice versa) is equal to multiplying the expectation of each error term, which is zero (e.g.,  $E[\boldsymbol{\varepsilon}^\circ] = 0$ ). Finally, using the definition

of the prior and observation error covariance ( $E[\boldsymbol{\varepsilon}^b (\boldsymbol{\varepsilon}^b)^T] = \mathbf{B}$ ,  $E[\boldsymbol{\varepsilon}^\circ (\boldsymbol{\varepsilon}^\circ)^T] = \mathbf{R}$ ) in Equation (4.10), one can write

$$E\left[(\mathbf{y}^\circ - E[\mathbf{y}^\circ])(\mathbf{y}^\circ - E[\mathbf{y}^\circ])^T\right] = \mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R} \quad (4.12)$$

Therefore, the marginal likelihood,  $p(\mathbf{y}^\circ)$ , has the form

$$p(\mathbf{y}^\circ) \sim N(\mathbf{H}\mathbf{x}_A, \mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R}), \quad (4.13)$$

or specifically,

$$p(\mathbf{y}^\circ) = \frac{1}{\sqrt{(2\pi)^m \det(\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R})}} \exp\left(-\frac{1}{2}(\mathbf{y}^\circ - \mathbf{H}\mathbf{x}_A)^T (\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R})^{-1} (\mathbf{y}^\circ - \mathbf{H}\mathbf{x}_A)\right). \quad (4.14)$$

We note that  $\mathbf{y}^\circ - \mathbf{H}\mathbf{x}_A$  is known as innovation, which corresponds to the part of  $\mathbf{y}^\circ$  that is not explained by the  $\mathbf{x}_A$  (i.e., model forecast).

Finally, we use the expression of the Bayes' theorem (Equation (4.1)) to obtain the posterior *pdf*,  $p(\mathbf{x} | \mathbf{y}^\circ)$ . To do this, we substitute Equations (4.3), (4.5), and (4.14) into Equation (4.1), and we have

$$p(\mathbf{x} | \mathbf{y}^\circ) = \frac{\det(\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R})^{1/2}}{(2\pi)^{m/2} \det(\mathbf{B})^{1/2} \det(\mathbf{R})^{1/2}} \exp\left(-\frac{1}{2}J(\mathbf{x})\right) \quad (4.15)$$

where  $J(\mathbf{x})$  is equal to

$$\begin{aligned} J(\mathbf{x}) = & (\mathbf{y}^\circ - \mathbf{H}\mathbf{x})^T (\mathbf{R})^{-1} (\mathbf{y}^\circ - \mathbf{H}\mathbf{x}) \\ & + (\mathbf{x} - \mathbf{x}_A)^T (\mathbf{B})^{-1} (\mathbf{x} - \mathbf{x}_A) \\ & - (\mathbf{y}^\circ - \mathbf{H}\mathbf{x}_A)^T (\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R})^{-1} (\mathbf{y}^\circ - \mathbf{H}\mathbf{x}_A) \end{aligned} \quad (4.16)$$

The optimal state estimate  $\hat{\mathbf{x}}$  can be obtained using two different estimation methods (or estimators), such as the maximum a-posteriori (MAP) estimator or the minimum variance (MV) estimator. The MAP estimation method consists of finding the

value of  $\mathbf{x}$  that maximizes the posterior probability of Equation (4.15). The maximum posterior probability, in fact, corresponds to the minimization of  $J(\mathbf{x})$  as we see in Equation (4.17).

$$\hat{\mathbf{x}}_{\text{MAP}} = \arg \max(p(\mathbf{x} | \mathbf{y}^\circ)) = \arg \min(J(\mathbf{x})). \quad (4.17)$$

This is the estimator that is usually used in the variational method (e.g., three-dimensional variational (3D-Var) and 4D-Var). In practice, we try to find the minimum of  $J(\mathbf{x})$  using a gradient search optimization method.

Another estimation method, known as minimum variance (MV) estimation, minimizes the error variance in the posterior probability space. For example,  $\mathcal{V}(\mathbf{x}^*) = \int_{-\infty}^{+\infty} (\mathbf{x} - \mathbf{x}^*)^T (\mathbf{x} - \mathbf{x}^*) p(\mathbf{x} | \mathbf{y}^\circ) d\mathbf{x}$  defines a variance in the posterior probability space, where  $\mathbf{x}^*$  is the variable tends to minimize  $\mathcal{V}$ . We can make it general by introducing a weighting function  $\mathbf{S}$  on the error variance of this form

$$\mathcal{V}(\mathbf{x}^*) = \int_{-\infty}^{+\infty} (\mathbf{x} - \mathbf{x}^*)^T (\mathbf{S})^{-1} (\mathbf{x} - \mathbf{x}^*) p(\mathbf{x} | \mathbf{y}^\circ) d\mathbf{x}. \quad (4.18)$$

The minimum variance estimate can be obtained by solving  $\left. \frac{\partial \mathcal{V}}{\partial \mathbf{x}^*} \right|_{\mathbf{x}^* = \hat{\mathbf{x}}_{\text{MV}}} = 0$ , where  $\hat{\mathbf{x}}_{\text{MV}}$

is the optimal estimate corresponding to the minimum variance estimator. Apply this condition to Equation (4.18), we obtain

$$\left. \frac{\partial \mathcal{V}}{\partial \mathbf{x}^*} \right|_{\mathbf{x}^* = \hat{\mathbf{x}}_{\text{MV}}} = +2\mathbf{S}^{-1} \int_{-\infty}^{+\infty} (\mathbf{x} - \hat{\mathbf{x}}_{\text{MV}}) p(\mathbf{x} | \mathbf{y}^\circ) d\mathbf{x} = 0, \quad (4.19)$$

and by solving Equation (4.19) for  $\hat{\mathbf{x}}_{\text{MV}}$ , we have

$$\int_{-\infty}^{+\infty} \mathbf{x} p(\mathbf{x} | \mathbf{y}^\circ) d\mathbf{x} = \int_{-\infty}^{+\infty} \hat{\mathbf{x}}_{\text{MV}} p(\mathbf{x} | \mathbf{y}^\circ) d\mathbf{x}, \quad (4.20)$$

which is independent of  $\mathbf{S}$ . The right-hand side of Equation (4.20) simplifies as

$$\int_{-\infty}^{+\infty} \hat{\mathbf{x}}_{\text{MV}} p(\mathbf{x} | \mathbf{y}^\circ) d\mathbf{x} = \hat{\mathbf{x}}_{\text{MV}} \int_{-\infty}^{+\infty} p(\mathbf{x} | \mathbf{y}^\circ) d\mathbf{x} = \hat{\mathbf{x}}_{\text{MV}}, \quad (4.21)$$

Since  $\int_{-\infty}^{+\infty} p(\mathbf{x} | \mathbf{y}^\circ) d\mathbf{x} = 1$ .

Thus, substituting Equation (4.21) into Equation (4.20), the expression for minimum variance estimation is obtained as

$$\hat{\mathbf{x}}_{\text{MV}} = \int_{-\infty}^{+\infty} \mathbf{x} p(\mathbf{x} | \mathbf{y}^\circ) d\mathbf{x} = E[\mathbf{x} | \mathbf{y}^\circ], \quad (4.22)$$

which is equal to the conditional mean ( $E[\mathbf{x} | \mathbf{y}^\circ]$ ). The conditional mean can be calculated by knowing the probability that is usually assumed to be Gaussian. Note that for a linear and Gaussian estimation problem, the solution obtained from MAP and MV methods are the same ( $\hat{\mathbf{x}}_{\text{MV}} = \hat{\mathbf{x}}_{\text{MAP}}$ ).

In this section, we present several atmospheric data assimilation (and inverse modelling) systems that use the MV or the MAP estimator to find optimal estimation. In fact, those approaches solve estimation problems and obtain the posterior state and its error statistics. They mostly arise from two data assimilation categories: Kalman filtering and variational data assimilation. In the inverse modelling context, these two categories are equivalent to analytical and adjoint methods, respectively (Kopacz et al. 2010; Brasseur and Jacob 2017). We describe these two approaches by focusing on the former and its subcategories. In the analytical approach, similar to KF, the gain matrix (or the Jacobian) is constructed explicitly, while in adjoint inversion, a variational analysis is performed without constructing the gain matrix.

#### 4.1 4-Dimensional Variational (4D-Var)

The CMAQ adjoint model (Hakami et al. 2007; Zhao et al. 2020) is defined based on a scalar cost function  $J$ . As demonstrated in Section 3.2.3, the adjoint model computes the gradient of the cost function with respect to the model state or model parameter  $\nabla_{\mathbf{x}}J$ . 4D-Var data assimilation (or inverse modelling) integrates the knowledge of the model processes with observations and their associated uncertainties over a particular time period,  $t_0 \rightarrow t_N$ , (assimilation window). It is a method of optimizing the model state in data assimilation or the model parameter in inverse modelling at the initial time  $t_0$ . Since model trajectory in 4D-Var provides an optimal fit to all observations together, it exerts a smooth correction on the state variables. Note that the model parameters are sometimes assumed time-invariant; thus, the time subscript for the parameter can be dropped in their formulation. The 4D-Var cost function follows Equation (4.23), accounting for all observations in the assimilation window,

$$J(\mathbf{x}) = \gamma \frac{1}{2} (\mathbf{x} - \mathbf{x}_A)^T \mathbf{B}^{-1} (\mathbf{x} - \mathbf{x}_A) + \sum_{t=0}^N \frac{1}{2} (\mathbf{y}_t^\circ - \mathbf{H}\mathbf{c}_t)^T \mathbf{R}^{-1} (\mathbf{y}_t^\circ - \mathbf{H}\mathbf{c}_t). \quad (4.23)$$

- For emissions estimation (in general parameter estimation in an inversion), the state vector  $\mathbf{x}$  denotes emissions (i.e.,  $\mathbf{x} = \mathbf{e}$ ),  $\mathbf{x}_A$  is emissions a-priori representing prior or baseline emissions (i.e.,  $\mathbf{x}_A = \mathbf{e}^b$ ),  $\mathbf{B}$  is equivalent to prior emissions error covariance (i.e.,  $\mathbf{B} = \mathbf{E}$ ), and  $\gamma$  is a tunable parameter.
- For concentrations estimation (in general model state estimation for assimilation), the state vector  $\mathbf{x}$  denotes initial concentrations (i.e.,  $\mathbf{x} = \mathbf{c}_0$ ),  $\mathbf{x}_A$  is the model forecast of initial concentrations (i.e.,  $\mathbf{x}_A = \mathbf{c}_0^f$ ),  $\mathbf{B}$  represents forecast error covariance, and  $\gamma=1$ .

We remark that since  $\mathbf{H}$  is applied to concentrations in Equation (4.23), it only represents the observation operator between concentrations and observations, thus does not integrate the model  $\mathbf{M}$ .  $\gamma$  is a regularization factor that is commonly used in emissions inversion to scale the impact of prior error covariance relative to the observation error covariance. In other words, the deviation from the prior estimates in the cost function is weighted against the deviation from the observations and model together (Hakami et al. 2005). In this research, we attempt to find the optimal value of  $\gamma$  based on the method used by Lu et al. (2021).

The evolution of the model state  $\mathbf{c}$  based on the atmospheric transport model ( $\mathbf{M}$ ) equation and stationary emissions  $\mathbf{e}$  is expressed as

$$\mathbf{c}_t = \mathbf{M}(\mathbf{c}_{t-1}, \mathbf{e}). \quad (4.24)$$

If the model is perfect, it provides strong constrain on the evolution of the state vector, which leads to a strong-constraint 4D-Var (or simply 4D-Var); however, the model can suffer from various types of errors ( $\boldsymbol{\varepsilon}_t$ ), such as numeral discretization error, representation error, and errors due to insufficient modelling of sub-processes. Another form of 4D-Var that allows for the evolution of model error is known as weak-constraint 4D-Var, where the evolution of the model is given by

$$\mathbf{c}_t = \mathbf{M}(\mathbf{c}_{t-1}, \mathbf{e}) + \boldsymbol{\varepsilon}_t. \quad (4.25)$$

Taking a similar definition of variables from the explanations in Equation (4.23), the cost function for the weak constraint 4D-Var is

$$J(\mathbf{x}, \boldsymbol{\varepsilon}_t) = \frac{1}{2}(\mathbf{x} - \mathbf{x}_A)^T \mathbf{B}^{-1}(\mathbf{x} - \mathbf{x}_A) + \sum_{t=0}^N \frac{1}{2} \boldsymbol{\varepsilon}_t^T \mathbf{Q}^{-1} \boldsymbol{\varepsilon}_t + \sum_{t=0}^N \frac{1}{2} (\mathbf{y}_t^\circ - \mathbf{H}\mathbf{c}_t)^T \mathbf{R}^{-1}(\mathbf{y}_t^\circ - \mathbf{H}\mathbf{c}_t), \quad (4.26)$$

where  $\mathbf{Q}$  is model error covariance. In this case, the cost function is minimized with respect to the state variable (either emissions or concentration) and model error simultaneously. Several studies applied this method in their assimilation or inverse modelling scheme and examined its capability (Stanevich et al. 2020; 2021). Despite the fact that a weak constrain 4D-Var aids in identifying the source of model bias over the whole model trajectory, the cost of implementing and simulating such a method is extremely high. Many studies also indicate that the impact of incorrect prior emissions is significantly higher than the model error throughout assimilation/inversion; hence, they do not account for this in their estimation problem. In addition, given the accumulation behaviour of the model errors over time, its impact can be even smaller for a short assimilation window. In this research, a strong 4D-Var method for methane emissions inverse modelling over a regional domain of CMAQ is considered.

The minimization of the cost function in Equation (4.23) requires the gradient  $\nabla_{\mathbf{x}}J$ , which is obtained by making a perturbation in the model parameter or initial state. The adjoint model provides an efficient way to compute the gradient (see Section 3.2.3). Accordingly, the gradients of  $J$  have a form

$$\nabla J(\mathbf{x}) = \gamma \mathbf{B}^{-1}(\mathbf{x} - \mathbf{x}_A) + \sum_{t=0}^N (\mathbf{M}'_t)^T \mathbf{R}^{-1}(\mathbf{y}_t^\circ - \mathbf{H}\mathbf{c}_t), \quad (4.27)$$

where  $\mathbf{M}' = \frac{\partial \mathbf{M}}{\partial \mathbf{x}_t}$  is a tangent linear model (equivalent to the Jacobian matrix in an explicit calculation), and  $(\mathbf{M}'_t)^T$  is the adjoint model. The minimization of the cost function in Equation (4.23) can be accomplished through variational analysis. A mathematical derivative of this method can be found in the literature using an adjoint method or a by applying the method of the Lagrangian multiplier (Sandu et al. 2005; Elbern et al. 2007;

Hakami et al. 2007; Henze et al. 2007). We note that the adjoint model does not necessarily involve in construction of an explicit form of  $(\mathbf{M}'_t)^T$ , instead, it acts as an operator application of  $(\mathbf{M}'_t)^T$  to a vector that initializes it, also known as adjoint forcing.

In practice, an adjoint system starts with a pass of a forward model simulation over the assimilation time window  $t_0 \rightarrow t_N$ , during which it computes the adjoint forcing,  $\mathbf{R}^{-1}(\mathbf{y}_t^\circ - \mathbf{H}\mathbf{c}_{t,A})$ , for all the observations over that period. The adjoint model involves a backward integration of the model for a new variable (i.e., adjoint variable). In the backward integration, we force the adjoint model with the collected forcing and propagate it backward in time. This process repeats for each new forcing along the way of integrating back in time from  $t_N \rightarrow t_0$ . At the final time of the backward integration, the cost function gradient is obtained as  $\nabla J(\mathbf{x}_A)$ . Using a gradient-based optimization algorithm (e.g., Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS), (Byrd et al. 1995)), one can find an updated estimate on  $\mathbf{x}$ . However, minimization of the cost function may require a number of iterations (due to nonlinearity of dynamic system or non-Gaussianity of error assumptions) until convergence.

4D-Var assimilation/inversion is an efficient method to constrain the model state/parameter at the native model resolution. It does not have a computational limitation on the size of the state vector or the observations, and the error pdf does not need to be Gaussian. However, besides the assumption of a perfect model, 4D-Var does not provide an estimate of the posterior (analysis) error covariances as part of the solution. We note that for an assimilation/inversion system that relies on different types of uncertainties in

the model and observations, it is crucial to find an accurate and realistic estimate of analysis error covariance corresponding to the optimal solution.

An alternative (in 4D-Var) for estimating the posterior error covariance is to obtain the Hessian matrix,  $\nabla_{\mathbf{x}}^2 J$  (the second derivative of the cost function). However, full construction of that requires  $N+1$  simulations, which is computationally prohibitive for a large state vector. Nevertheless, different approximation methods have been introduced to fully or partially compute the posterior error covariance, mainly divided into implicit (deterministic) and explicit low-rank estimation (stochastic approaches). Bousserez et al. (2015) and Bousserez and Henze (2018) detailed the theoretical analysis of those methods with a focus on the explicit approach.

## 4.2 Ensemble Kalman Filter (EnKF)

Another category of data assimilation approach exists that does not rely on optimization based on variational analysis. It is derived from linear estimation theory and is widely known as Kalman filtering. The standard Kalman filter (KF) and its variants, including the extended Kalman filter (EKF) and RTS extended Kalman smoother (RTS-EKS), are described in detail in Appendix B, mainly since they provide a limited application for the high-dimensional state space problem of atmospheric models. The ensemble Kalman filter (EnKF) is another extension to the Kalman filter method that is designed to deal with nonlinearity (and non-Gaussianity) in large state space estimation problems; thus, it is practical for the application of chemical data assimilation and atmospheric inversions. The main difference between EnKF and EKF (see Appendix B) is that EnKF, contrary to EKF, does not rely on the linearity of the system. In other words, the model  $M$  and the observation operator  $H$  are used in their default nonlinear form; hence, a linearization is

not required. In order to keep the computational cost of estimating the state and error covariance in EnKF low, a limited number of realizations (or ensembles) of the state are initialized, typically by random sampling. One major characteristic of EnKF with a limited number of ensembles and random perturbations that distinguish it from a typical metaheuristic optimization algorithm (e.g., genetic algorithms) is that EnKF updates the states and error covariances by providing new information (observations) over time; thus, the estimation is built upon the past and current state of the system (and is also time dependent).

In atmospheric data assimilation, the idea of EnKF was introduced by (Evensen 1994) to respond to the two main challenges in the previous form of Kalman filtering in the high dimensional system. (i) high computational cost of constructing and propagating the error covariance matrix using the TLM, and (ii) approximation of the nonlinearity of a system by TLM, which may lead to underestimated error covariance estimate (i.e., filter divergence) depending on the degree of nonlinearity of the system and time discretization. The EnKF algorithm was amended later (Evensen 2009a; Houtekamer and Mitchell 1998; 2005), and several subcategories have been developed to improve the performance of this approach and to address its issues. A detailed review of these methods is provided in Houtekamer and Zhang (2016). We present a brief illustration of the EnKF algorithm below.

The EnKF method generally approximates the state and error covariance of a high-dimensional system with a limited  $N$  number of random samples or ensembles, where each realization requires an individual model integration. In this case, the ensemble mean and its pdf represent the model state and its uncertainty. Given  $\mathbf{x}(i)$  the model state for the  $i$ th

ensemble member, the mean state of the system  $\bar{\mathbf{x}}_t$  and the error covariances  $\mathbf{P}_t^f$  are computed from ensembles. For  $i=1 \dots N$  and timestep  $[t, t+1]$ , the two steps of the algorithm are as follows

- Analysis step

$$\mathbf{x}_t^a(i) = \mathbf{x}_t^f(i) + \mathbf{K}_t (\mathbf{y}_t^\circ - H_t \mathbf{x}_t^f(i)) \quad (4.28)$$

$$\left. \begin{aligned} \mathbf{P}_t^f H_t^T &= \sum_{i=1}^N \left( \mathbf{x}_t^f(i) - \bar{\mathbf{x}}_t^f \right) \left( \mathbf{y}_t^\circ - \overline{H_t \mathbf{x}_t^f} \right)^T \\ H_t \mathbf{P}_t^f H_t^T &= \sum_{i=1}^N \left( \mathbf{y}_t^\circ - \overline{H_t \mathbf{x}_t^f} \right) \left( \mathbf{y}_t^\circ - \overline{H_t \mathbf{x}_t^f} \right)^T \end{aligned} \right\} \Rightarrow \mathbf{K}_t = \mathbf{P}_t^f H_t^T \left( H_t \mathbf{P}_t^f H_t^T + \mathbf{R} \right)^{-1} \quad (4.29)$$

- Forecast step

$$\mathbf{x}_{t+1}^f(i) = M_{t+1} \mathbf{x}_t^a(i) + \varepsilon \quad (4.30)$$

$$\begin{aligned} \bar{\mathbf{x}}_t^f &= \sum_{i=1}^N \mathbf{x}_t^f(i), \quad \overline{H_t \mathbf{x}_t^f} = \sum_{i=1}^N H_t \mathbf{x}_t^f(i), \\ \mathbf{P}_t^f &= \sum_{i=1}^N \left( \mathbf{x}_t^f(i) - \bar{\mathbf{x}}_t^f \right) \left( \mathbf{x}_t^f(i) - \bar{\mathbf{x}}_t^f \right)^T \end{aligned} \quad (4.31)$$

where  $M$  and  $H$  represent a nonlinear model and observation operator, and  $\mathbf{K}$  denotes the Kalman gain matrix. Equivalent to the forecast equations, the mean value of the analysis,  $\bar{\mathbf{x}}_t^a$ , and the analysis error covariance  $\mathbf{P}_t^a$  can be obtained as part of the solution.

The Ensemble Kalman filter method has been used extensively in data assimilation (and inverse modelling), mainly due to its lower computational expenses than those that require the adjoint model or TLM. However, in order to obtain a good approximation of the error statistics of the system, adequate realizations (ensemble members) are required depending on the size, scale, and application of the problem. Theoretically, increasing the number of ensembles more effectively represents the underlying pdf, yet in practice, a few dozen ensembles are usually appropriate, depending on the system's complexity. For

example, in atmospheric chemistry, Zubrow et al. (2008) used 20 ensembles for their CO assimilation with CMAQ; Miyazaki et al. (2012) found that they could achieve an optimal estimation with 32 ensemble members, and Peng et al. (2015) used 48 ensembles for CO<sub>2</sub> flux inverse modelling with CMAQ.

Besides the computational cost of generating ensembles, EnKF faces the challenge of sampling error due to a low-rank approximation (i.e., a limited number of ensembles that represent the full error covariance of a high dimensional system). The low-rank approximation is often accompanied by a spurious long-range error correlation (Evensen 2009b), which can strongly affect the performance of EnKF assimilation/inversion (e.g., result in filter divergence). Yet, some techniques aid in dealing with that problem, particularly localization (Bocquet 2016; Houtekamer and Mitchell 2005) and filtering of variance or correlation length scale (Raynaud et al. 2009; Berry and Sauer 2013). Another challenge with EnKF is that the limited number of ensembles can result in an underestimation of ensemble spread, as a form of error variance loss (Menard et al. 2021), which again negatively impacts the performance of the assimilation. Inflation of error variance is a technique commonly used to alleviate this kind of problem (Miyoshi 2011). Although the localization and inflation techniques make EnKF perform more stable than before, there is not a well-adapted adjustment for all problems; in other words, it requires a careful assessment of the assimilation system behaviour, depending on the domain and application. All these considered, another type of Kalman filtering assimilation exists that avoids facing the shortcomings of EnKF. This category relies on the parametrization of the covariance matrices, also known as covariance modelling, which attempts to replicate the form and evolution of error covariances explicitly in an assimilation/inversion system.

Before describing this new assimilation system upon which our thesis analysis is built, we need to identify how a covariance model performs in an assimilation framework, and particularly, how appropriate parameters corresponding to the optimal solution are determined in covariance models. Hence, it is recommended to review the materials in Section 3.3 about covariance modelling and in Section 3.4 and Appendix A about covariance parameter estimation.

### **4.3 Parametric Variance Kalman Filter (PvKF)**

The parametric variance Kalman filter is a simplified version of the Parametric Kalman Filter (PKF). Instead of evolving error covariance matrices as in Kalman Filter, the PKF evolves the characteristic parameters of the error covariance. An error covariance function is often characterized by its variance and correlation length scales. The PKF offers evolution equations to the error variance and correlation length scales (in time and space). With the parametric variance Kalman filter (PvKF), only the error variance is evolved (in time and space), and the error correlation is assumed to be homogeneous, isotropic, and stationary.

PvKF also meets the fundamental assumptions generally used in Kalman filtering, which are briefly reviewed as follows. First, all errors in our state space, including the model and observations errors, are assumed to be white noise, meaning that they are sampled from a Gaussian distribution with zero mean. Second, observation errors are typically assumed to be uncorrelated to each other in space and time. Third, observation errors are also uncorrelated with model (forecast) errors, mainly because the model operator ( $M$ ) and the observation operator ( $H$ ) are independent. Fourth, methane chemistry

is assumed to be linear (see Section 2.1.2), so that PvKF as a linear estimation system is expected to reasonably perform for methane assimilation or inversion.

This thesis in Chapters 5 and 6 presents the development and results of the PvKF assimilation system using GOSAT methane and the hemispheric CMAQ model (Voshtani et al. 2022a, 2022b). It is shown that the PvKF algorithm, besides its simplicity, is well-adapted for the assimilation of long-lived species without loss of variance. In fact, the total variance field is preserved for a certain revisit time of the satellite retrieval (e.g., 3 days). Since PvKF assimilation is proposed in this thesis, more details prior to its development with hemispheric CMAQ and GOSAT methane (presented in Chapter 5), including the background information and formulations, are needed to be reviewed. First, some important theoretical background of the PvKF, particularly the derivation of the continuum representation of error propagations, are provided in Sections 4.3.1 to Section 4.3.3. This includes an example of simple dynamics, followed by two forms of solutions: (i) solutions by operator splitting and (ii) solutions by method of characteristics. Second, the algorithm of PvKF assimilation are presented in Section 4.3.4.

### **4.3.1 Continuum Representation of Covariances: A Simple Dynamical Model**

The PKF approach features the continuum representation of covariance matrices in time and space. In a 1D problem, for example, covariances have a functional form (operators) of a pair of spatial coordinates ( $\mathbf{x}_1, \mathbf{x}_2$ ) and time  $t$ . The time evolution of such a covariance function can be divided into two separate functions, each representing the dynamics with respect to one spatial coordinate. Using an operator splitting technique (Strang 1968), one can integrate the two operators with a dynamical model (e.g., advection-diffusion transport). Nevertheless, for a pure advection system, a solution to the original

problem can also be obtained without operator splitting. This method, known as the method of characteristics (Cohn 1993), is applied in PKF assimilation to propagate the error variance using a purely advective dynamical scheme. In fact, it is shown that for a system with no model error, the error variance obeys the advection equation; thus, it can be propagated without knowledge of error correlations. As this approach frames the core of the PKF algorithm (and its variants), including the assimilation scheme applied in this thesis, we describe it as follows.

First, a simple form of a dynamical model (i.e., 1D linear advection model) is shown, then solutions to that model by both the operator splitting method and the method of characteristics are expressed.

Let us define a 1D linear advection equation in a form

$$\frac{\partial c}{\partial t} + u(x, t) \frac{\partial c}{\partial x} = 0, \quad (4.32)$$

where  $c$  represents the concentration of a tracer and  $u(x, t)$  denotes the 1D wind field. Given that the error only originates from concentrations (not wind), it follows the same differential equation as

$$\frac{\partial \sigma}{\partial t} + u(x, t) \frac{\partial \sigma}{\partial x} = 0, \quad (4.33)$$

where  $\sigma$  is the error of  $c$  with respect to the truth (i.e.,  $\sigma = c - c^t$ ). For a continuous representation of  $\sigma(x, t)$ , the error covariance function  $P(x_1, x_2, t)$  at time  $t$ , is the covariance of  $\sigma(x_1, t)$  with  $\sigma(x_2, t)$  at time  $t$ , (i.e.,  $P(x_1, x_2, t) = E[\sigma(x_1, t)\sigma(x_2, t)]$ ), where  $E$  denotes the expectation operator, and  $x_1$  and  $x_2$ , represent two spatial coordinates. Note that given the dimension of the spatial space  $N$ , the covariance  $P$  has  $N \times N$  dimension; hence, for the 1D advection problem, the covariance evolution equation will be in 2D. Following (Cohn

1993), we specified an extended spatial space  $(x_1, x_2)$ , where  $\sigma_1 = \sigma(x_1, t)$  and  $\sigma_2 = \sigma(x_2, t)$  both follow the advection formulation in Equation (4.33). Multiplying the evolution of the error equation of  $\sigma_1$  (or  $\sigma_2$ ) in Equation (4.33) with the error  $\sigma_2$  (or  $\sigma_1$ ) yields

$$\begin{aligned}\sigma_2 \frac{\partial \sigma_1}{\partial t} + u(x_1, t) \frac{\partial (\sigma_1 \sigma_2)}{\partial x_1} &= 0, \\ \sigma_1 \frac{\partial \sigma_2}{\partial t} + u(x_2, t) \frac{\partial (\sigma_2 \sigma_1)}{\partial x_2} &= 0,\end{aligned}\tag{4.34}$$

where  $u(x_1, t)$  and  $u(x_2, t)$  are the same wind field, but applied to different coordinates. Adding the two equations (in Equation (4.34)), and taking the expectation of each term, we obtain

$$\frac{\partial P}{\partial t} + L_1(P) + L_2(P) = 0,\tag{4.35}$$

where  $L_1$  and  $L_2$  are linear operators of the form

$$L_k = u(x_k, t) \frac{\partial}{\partial x_k} \quad (k=1,2).\tag{4.36}$$

Further details and the generalization of this approach to higher dimensions are detailed by Cohn (1993) and later discussed in Ménard (2000) and Menard et al. (2021). The solution to Equation (4.35) can be demonstrated using two approaches: (i) operator splitting and (ii) method of characteristics as follows.

### 4.3.2 Solutions by Operator Splitting

An important property of linear operators  $L_1$  and  $L_2$  is that they can commute over an arbitrary function  $f(x_1, x_2, t)$ , such that  $L_1 L_2 f = L_2 L_1 f$ . Accordingly, the solution of Equation (4.35) can be obtained using the operator splitting technique (Strang 1968). To demonstrate this, let  $\psi(x_1, x_2, t_f; 0)$  be the solution to Equation (4.35) from  $[0, t_f]$ ; we have

$$P(x_1, x_2, t_f) = \psi(x_1, x_2, t_f; 0)P(x_1, x_2, 0). \quad (4.37)$$

One can show that the solution  $\psi(x_1, x_2, t_f; 0)$  can be decomposed and expressed in a form

$$\psi(x_1, x_2, t_f; 0) = \psi_1(x_1, x_2, t_f; 0)\psi_2(x_1, x_2, t_f; 0), \quad (4.38)$$

in a way that  $\psi_1(x_1, x_2, t_f; 0)$  and  $\psi_2(x_1, x_2, t_f; 0)$  become the solutions to equations

$$\frac{\partial P}{\partial t} + L_1(P) = 0, \quad (4.39)$$

$$\frac{\partial P}{\partial t} + L_2(P) = 0. \quad (4.40)$$

Thus, with operator splitting, the solution to Equation (4.39) is expressed as

$$P^*(x_1, x_2, t_f; 0) = \psi_1(x_1, x_2, t_f; 0)P(x_1, x_2, 0), \quad (4.41)$$

while it is used as an initial condition for solving Equation (4.40), yielding

$$P(x_1, x_2, t_f; 0) = \psi_2(x_1, x_2, t_f; 0)P^*(x_1, x_2, 0). \quad (4.42)$$

Finally, the solution of the error covariance propagation equation (Equation (4.35)) can be obtained by combining Equations (4.39) and (4.40). In fact, the solution is obtained by propagating the error covariance in  $x_1$  followed by the propagation in  $x_2$ . Due to the commutativity of operators, the order can be reversed, and furthermore, the solution to Equation (4.35) is exact.

Operator splitting can be applied similarly to the discrete formulation. Considering the vector of initial error (error standard deviation),  $\boldsymbol{\sigma}(0)$ , the propagator (i.e., model matrix) of the error  $\mathbf{M}_{0,t}$  leads to computing the error at time  $t$ ,  $\boldsymbol{\sigma}(t)$ . Applying the model matrix on the initial error covariance matrix,  $\mathbf{P}(0) = E[\boldsymbol{\sigma}(0)\boldsymbol{\sigma}^T(0)]$ , a discretized form of the solution in Equation (4.41) and Equation (4.42) from time  $0$  to  $t$  are obtained as

$$\mathbf{P}^*(t) = \mathbf{M}\mathbf{P}(0), \quad (4.43)$$

and

$$\mathbf{P}(t) = \mathbf{M}[\mathbf{P}^*(t)]^T. \quad (4.44)$$

The transpose in Equation (4.44) emphasizes the action of  $\mathbf{M}$  on the second spatial coordinate in the covariance matrix (rows of  $\mathbf{P}^*(t)$ , instead of columns). Thus, combining Equations (4.43) and (4.44), the discrete solution of Equation (4.35) at time  $t$  is expressed as

$$\mathbf{P}(t) = \mathbf{M}[\mathbf{P}^*(t)]^T = \mathbf{M}[\mathbf{M}\mathbf{P}(0)]^T = \mathbf{M}\mathbf{P}(0)\mathbf{M}^T. \quad (4.45)$$

Equation (4.45), in fact, represents the error covariance propagation of the standard Kalman filter for a discrete space and model.

### 4.3.3 Solutions by Method of Characteristics

Let us consider that a particle  $\xi$  in the air starts moving along the trajectory  $\Phi_t$ , and its position  $x$  at time  $t$  can be expressed as

$$x(t, \xi) = \Phi_t(\xi). \quad (4.46)$$

By solving the characteristics equation (Equation (4.47)), given the wind field  $u(x, t)$ ,

$$\frac{dx}{dt} = u(x, t), \quad (4.47)$$

one can obtain the trajectory for a specified time window  $[0, t]$ . Equation (4.35) is, in fact, a hyperbolic equation for a pair of particles  $(\xi_1, \xi_2)$  with trajectories  $\Phi_t(\xi_1)$  and  $\Phi_t(\xi_2)$ , and characteristics  $x_1 = x(t, \xi_1)$  and  $x_2 = x(t, \xi_2)$  that are solutions of

$$\frac{dx_1}{dt} = u(x_1, t), \quad \frac{dx_2}{dt} = u(x_2, t). \quad (4.48)$$

Hence, the covariance function between these pairs of particles (i.e.,  $P(x_1, x_2, t) = P(t, \xi_1, \xi_2)$ ) satisfies the conservation equation:

$$\frac{dP}{dt} = 0, \quad (4.49)$$

which indicates that the covariance function between any arbitrary pair of particles along the trajectory is conserved. Note that all trajectories use the same wind field  $u$ , yet can apply to different positions.

One can show that the conservation property in Equation (4.49) is always valid for the error variances. To show this, we assume that the position of two particles coincides initially, such that  $x_1(0) = x_2(0)$ ; then, according to Equation (4.49), they remain coincident after any period of time  $t$  (i.e.,  $x_1(t) = x_2(t)$ ). Knowing that the covariance evaluated at the same coordinates represents the variance (i.e.,  $P(x, x, t) = V(x, t)$ ), the variance is conserved and follows the advection equation,

$$\frac{dV}{dt} = \frac{\partial V}{\partial t} + u(x, t) \frac{\partial V}{\partial x} = 0. \quad (4.50)$$

We note that all these properties are derived in Cohn (1993), where further details and generalizations to a higher dimension are demonstrated.

#### 4.3.4 PvKF Algorithm

PvKF maintains a similar strategy used previously in sequential filtering as an efficient technique for high dimensional data assimilation systems, particularly for estimating long-lived chemical species such as stratospheric ozone. The error correlation in this method is computed at every assimilation cycle on the fly, as needed, using the correlation functions between a pair of points without the need to store a large matrix in

state-space. A 3D covariance function between two model state gridpoints ( $\mathbf{x}(x, y, z), \mathbf{x}'(x, y, z)$ ) thus has a form

$$P(\mathbf{x}, \mathbf{x}', t) = \sigma(\mathbf{x}, t) C(\mathbf{x}, \mathbf{x}', t) \sigma(\mathbf{x}', t) \quad (4.51)$$

where  $\sigma$  is the standard deviation at a point and time  $(\mathbf{x}, t)$ , and  $C(\mathbf{x}, \mathbf{x}', t)$  denotes a correlation function with a common form (e.g., Gaussian, FOAR, SOAR, diffusion model) with a determinable correlation length scale.

A summary of how analysis and forecast assimilation steps are configured within PvKF for timestep  $[t, t+1]$  are as follows (Appendix C2 details the algorithm).

- Analysis step

$$\mathbf{x}_t^a = \mathbf{x}_t^f + \mathbf{K}_t (\mathbf{y}_t^\circ - H_t \mathbf{x}_t^f) \quad (4.52)$$

$$\mathbf{P}_t^a = \mathbf{P}_t^f - \mathbf{K}_t \mathbf{H}_t \mathbf{P}_t^f = \mathbf{P}_t^f - (\mathbf{H}_t \mathbf{P}_t^f)^T (\mathbf{H}_t \mathbf{P}_t^f \mathbf{H}_t^T + \mathbf{R})^{-1} (\mathbf{H}_t \mathbf{P}_t^f) \quad (4.53)$$

$$\left. \begin{array}{l} H_{\mathbf{x}}[P_t^f(\mathbf{x}, \mathbf{x}', t)] \Leftrightarrow \mathbf{H} \mathbf{P}^f \\ H_{\mathbf{x}'}[P_t^f(\mathbf{x}, \mathbf{x}')] \Leftrightarrow \mathbf{P}_t^f \mathbf{H}_t^T \\ H_{\mathbf{x}}[H_{\mathbf{x}'}(P_t^f(\mathbf{x}, \mathbf{x}'))] \Leftrightarrow \mathbf{H}_t \mathbf{P}_t^f \mathbf{H}_t^T \end{array} \right\} \Rightarrow \mathbf{K}_t = (\mathbf{H}_t \mathbf{P}_t^f)^T (\mathbf{H}_t \mathbf{P}_t^f \mathbf{H}_t^T + \mathbf{R})^{-1} \quad (4.54)$$

- Forecast step

$$\mathbf{x}_{t+1}^f(i) = M_{t+1} \mathbf{x}_t^a(i) \quad (4.55)$$

$$\boldsymbol{\sigma}_{t+1}^f(i) = M_{t+1}^* \boldsymbol{\sigma}_t^a(i) \quad (4.56)$$

where  $M$  and  $M^*$  are the forecast and advection scheme of the native forecast model, respectively, and  $\boldsymbol{\sigma}$  is a vector of error variance  $\sigma$  at every point. The forecast error variance vector is accompanied by a similar correlation model as used in Equation (4.51), resulting in an updated error covariance after the forecast.

#### 4.4 General Form of Parametric Kalman Filter (PKF)

A general form of Parametric Kalman Filter (PKF) is developed by Pannekoucke et al. (2016), where the evolution of error covariance, including the error variance and correlations, is characterized by a certain number of parameters. In other words, the full description and evolution of a large error covariance matrix are approximated with a local characteristic of variance and the local shape of the correlation function. Accordingly, a diffusion covariance model, where the covariance matrix  $\mathbf{P}$  is replaced by a diffusion covariance model  $\mathbf{P}_{diff.}$  is developed, such that

$$\mathbf{P} = \mathbf{P}_{diff.} = \Sigma \mathbf{L} \mathbf{L}^T \Sigma^T, \quad (4.57)$$

where  $\Sigma$  is a diagonal matrix of error standard deviations, and  $\mathbf{L}$  denotes the propagator of the diffusion equation with a function form

$$\mathbf{L} = e^{\frac{1}{2} \nabla \cdot (\nu \nabla)}. \quad (4.58)$$

In Equation (4.58),  $\nu$  is a local diffusion tensor and  $V = \Sigma^2$  represents the variance field. The diffusion covariance model allows for constructing heterogeneous correlation functions, which have a particular property in designing a local diffusion tensor. The dynamic of the full covariance matrix is then reduced to the dynamics of  $\nu$  and  $V$ . Pannekoucke et al. (2016, 2018a) showed how the forecast and analysis steps of assimilation are devised based on these two parameters using the application of a simple advection-diffusion scheme and nonlinear diffusive of Burgers equation. Compared with a full Kalman filter, it was shown that the parametric formulation is able to reproduce the main features in the evolution of real variance and correlation length scales (even using an anisotropic formulation of the covariance function) (Pannekoucke et al. 2018a; Pannekoucke et al. 2021). The application of this method has been examined for 1D/2D

assimilation problems, yet has not been implemented in realistic assimilation of a particular observation network. Despite the approximation of the dynamic of error covariances, which is also taken into account explicitly/implicitly in EnKF/4D-Var, the PKF method can dramatically reduce the computational cost of the covariance forecast. It neither needs ensembles nor the adjoint model, while allowing for the inclusion of the modelling error growth (e.g., transport error) within the assimilation system (Voshtani et al. 2022a), which is not the case in standard 4D-Var.

## **Chapter 5: Assimilation of GOSAT Methane in the Hemispheric**

### **CMAQ; Part I: Design of the Assimilation System**

In this Chapter 5, a parametric Kalman filter data assimilation system using GOSAT methane observations within the hemispheric CMAQ model is demonstrated. The assimilation system produces forecasts and analyses of concentrations and explicitly computes its evolving error variance while remaining computationally competitive with other data assimilation schemes such as 4-dimensional variational (4D-Var) and ensemble Kalman filter (EnKF). The error variance in this system is advected using the native advection scheme of the CMAQ model and updated at each analysis while the error correlations are kept fixed. We discuss extensions to the CMAQ model to include methane transport and emissions (both anthropogenic and natural) and perform a bias correction for the GOSAT observations. The results using synthetic observations show that the analysis error and analysis increments follow the advective flow while conserving the information content (i.e., total variance). We also demonstrate that the vertical error correlation contributes to the inference of variables down to the surface. In the next chapter, we use this assimilation system to obtain optimal assimilation of GOSAT observations.

#### **5.1 Introduction**

Methane is the second most important greenhouse gas (GHG) after CO<sub>2</sub>, contributing to about 16% of the anthropogenic radiative forcing of all types of GHGs (Myhre et al. 2013; Etminan et al. 2016; Saunio et al. 2020). The globally averaged methane concentration has risen sharply since 2007, while the largest annual increase of methane ( $15.85 \pm 0.47$  ppb) was recorded in 2020 (Butler and Montzka 2020). Owing to a stronger warming potential and a significantly shorter lifetime compared to CO<sub>2</sub>, the

reduction of methane has drawn much attention in GHG mitigation policy (Dlugokencky et al. 2011; National Academies of Sciences 2018; Fletcher and Schaefer 2019; Nisbet et al. 2020). Furthermore, a higher methane concentration increases tropospheric ozone production, which is a critical problem, especially in populated areas in the Northern Hemisphere (Fiore et al. 2002; Forster et al. 2007b). Methane is also identified as a major factor in the production of stratospheric water vapor, which indirectly affects the oxidation of other pollutants in the atmosphere (Brasseur and Jacob 2017).

Because of its impact on atmospheric chemistry and climate, better characterization of methane concentrations, and subsequently, inverse modelling of its emissions, has received significant attention over the past decade (Jacob et al. 2016; Turner et al. 2019; Wang et al. 2019; Kuze et al. 2020; Saunio et al. 2020; Lu et al. 2021; Qu et al. 2021). While most past studies have focused on the global inversion of methane sources at low spatial resolution, inversions over regional domains and at higher resolutions are important for constraining anthropogenic sources, such as fugitive and high-emitting sources from industry (e.g., oil and gas fracking, natural gas production) (Zavala-Araiza et al. 2015; Johnson et al. 2017; Zavala-Araiza et al. 2017; Fox et al. 2019). Inverse modelling of methane on a highly resolved regional domain requires an accurate estimate of both the initial concentrations within the domain and the inflows of methane at the lateral boundaries of the regional model and their uncertainties. In other words, the contribution from emissions on a short time scale (e.g., a week to a month) of a regional domain can be significantly smaller than the inflow mass from the lateral boundary (Wecht et al. 2014). From an estimation point of view, the emissions signal (in inverse modelling) is masked by a much larger contribution from the state. Therefore, it is important to reduce the state

uncertainty to a level comparable to the signal that we want to extract in inverse modelling. The scheme we have developed and verified by cross-validation in Chapters 5 and 6 permits us to estimate the error variance of the state, which hopefully will be reduced enough by assimilation to extract the signal of the emissions.

Estimating the concentrations (i.e., in our case, the model state), whether for the purpose of constraining the initial conditions or boundary conditions or the whole domain in space and time, is a typical problem of data assimilation. We may contrast it with an inverse modelling problem where the variable to be estimated is not the model state, but rather a model parameter, such as the emissions. Furthermore, we should remark that the forward model in the data assimilation usually relates the estimated variable (i.e., a model concentration) with the observed variable (e.g., a column measurement). Hence, the forward model in the data assimilation context depends only on the measurement or instrument characteristics, such as the vertical averaging kernel or the radiative transfer problem when the radiance measurement is the observation (i.e., Level 1B data acquired by satellites). In inverse modelling, the forward model relates, instead, the estimated parameter (i.e., emissions) with observations, and thus has to include the transport model (i.e., dynamic model of the atmosphere). See the tutorials in Kasibhatla et al. (2000) for the commonalities and differences between inverse modelling and data assimilation.

Data assimilation has issues of its own, and several methods have been developed, such as the ensemble Kalman filter (EnKF), 4-dimensional variational (4D-Var), and hybrid ensemble-variational methods (see Asch et al. 2016) for a detailed review). These methods require several model integrations (on the order of 30 to 100) for the state estimation. Ensemble methods are well-adapted for nonlinear atmospheric models,

whereas variational methods are well-fitted for nonlinear observation operators (i.e., forward model in data assimilation context). However, regardless of nonlinearity, all data assimilation methods face the challenge of their applicability to large state-space models (e.g., atmospheric models state-space). For 3D- and 4D-Var, the applicability of the assimilation algorithm to large state-space is made possible with the introduction of the adjoint of operators combined with the use of an initial or background error correlation, assumed to be homogeneous and isotropic. In this case, the effect of multiplication of an (homogeneous and isotropic) error covariance with a state vector can be obtained as a sequence of operators without involving the storage of extremely large covariance matrices (Gauthier et al. 1999; Gauthier et al. 2007; Errera et al. 2008; Errera and Menard 2012) (see Massart et al. (2014) for 4D-Var assimilation of methane). For EnKF methods, applicability to large state-space arises from the use of a limited number of model integrations (e.g., 30–100 ensemble members) in combination with a localization performed with a Schur product of a tapering correlation function. The localization not only eliminates the spurious correlations at large distances but also increases the rank of the sample covariance to a value larger than the dimension of the state-space (as required by the assimilation algorithm), thus providing a full rank forecast error covariances to assimilate observations (Houtekamer and Zhang 2016). A comparison between EnKF and 4D-Var chemical data assimilation methods (i.e., to estimate the state, not the emissions) was conducted using a stratospheric chemistry transport model (Skachko et al. 2014; Skachko et al. 2016).

Another approach designed for the assimilation of long-lived chemical species is based on simplified Kalman filtering. This method is well-adapted to large state-space and

linear problems and exploits the properties of the continuity equation. In this scheme, the error variance is computed explicitly using only a single model integration. This assimilation method, historically also known as the sequential filter, was first used in the assimilation of long-lived chemical species in the stratosphere, both in case studies and over long time periods. In particular, it was used for the assimilation of MOPITT CO to perform a reanalysis of stratospheric ozone and to diagnose model error (Khattatov et al. 2000; Menard and Chang 2000; Menard et al. 2000; Eskes et al. 2003; Rosevall et al. 2007; van der A et al. 2010). This scheme is, in fact, part of a larger class of algorithms under development called parametric Kalman filters (PKF), where additional covariance parameters, such as correlation lengths, are evolved over time (Pannekoucke et al. 2016; Pannekoucke et al. 2018a; Pannekoucke et al. 2021). In its simplest form, where only the error variance is evolved, we will call this algorithm the PvKF for parametric variance Kalman filter. The PvKF requires only two model integrations, one for the state estimate and the other for its error variance (or uncertainty) and seems well-adapted for the assimilation of methane as a long-lived species. Considering a lifetime of about 10 years (Prinn et al. 2005; Prather et al. 2012) and limited atmospheric chemistry (Jacob et al. 2016), together with a linear observation mapping with smooth averaging kernels (such as those from GOSAT), renders the whole assimilation problem quasi-linear. In terms of algorithm computational efficiency, the applicability of the parametric Kalman filter (or PvKF) to large state-space is made from the correspondence between the continuous and discrete representations of the different operators, more specifically between spatial correlation functions and the corresponding correlation matrices. This idea was originally developed for the Optimal Interpolation (OI) method. Although OI is often used in

conjunction with a data selection procedure to process a large number of observations by batches, we may not need such a procedure (data selection) if the number of profiles per time step is relatively small so that matrices in observation space can be directly inverted. The main idea underlying the applicability of OI to large state-space comes from computing the error correlations on the fly, as needed, using a continuous correlation function between a pair of points without the need for the storage of a matrix in state space. It also has the advantage that spatial correlation functions can be represented on any grid (Gaspari and Cohn 1999). Note that the comparison of the computational cost mentioned above is for different assimilation schemes (not emissions inversion). The purpose of this chapter is to design the assimilation of GOSAT observations using the PvKF approach with the hemispheric version of the Community Multiscale Air Quality (CMAQ) model. In the next chapter, we use this assimilation system to obtain optimal assimilation of GOSAT observations and realistic error covariance parameters, including observation and model error and correlation lengths. We envision that the development of the current hemispheric assimilation system will be used in the future on a limited domain for joint data assimilation and emission inversion using the regional CMAQ model.

The organization of Chapter 5 is as follows. In Section 5.2, first, we present the modifications made to CMAQ to include methane as an evolving species and, in particular, how anthropogenic and natural emissions of methane can be incorporated into the SMOKE processing system. Then, we explain how to use the observation averaging kernels and a priori provided by the GOSAT data in a form useful for the hemispheric CMAQ grid and for assimilation. In Section 5.3, we describe the PvKF and how the scheme can be adapted for a large state space, such as the CMAQ model. Section 5.4 details other aspects of the

assimilation system, specifically, the initial conditions, the observation bias correction, and the formulation of homogeneous isotropic correlation models in the uniform hemispheric CMAQ grid. We close this section with a description of the parameters used in the error covariance matrices. Finally, in Section 5.5, we conduct simulated observation experiments, in particular, the one-observation experiment that permits the verification of the algorithm and the assumptions used for propagation of error variance. We conclude Chapter 5 with an estimation of the computational cost of such an algorithm.

## **5.2 Model and Observation Operator**

### **5.2.1 Modifications of CMAQ to Handle Methane Transport and Emissions**

CMAQ is a limited-area model developed by the U.S. Environmental Protection Agency (EPA) (Byun and Schere 2006) that is used as a regional model driven at the lateral boundaries by the hemispheric version of this model (referred to as H-CMAQ) (Mathur et al. 2017). H-CMAQ, based on CMAQ v5.3, is used here to simulate and assimilate methane in the Northern Hemisphere. For all practical purposes, a hemispheric model does not need to account for the inter-hemispheric exchange of mass for time scales up to several months (Brasseur and Jacob 2017). The domain covers the Northern Hemisphere on a polar stereographic projection (see Figure C.1 in Appendix C1), which includes  $187 \times 187$  grid cells horizontally with a 108 km grid spacing and 44 vertical layers extended from the surface to the model top at 50 hPa. The modified vertical structure of H-CMAQ maintains finer resolution than regional CMAQ above the boundary layer, aiming at a better representation of transport processes, also suitable for long-lived species (Mathur et al. 2017). We modified H-CMAQ to account for methane concentrations and emissions. Methane can be configured either as an inert trace gas or a reactive gas-phase species

oxidized by hydroxyl radical (OH). Figure C.3 (Appendix C4) shows an example of concentration loss due to chemical reactions after two weeks. We considered reactive methane with the gas-phase model within CMAQ v5.3 and based on the CB06 chemical mechanism (CMAQ tutorials 2021). The error variance is treated as a chemically inert tracer in CMAQ, subject only to an advection-only transport scheme.

Our data assimilation system includes methane emissions of both anthropogenic (~60%) and natural (~40%) sources. For the anthropogenic emissions, we obtained all available source sectors from the Emission Database for Global Atmospheric Research (EDGAR v6) (Crippa et al. 2020; Crippa et al. 2021). These sources come with  $0.1^\circ \times 0.1^\circ$  horizontal resolution and monthly temporal resolution emission grid maps. We processed them using Sparse Matrix Operator Kernel Emissions (SMOKE v3.6) (UNC 2017) to provide hourly gridded methane emissions into the model. Wetlands are the primary source of natural emissions, accounting for about 80% of the total (Cressot et al. 2014; Jacob et al. 2016). We used monthly wetlands emissions from WetCHARTs v1.0 with the full ensemble mean (Bloom et al. 2017) and mapped it into our domain using a flat hourly/weekly temporal profile.

### **5.2.2 GOSAT Observation Operator for Data Assimilation**

Greenhouse Gas Observing Satellite (GOSAT) was launched in January 2009 by the Japanese Space Agency (JAXA) (Kuze et al. 2009). It is in a Sun-synchronous orbit at an altitude of 666 km with a 3-day revisit time and an equator overpass time at about 13:00 local time. One main goal of GOSAT is monitoring the abundance of methane in Earth's atmosphere. Due to the high sensitivity of the shortwave infrared (SWIR) GOSAT retrieval

at the surface, coupled with a suitable spatiotemporal resolution, the assimilation of atmospheric methane and inverse modelling of its sources and sinks are suitable (Butz et al. 2011; Schepers et al. 2012; Turner et al. 2015; Parker et al. 2015; Buchwitz et al. 2017).

The instrument, TANSO-FTS, onboard GOSAT, has a field of view with a 10.5 km diameter footprint operating in a cross-track scanning mode. It measures the abundance of methane by analyzing the backscattered solar radiance spectrum in the SWIR near 1.6  $\mu\text{m}$ . A column-average dry-mole fraction of methane ( $X_{\text{CH}_4}$ ) represents the instrument retrieved observation, which corresponds to the methane average volume mixing ratio ( $VMR \equiv y^o$ ) of a partial column atmosphere with a given surface,  $p_S$ , and top pressure,  $p_T$ . Two approaches are used to derive retrieval algorithms: Full Physics (FP) and Proxy (PR). The FP method integrates a sophisticated radiative transfer model and solely relies on methane modelling and its corresponding errors, while the PR algorithm provides more data points but relies on an accurate  $\text{CO}_2$  model simulation and its retrieval ( $X_{\text{CO}_2}$ ) (Schepers et al. 2012). Both algorithms were developed at the Netherlands Institute for Space Research (SRON) and Karlsruhe Institute for Technology (KIT) (Butz et al. 2011), and their products are available through the ESA GHG-CCI initiative, <https://climate.esa.int/en/projects/ghgs/> (accessed on 10 January 2021) (Buchwitz et al. 2017). We use both products, but mainly focus on PR due to its higher density and/or better coverage (Butz et al. 2011; Schepers et al. 2012).

For assimilation purposes, in addition to the retrieval data, we use supplementary products (Bovensmann et al. 1999; Buchwitz et al. 2017). Thus, each retrieval also includes vectors of the normalized column-average kernel,  $\mathbf{A}$ , pressure levels,  $p_l$ , at which the average kernels are derived, and the corresponding vector of a priori,  $\mathbf{y}^p$ . PR and FP

retrieval algorithms offer 5 and 13 levels of data in the vertical, respectively. A layer-based approach described in Bergamaschi et al. (2007) is applied to compute the model partial-column value,  $y^m$ , equivalent to the retrieval,  $y^o$

$$y^m = [(\mathbf{1} - \mathbf{A})\mathbf{y}^p + \mathbf{A}\mathbf{y}^m] \boldsymbol{\omega}^T \quad (5.1)$$

$\mathbf{y}^m$  represents the mapped concentration of H-CMAQ on the pressure layers of the observation, and  $\boldsymbol{\omega}$  is the vector of pressure layers weights, whose elements are expressed as

$$\omega_l = \frac{p_l - p_{l+1}}{p_S - p_T}, \quad (5.2)$$

where  $p_S$  is the surface pressure and  $p_T$  is the model top pressure. Note that Equations (5.1) and (5.2) are in the retrieval space so that all the vectors represent the vertical grids within a single retrieval. Now, let us consider the 3D model estimate of methane concentrations at time,  $t$ , in a vectorized form,  $\mathbf{X}$ , with  $N_x$  dimension. The observation vector,  $\mathbf{Y}^o$ , consists of a set of retrievals at approximately the same time but in different locations from model grid points. It also provides the same type of quantity but with a dimension quite smaller than the model ( $N_y \ll N_x$ ). Assuming a linear relationship between  $\mathbf{Y}^o$  and  $\mathbf{X}$ , we have a linear observation operator,  $\mathbf{H}$ .

$$\mathbf{Y}^o = \mathbf{H}(\mathbf{X}) + \boldsymbol{\varepsilon}^o. \quad (5.3)$$

In fact,  $\mathbf{H}$  is a combination of two linear operators, comprising a horizontal,  $\mathbf{H}_h$ , and a vertical,  $\mathbf{H}_v$ , operator,

$$\mathbf{H}(\mathbf{X}) = \mathbf{H}_v(\mathbf{H}_h(\mathbf{X})), \quad (5.4)$$

where  $\mathbf{H}_h$  interpolates the components of  $\mathbf{x}$  from the horizontal model grid points to observation locations using a bi-linear interpolation function, and  $\mathbf{H}_o$  transforms  $\mathbf{x}$  at the geographical location of observations to a vector equivalent to  $\mathbf{Y}^o$  (using averaging kernels). Thus, the assimilation problem consists of finding the best estimate of  $\mathbf{x}$  with its error statistics, which together are called the analysis.

### 5.3 Data Assimilation System

#### 5.3.1 Background of the Assimilation Scheme

First, let us define a correlation function on an arbitrary grid following Gaspari and Cohn (1999). It is sufficient to define a correlation function in  $\mathbb{R}^3 \times \mathbb{R}^3$ , so that any subspace (or manifold) of  $\mathbb{R}^3$  (e.g., the surface of a sphere) also define a correlation on that subspace. This property was used in this study to define underlying homogeneous isotropic correlation functions (with periodicity on a sphere), which are then mapped onto the polar stereographic grid of H-CMAQ, which has a uniform grid spacing on the projected plane. PvKF uses these concepts, together with the property that for the advection equation, the error variance can be forecast without knowing the error correlation (Cohn 1993). Thus, in a PvKF framework, the error variances are dynamically evolving according to the model's advection scheme, but the spatial error correlations are kept fixed and are computed using the same approach as taken by OI.

The algorithm of PvKF is decomposed into two steps; a forecast step and an analysis step, which accounts for the effect of observations. A covariance function,  $P(\mathbf{x}, \mathbf{x}', t)$ , is a function of a pair of points,  $\mathbf{x} = (x, y, z)$  and  $\mathbf{x}' = (x', y', z')$ , at time  $t$ . It is related to the correlation function,  $C(\mathbf{x}, \mathbf{x}', t)$ , through the standard expression

$$P(\mathbf{x}, \mathbf{x}', t) = \sigma(\mathbf{x}, t) C(\mathbf{x}, \mathbf{x}', t) \sigma(\mathbf{x}', t), \quad (5.5)$$

where  $\sigma$  is the error standard deviation (i.e.,  $\sigma^2(\mathbf{x}, t)$  is the error variance at the point  $\mathbf{x}$  and time  $t$ ). In PvKF, the error correlation is generally time-invariant, homogeneous and isotropic in the horizontal (i.e., depends only on the horizontal distance). 3D spatial correlations are constructed using a horizontal/vertical separability assumption,

$$C(x, y, z, x', y', z') = C_h(x, y, x', y') C_v(z, z'), \quad (5.6)$$

and for the horizontal correlation

$$C_h(x, y, x', y') = C_h(\|(x, y) - (x', y')\|), \quad (5.7)$$

is assumed to be homogeneous and isotropic. Note that the separability assumption has been verified for long-lived species in the stratosphere by Menard et al. (2019) (see Figure S3 therein).

### 5.3.2 Forecast Step

The forecast step consists of two model integrations, one for the state and the other for the error variance,

$$\begin{aligned} X(\mathbf{x}, t) &= \mathbf{M}(X(\mathbf{x}, t - \delta t)) \\ \sigma^2(\mathbf{x}, t) &= \mathbf{M}^*(\sigma^2(\mathbf{x}, t - \delta t)) \end{aligned} \quad (5.8)$$

where  $\mathbf{M}$  represents the chemical transport model based on the atmospheric diffusion equation (Jacobson 2020; Seinfeld and Pandis 2016) of CMAQ, and  $\mathbf{M}^*$  denotes the advection model of error variances based on the CMAQ native advection scheme.  $X(\mathbf{x}, t)$  represents the chemical concentration as a function of 3-dimensional model space  $\mathbf{x}=(x, y, z)$  and time  $t$ . Note that the error covariance forecast step of the PvKF does not follow the standard Kalman filter equation, but uses the continuous formulation of

propagation of error covariances, which can be solved by using the method of characteristic as discussed in Cohn (1993). This approach was applied in several studies on the assimilation of long-lived species (Khattatov et al. 2000; Menard and Chang 2000; Menard et al. 2000; van der A et al. 2010). Using the advection of error variance also has two advantages. First, it avoids the loss of error variance in standard Kalman filter formulation, such as in EnKF. In fact, the main cause of the loss of error variance is due to propagating error in a discretized form of Kalman filtering (see Section 2 in Menard et al. (2021); also see Menard and Chang (2000)). Secondly, it significantly reduces the computational cost compared to other methods (e.g., EnKF). A demonstration of how the PvKF assimilation can avoid the loss of error variance is presented in Section 5.5.1, and its computational cost is compared with the model forecast in Section 5.5.2.

The  $\mathbf{M}^*$  operator is, in fact, the integration from  $t - \delta t$  to  $t$  of

$$\frac{\partial \sigma^2}{\partial t} + \mathbf{V} \cdot \nabla \sigma^2 = q, \quad (5.9)$$

where  $\mathbf{V}$  is the 3-dimensional advection wind, and  $q(\mathbf{x}, t)$  accounts for the model error variance growth (e.g., errors of any other processes that are not advection; we discuss further in Section 5.4.4 and in Section 6.4). It is important to note that in the context of data assimilation, the effect of the observations usually has the largest impact on error covariances, so that it is common to use a simpler model to forecast the error covariances or to run the adjoint model (e.g., incremental 4D-Var by Courtier et al. (1994)) used operationally for weather forecasting using a lower resolution model with simplified physics). Here, as in the incremental approach, we use only advection in the forecast of error variances.

After the integration time,  $\delta t$ , we obtain a forecast concentration,  $X^f(\mathbf{x}, t)$ , and a forecast error variance,  $(\sigma^f)^2$ , from which the covariance function can be reconstructed by applying the correlation function to Equation (5.5). Therefore, a forecast error covariance function,  $P^f(\mathbf{x}, \mathbf{x}', t)$ , is obtained, considering that the correlation function is time-invariant (i.e., stationary). Thus,

$$\begin{aligned} P^f(\mathbf{x}, \mathbf{x}', t) &= \sigma^f(\mathbf{x}, t) C(\mathbf{x}, \mathbf{x}') \sigma^f(\mathbf{x}', t) \\ C(\mathbf{x}, \mathbf{x}') &= C_h(\|(x, y) - (x', y')\|) C_v(z, z') \end{aligned} \quad (5.10)$$

and equivalently in matrix notation

$$\begin{aligned} \mathbf{P}^f(t) &= \mathbf{\Sigma}^f(t) \mathbf{C} \mathbf{\Sigma}^f(t) \\ \mathbf{C} &= \mathbf{C}_h \otimes \mathbf{C}_v \end{aligned} \quad (5.11)$$

where  $\mathbf{\Sigma}^f$  is a diagonal matrix of forecast error standard deviations and  $\otimes$  is the Kronecker product of matrices. The matrix form is useful to write out the analysis step.

### 5.3.3 Analysis Step

Let us start as if  $\mathbf{H}$  is only a horizontal operator (we will discuss the vertical aspects in the following Section 5.3.4). A computational simplification in the analysis step arises when observations are considered horizontally as point measurements. This allows simplification of the Kalman gain matrix and the analysis error covariance and variance. Horizontally, a point measurement observation operator can be modelled as a delta function at the observation location,  $\mathbf{x}_o$ . Specifically, if  $f(\mathbf{x})$  is a continuous function of space,  $\mathbf{x}$ , and the horizontal observation operator,  $H_h$  (as previously defined in Section 5.2), applied on  $f$  ( $H_h[f]$ ), gives the value of the function at the observation location,  $f(\mathbf{x}_o)$ , we can write

$$H_h[f] = f(\mathbf{x}_o) = \int f(\mathbf{x}) \delta(\mathbf{x} - \mathbf{x}_o) d\mathbf{x} \quad (5.12)$$

The  $H_h[\cdot]$  operator is thus associated with a delta function,  $\delta(\mathbf{x} - \mathbf{x}_o)$ . Although we never use this representation of the observation operator on the state fields themselves since they are represented discretely on a grid, we do apply this definition when we consider the application of the observation operator on error covariance functions. Thus, we have the following,

$$[\mathbf{HP}^f]_{o,j} \Leftrightarrow H_h[P^f]_{o,j} = \int P^f(\mathbf{x}, \mathbf{x}'_j) \delta(\mathbf{x} - \mathbf{x}_o) d\mathbf{x} = P^f(\mathbf{x}_o, \mathbf{x}'_j), \quad (5.13)$$

$$[(\mathbf{HP}^f)^T]_{i,o} = [\mathbf{P}^f \mathbf{H}^T]_{i,o} \Leftrightarrow H_h^T[P^f]_{i,o} = \int P^f(\mathbf{x}_i, \mathbf{x}') \delta(\mathbf{x}' - \mathbf{x}'_o) d\mathbf{x}' = P^f(\mathbf{x}_i, \mathbf{x}'_o), \quad (5.14)$$

where the subscript  $i$  and  $j$  refer to model grid points, whereas subscript  $o$  refers to the position of the observations.

Considering the covariance function between any pair of observation locations,  $\mathbf{x}_{o(1)}, \mathbf{x}_{o(2)}$

, we get an element of the matrix  $\mathbf{HP}^f \mathbf{H}^T$ . That is

$$\begin{aligned} [\mathbf{HP}^f \mathbf{H}^T]_{o(1),o(2)} &\Leftrightarrow H_{h,o(2)}[H_{h,o(1)}[P^f]]_{o(1),o(2)} \\ &= \int P^f(\mathbf{x}, \mathbf{x}') \delta(\mathbf{x} - \mathbf{x}_{o(1)}) \delta(\mathbf{x}' - \mathbf{x}'_{o(2)}) d\mathbf{x} d\mathbf{x}' = P^f(\mathbf{x}_{o(1)}, \mathbf{x}_{o(2)}) \end{aligned} \quad (5.15)$$

Suppose now we have  $K$  observations (with different locations) that are used in an analysis. The error covariance matrix in observation space (i.e., between each pair of observations) is  $\mathbf{HP}^f \mathbf{H}^T$  (a  $K \times K$  matrix), where each element of this matrix is given by Equation (5.15). With this preamble, we are now able to formulate the analysis step for both the state estimate and the error variance. The analysis state is written as usual and requires the computation of the Kalman gain matrix  $\mathbf{K}$ ,

$$\mathbf{X}^a = \mathbf{X}^f + \mathbf{K}(\mathbf{Y}^o - \mathbf{H}\mathbf{X}^f), \quad (5.16)$$

where

$$\mathbf{K} = (\mathbf{HP}^f)^T (\mathbf{HP}^f \mathbf{H}^T + \mathbf{R})^{-1}, \quad (5.17)$$

$\mathbf{R}$  represents the observation error covariance matrix. We should remark that in data assimilation, the observation error is not simply the instrument error, but should also contain an error of representativeness since, ultimately, observations are used to update the model state (see Section 5.4.4)

All state quantities (i.e., the forecast and the analysis) are written as vectors of dimension  $N_x$ , where  $N_x$  is the number of model grid points. We recall that  $\mathbf{Y}^o$  is a vector of dimension  $K$  that contains all observations processed in data assimilation. The matrix  $\mathbf{HP}^f$  is sometimes called the matrix of representors, where each observation location comes with all the model grid points. We can think of it as a vector of *impact covariance functions*.  $\mathbf{HP}^f$  can be denoted in column form as

$$\mathbf{HP}^f = (\mathbf{p}_1 \quad \mathbf{p}_2 \quad \cdots \quad \mathbf{p}_N), \quad (5.18)$$

$$\mathbf{p}_i = \begin{pmatrix} P^f(\mathbf{x}_{o(1)}, \mathbf{x}_i) \\ P^f(\mathbf{x}_{o(2)}, \mathbf{x}_i) \\ \cdots \\ P^f(\mathbf{x}_{o(K)}, \mathbf{x}_i) \end{pmatrix}, \quad (5.19)$$

where each  $\mathbf{p}_i$  is a column vector ( $K \times 1$ ) that represents the forecast error between all observations and a single model grid point that we call the *sensitivity error covariance function*— in matrix form,  $(\mathbf{HP}^f)^T$ . The analysis update of the error covariance, called the analysis error covariance in data assimilation, takes the form,

$$\mathbf{P}^a = (\mathbf{I} - \mathbf{KH})\mathbf{P}^f = \mathbf{P}^f - (\mathbf{HP}^f)^T (\mathbf{HP}^f \mathbf{H}^T + \mathbf{R})^{-1} (\mathbf{HP}^f), \quad (5.20)$$

is derived using Equation (5.17).

In particular, an element  $(i, j)$  of the analysis error covariance matrix,  $\mathbf{P}^a$ , can be conveniently written as

$$\mathbf{P}_{ij}^a = \mathbf{P}_{ij}^f - \mathbf{p}_i^T \mathbf{\Gamma}^{-1} \mathbf{p}_j, \quad (5.21)$$

where

$$\mathbf{\Gamma} = \mathbf{H} \mathbf{P}^f \mathbf{H}^T + \mathbf{R}, \quad (5.22)$$

represents the modelled innovation covariance matrix. The analysis error variance, i.e., when  $j = i$ , is simply computed as

$$(\sigma_i^a)^2 = (\sigma_i^f)^2 - \mathbf{p}_i^T \mathbf{\Gamma}^{-1} \mathbf{p}_i. \quad (5.23)$$

The analysis error variance is then used as an initial condition to integrate the variance through Equation (5.9). This completes the algorithm. The next section deals with the three-dimensional observation operator (Equation (5.4)), which includes the vertical structure of the averaging kernel, rendering the analysis update more complex.

### 5.3.4 Analysis Step with 3D Observation Operator Using Averaging Kernels

The three-dimensional observation operator is given in Equation (5.4). First, it consists of interpolating the 3D field horizontally, as in Equation (5.12), to get vertical profiles at the observation location and then applying the vertical observation operator,  $\mathbf{H}_v$ , using the instrument averaging kernel to get the observation equivalent quantity. The vertical observation operator in function form is denoted as  $H_{v(z)}$ . To apply it, first, we use a mapping function (i.e., vertical interpolation,  $V_z$ , on  $z$ ) to convert the model vertical concentrations to the vertical layers of the observation retrieval, so that  $\mathbf{y}^m$  (see Equation (5.1)), the a priori  $\mathbf{y}^p$ , and the averaging kernel  $\mathbf{A}$  are in the same vertical coordinates. Thus, in an observation location, we get

$$V_z[f_o(z,t)] \Rightarrow \mathbf{y}^m(z^*,t), \quad (5.24)$$

where  $f_o$  denotes the vertical model concentrations at the observation location and  $z^*$  is the vertical coordinate as in averaging kernels.

In the second step,  $H_{v(z)}$  carries out  $\mathbf{y}^m(z^*,t)$  in Equation (5.1) and computes a single value equivalent to the retrieval,  $y^m$  (see Section 5.2.2 for an explanation to Equation (5.1)).

$$H_{v(z)}[V_z[f_o(z,t)], \mathbf{A}, \mathbf{y}^p, \boldsymbol{\omega}] \Rightarrow y^m \quad (5.25)$$

Note that the quantities in bold ( $\mathbf{A}, \mathbf{y}^p, \boldsymbol{\omega}, \mathbf{y}^m$ ) are defined on the vertical layers of the averaging kernel, and the non-bold ( $y^m$ ) is, in fact, a scalar quantity representing the column-averaged quantity. Therefore,  $H_{v(z)}$  maps a vector in the model vertical space to a scalar in the observation space.

The observation operator  $H$ , which is applied on functions, is the composition of  $H_{v(z)}$  with  $H_h$ ,

$$H = H_{v(z)} \circ H_h \Leftrightarrow \mathbf{H}_v(\mathbf{H}_h(\quad)), \quad (5.26)$$

which is equivalent to the matrix form of the  $\mathbf{H}$  in Equation (5.4).

Before applying the observation operator on a covariance function, we need to specify on which variable it is applied. If it is applied on the spatial variable,  $\mathbf{x}$ , it is equivalent, in matrix form, to a left multiplication of the observation operator on the covariance matrix,

$$H_{\mathbf{x}}[P^f(\mathbf{x}, \mathbf{x}', t)] \Leftrightarrow \mathbf{H} \mathbf{P}^f. \quad (5.27)$$

However, if the observation operator is applied to the spatial variable,  $\mathbf{x}'$ , it is then equivalent to a right multiplication of the transpose of the observation operator on the covariance matrix,

$$H_{\mathbf{x}'}[P^f(\mathbf{x}, \mathbf{x}', t)] \Leftrightarrow \mathbf{P}^f \mathbf{H}^T. \quad (5.28)$$

Using the separable form of the covariance function in Equation (5.10), we thus get

$$\begin{aligned} H_{\mathbf{x}}[P^f(\mathbf{x}, \mathbf{x}', t)] &= H_{v(z)} \left\{ \sigma^f(x_o, y_o, z, t) C_h(x_o, y_o, x', y') C_v(z, z') \sigma^f(x', y', z', t) \right\}, \\ &= H_{v(z)} [\sigma_o^f(z) C_v(z, z')] \times C_h(x_o, y_o, x', y') \sigma^f(x', y', z', t) \end{aligned} \quad (5.29)$$

where  $H_{v(z)}$  is to denote, in functional form, the vertical observation operator operating on  $z$  (and not  $z'$ ). However, to use the vertical observation operator, the expression in square brackets in Equation (5.29) has to be written in vector/matrix form. Note that in matrix form,  $\sigma_o^f(z, t) C_v(z, z')$  is an element of an  $N_{lev} \times N_{lev}$  matrix, where  $N_{lev}$  is the number of vertical levels in the model. This matrix can in fact be written as  $\mathbf{diag}(\sigma_o^f) \mathbf{C}_v$ , where  $\mathbf{diag}(\sigma_o^f)$  is an  $N_{lev} \times N_{lev}$  diagonal matrix where the elements on the diagonal are  $\sigma_o^f(z, t)$ . Therefore, the application

$$H_{v(z)} [\sigma_o^f(z, t) C_v(z, z')] = f_o(z', t) \Rightarrow \mathbf{f}_o^T \text{ a row vector } 1 \times N_{lev}, \quad (5.30)$$

is a row vector that depends on the (second) vertical coordinate,  $z'$ . Overall, the application of the observation operator in Equation (5.29) produces a 3D spatial field of elements,

$$H_{\mathbf{x}}[P^f(\mathbf{x}, \mathbf{x}', t)] = C_h(x_o, y_o, x', y') f_o(z', t) \sigma^f(x', y', z', t). \quad (5.31)$$

Based on Equation (5.28), a similar expression is obtained

$$H_{\mathbf{x}'}[P^f(\mathbf{x}, \mathbf{x}', t)] = C_h(x, y, x'_{o(2)}, y'_{o(2)}) f_{o(2)}(z, t) \sigma^f(x, y, z, t), \quad (5.32)$$

where we let the reference observation location be  $o(2)$ . To compute the (simulated) innovation covariance matrix (i.e.,  $\mathbf{HP}^f\mathbf{H}^T$ ) we combine Equation (5.27) with Equation (5.28),

$$H_{\mathbf{x}}\left[H_{\mathbf{x}'}\left(P^f(\mathbf{x},\mathbf{x}',t)\right)\right] \Leftrightarrow \mathbf{HP}^f\mathbf{H}^T, \quad (5.33)$$

and account for a pair of non-coincident observations,  $o(1)$ ,  $o(2)$ , we get

$$\begin{aligned} & H_{\mathbf{x}}[C_h(x,y,x'_{o(2)},y'_{o(2)})f_{o(2)}(z,t)\sigma^f(x,y,z,t)] \\ &= H_{v(z)}\left\{C_h(x_{o(1)},y_{o(1)},x'_{o(2)},y'_{o(2)})f_{o(2)}(z,t)\sigma_{o(1)}^f(z,t)\right\}, \\ &= C_h(x_{o(1)},y_{o(1)},x'_{o(2)},y'_{o(2)})H_{v(z)}[f_{o(2)}(z,t)\sigma_{o(1)}^f(z,t)] \\ &= C_h(x_{o(1)},y_{o(1)},x'_{o(2)},y'_{o(2)})\times\beta_{o(1),o(2)}(t) \end{aligned} \quad (5.34)$$

where  $\beta_{o(1),o(2)}$  is an element of a  $N_{lev}\times N_{lev}$  covariance matrix,

$$\beta_{o(1),o(2)} = H_{v(z)}\left\{H_{v(z')}\left[\sigma_{o(1)}^f(z,t)\sigma_{o(2)}^f(z',t)C_v(z,z')\right]\right\}. \quad (5.35)$$

### 5.3.5 An Overview of the Assimilation Algorithm

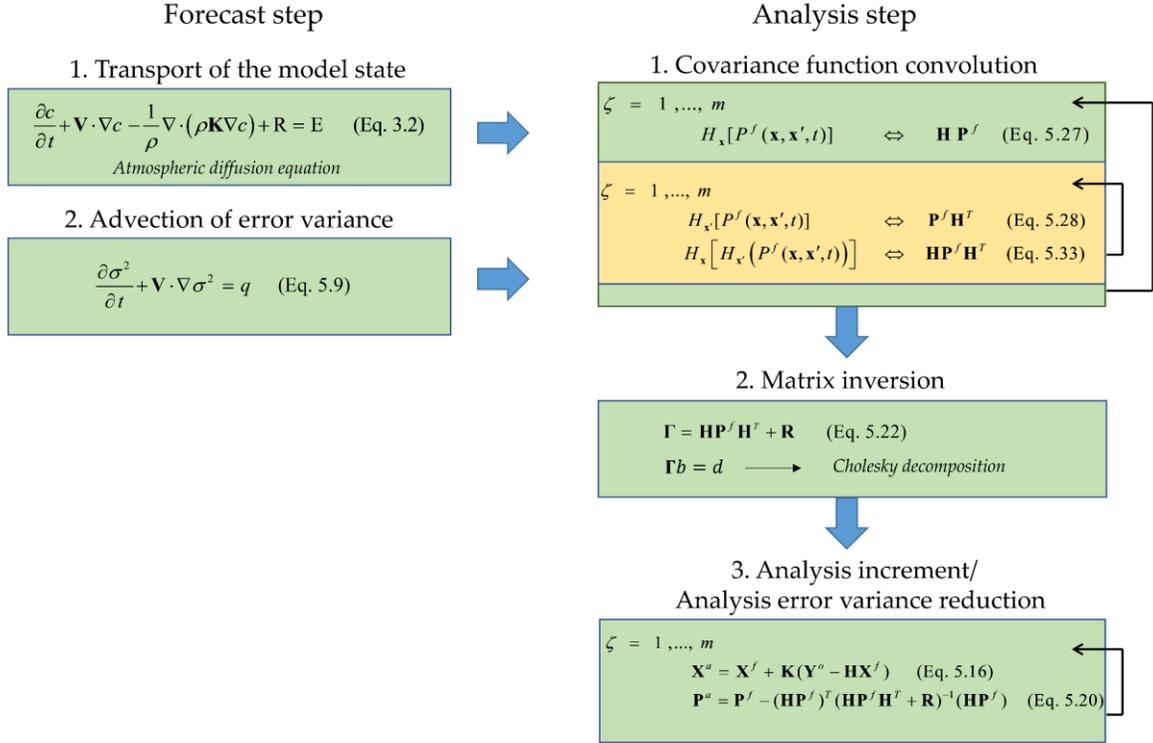
Figure 5.1 presents a flowchart of the forecast and analysis steps of the PvKF algorithm described above. The algorithm's forecast step (left side) involves two parallel model simulations, one for the state's transport with the atmospheric diffusion equation (Jacobson 2020; Seinfeld and Pandis 2016) and the other for the advection of the error variance (Equation (5.9)). In the first part of the analysis step (right side), the covariance function is convolved with the observation operator on the left for a series of available observations at a given time (Equation (5.27)). This is followed by a second convolution with the observation operator on the right for the same observations to obtain the innovation covariance matrix (Equations (5.28) and (5.33)–(5.35)). The second part of the analysis step (Equation (5.22)) performs a Cholesky decomposition and inversion of the

modelled innovation covariance matrix. Finally, the analysis and analysis error variance increment are computed for each set of observations at a time  $t$  (Equations (5.16) and (5.23)). A complete PvKF assimilation algorithm in a matrix form is presented in Appendix C2.

As mentioned earlier, the PvKF assimilation relies on the continuous properties of operators, which makes it a suitable scheme for evolving the analysis and its error variance. An example of the error variance (i.e., the uncertainty of the state estimation) evolution with synthetic GOSAT observations over land is shown in Figure 5.2, emphasizing the combination of the 3<sup>rd</sup> part of the analysis step and the 2<sup>nd</sup> part of the forecast step of the flowchart (Figure 5.1) in creating the analysis error variance evolution. In Figure 5.2, the error variance is initiated with a constant field using a 5% error (standard deviation) of the globally averaged methane concentration (Figure 5.2a). The reduction of error variance occurs in the presence of every observation at a particular time and location based on the correlation function (Figure 5.1–analysis step) while the updated error variance field is propagated with the flow (Figure 5.1–forecast step). The reduction gradually grows over land (Figure 5.2b) by assimilating a new batch of observations over the same regions according to GOSAT revisit time (i.e., 3 days).

After 8 days, the analysis error variance field (Figure 5.2c) shows a noticeable reduction over Western Asia, followed by Central Asia, Eastern Africa, and the southern/eastern part of North America, mainly due to denser GOSAT observations and higher innovations at those regions. Then, owing to the advection, the reduction of error variance spread out with the flow over other regions with fewer or no observations, such as oceans and higher latitudes. On day 12 (Figure 5.2d), most of the lands, except polar

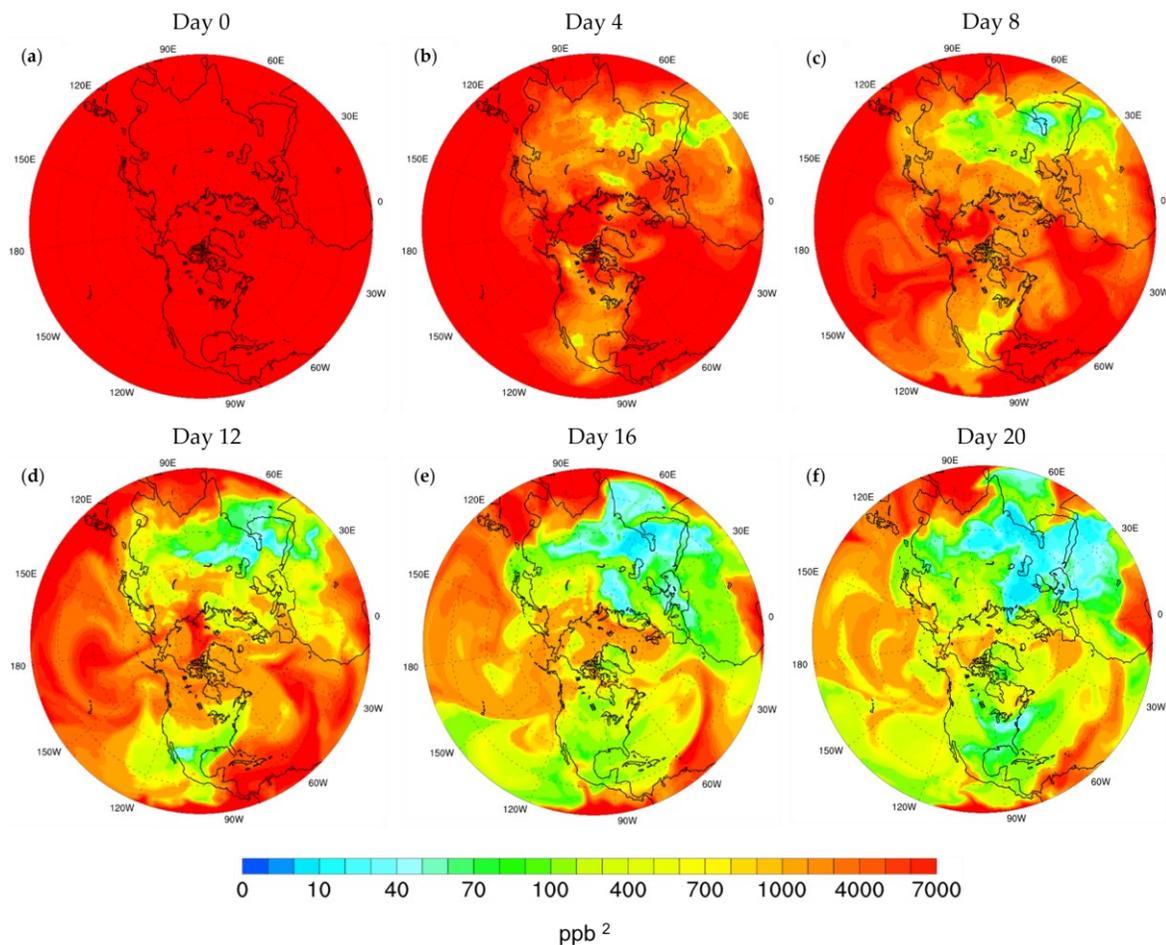
regions and near the Equator, are subjected to a significant error reduction, suggesting a higher impact of the GOSAT observations due to their density over those areas.



**Figure 5.1.** Flowchart of the analysis and forecast steps of the PvKF assimilation. In the forecast step, the transport of the model state uses the atmospheric diffusion equation (Jacobson 2020; Seinfeld and Pandis 2016).

The maximum reduction (light blue) indicates a 20–40 times reduction in the initial error, while assimilating more observations within PvKF after 16 days does not provide a tangible reduction in the analysis error (Figure 5.2e). Nonetheless, over time, the variance reduction significantly spreads out over the entire domain, particularly over the oceans with no observations (Figure 5.2f). This highlights the behaviour of the assimilation system that provides an error estimation over the whole domain. Note that in the experiment of Figure 5.2, we consider a uniform and high value of initial error variance to emphasize and isolate the impact of adding more information by observations on the estimation uncertainties over

time. In Section 5.5.1 and Chapter 6, the initial field of uncertainties are not uniform but is more realistic and is obtained through the initial concentration field.



**Figure 5.2.** Evolution of the analysis error variance on (a) Day 0, (b) Day 4, (c) Day 8, (d) Day 12, (e) Day 16, and (f) Day 20.

## 5.4 System Setup

### 5.4.1 Initial Conditions

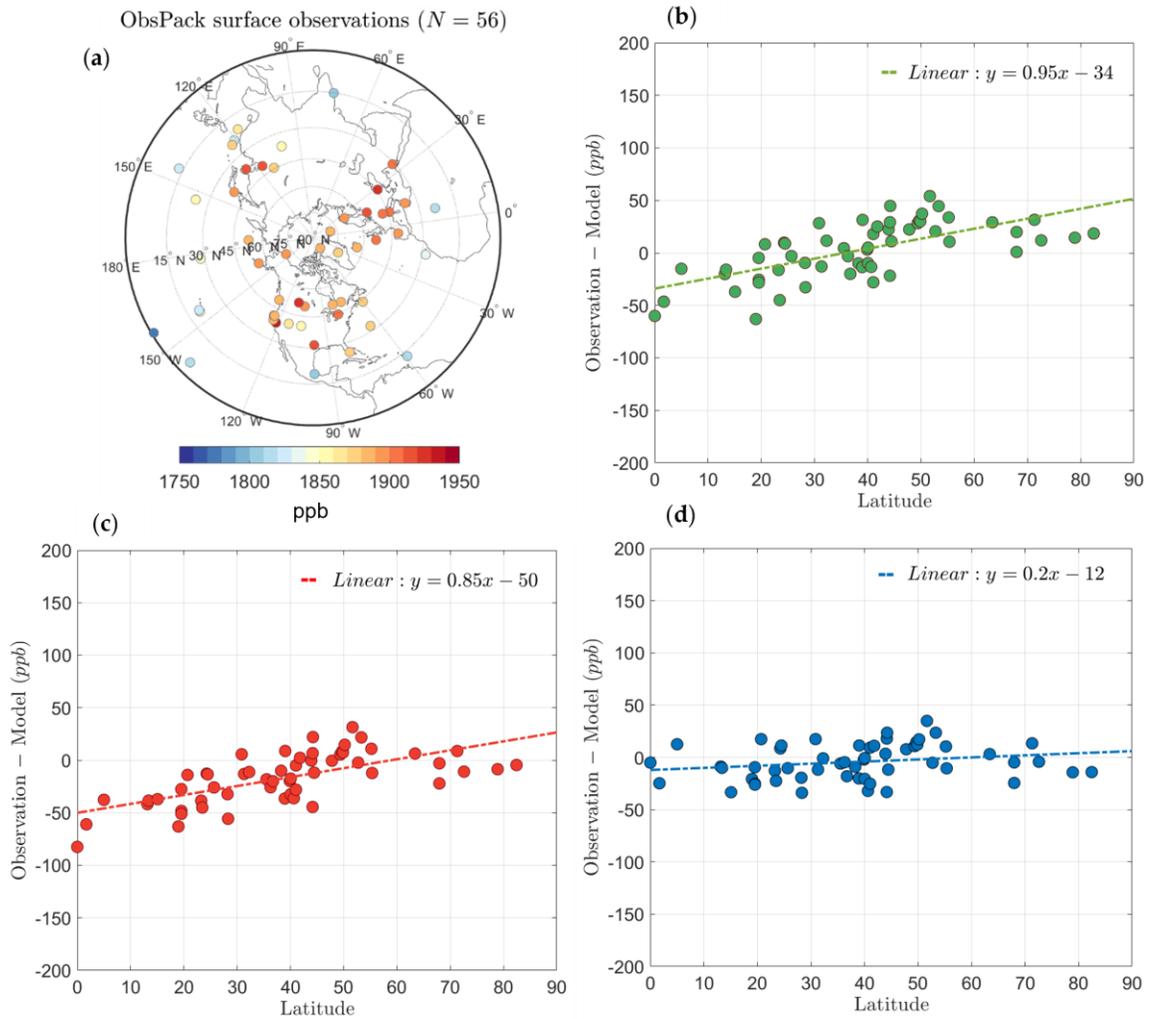
The impact of the initial condition can persist long in the model. For different species, it is recommended that H-CMAQ be initialized to the clean background and driven based on its emissions and chemical transport (Mathur et al. 2017). However, due to an almost well-known background field of methane, the long lifetime of atmospheric methane

(~10 years) (Prinn et al. 2005), and the high level of uncertainties in its source and sink (Kirschke et al. 2013), conducting a lengthy model integration might be unnecessary. Furthermore, in the context of data assimilation, the sensitivity to the initial conditions disappears rapidly with time (Menard et al. 2000).

In this study, we initialize the model using prescribed vertical profiles as well as a 2D concentration field obtained from previous global analyses. Our model setup uses the approach of Olsen et al. (2013), which is derived from a nonlinear polynomial fit to global models and various types of measurements from 2003 to 2006. This initial field, taken as a first guess, is a function of latitude and altitude varying smoothly from the North Pole to the Equator with no temporal and longitudinal variation. The details of this preparation can be found in Xiong et al. (2008). Maintaining the shape of the vertical profiles, we rescale the initial field to our simulation period, April 2010, based on the annual/monthly increase in globally-averaged atmospheric methane (Dlugokencky 2022).

The bias may have different origins; here, it mainly depends on the initial field provided as input to the model since it does not necessarily represent the specific month of the simulation. With the aim to remove the potential biases in the model, we obtain accurate surface observations during the assimilation period from GLOBALVIEWplus CH<sub>4</sub>-ObsPack v3.0 (Schuldt et al. 2021) compiled by National Oceanic and Atmospheric Administration (NOAA) (Figure 5.3a). In our bias correction, we assume that a significant part of bias arises from the emissions at the surface, and the rate of change of concentration with height remains the same (shape of the vertical profile). In fact, we multiply the vertical profiles by a constant that depends on the discrepancy at the surface (we assume that vertical mixing in the troposphere is unchanged). Note that the model vertical coordinate

is extended up to the upper troposphere/lower stratosphere, and the profiles are adapted from the analysis of global methane studies (Xiong et al. 2008; Olsen et al. 2013). This assumption is made mainly due to the fact that other accurate observations, such as the Total Carbon Column Observing Network (TCCON), are used for the validation of our assimilation results (see Section 6.5).



**Figure 5.3.** Difference between (a) methane surface observations (GLOBALVIEWplus surface flask and tower, shown top left) and model equivalent values of concentration for (b) the initial guess (green dots/line), (c) rescaled initial to April 2010 but before bias correction (red dots/line), and (d) after bias correction (blue dots/line).

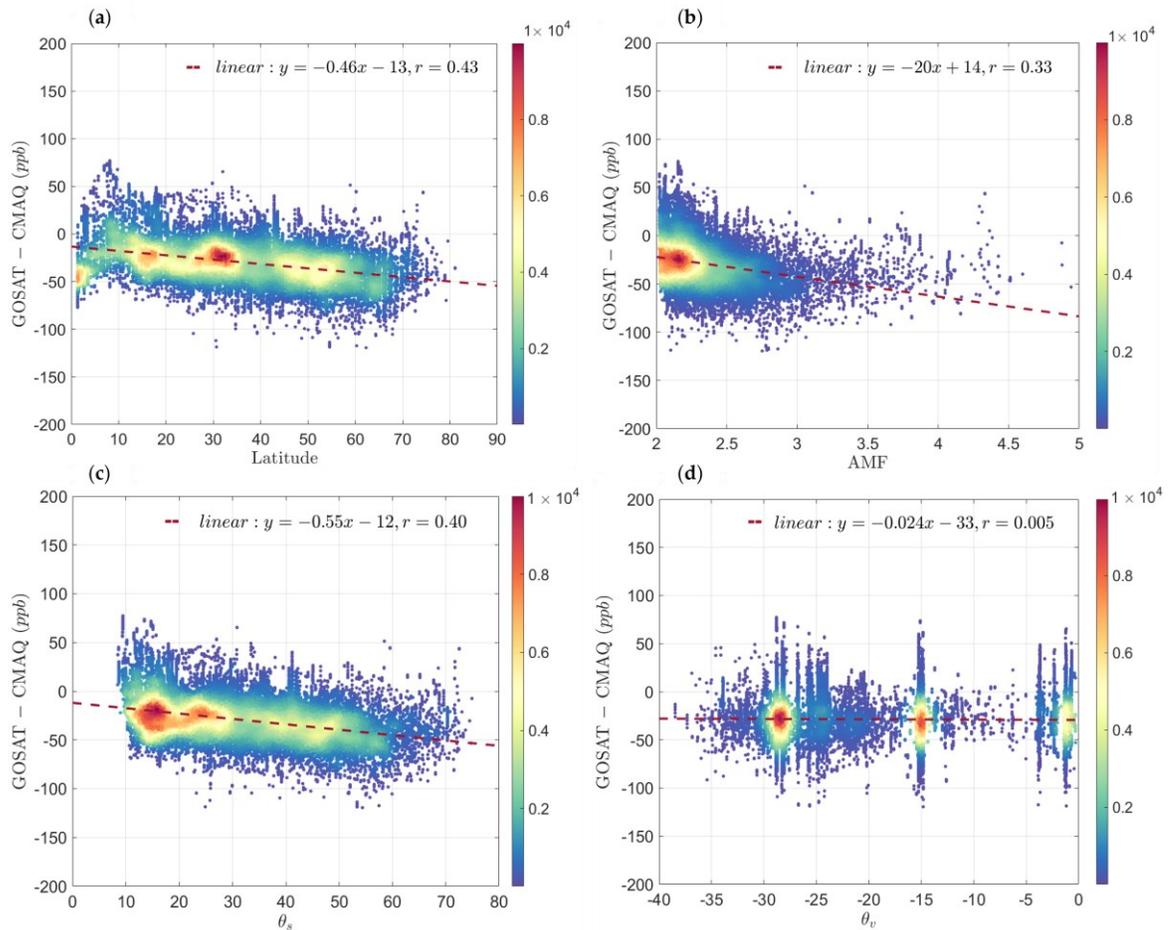
The bias removal involves adjusting the initial concentration field (i.e., initial guess) using a latitudinal linear regression model to match the surface observations (i.e., flasks and towers). Accordingly, first, the monthly averaged linear fit of Observation – Model against latitude using the first guess model initial conditions is obtained (Figure 5.3b–green dots). Next, the achieved fit is rescaled (with respect to globally averaged concentration from NOAA (Dlugokencky 2022)) to our simulation period in April 2010 (Figure 5.3c–red dots), and eventually, the fit is employed to remove the bias of the entire space of the initial conditions (Figure 5.3d–blue dots). It indicates that the latitude-based bias in the initial conditions is subtracted, resulting in a better (unbiased) agreement between the model and the surface observations. This could imply that the polynomial function (Olsen et al. 2013) deriving our first guess initial conditions is too smooth; thus, it tends to overpredict the lower latitude concentrations compared to the surface in-situ observations. In addition, the hemispheric absolute mean bias is decreased from 17.1 ppb for the rescaled initial to 4.3 ppb for the bias-corrected initial. Note that the negative values in Figure 5.3b–d correspond to the positively biased model forecasts (i.e., model overestimation). We consider the model that is initialized with the bias-corrected field as our nearly unbiased model and rely on this to remove the potential bias between GOSAT and H-CMAQ. An illustration of the bias correction of GOSAT with respect to H-CMAQ is provided in the following subsection. Note that the bias correction here is only applied for one month (i.e., April 2010) and may not be representative for another month or a longer period; thus, one needs to take into account the same type of bias correction for the time of assimilation.

## 5.4.2 Observation Bias Correction

GOSAT provides high accuracy retrieval data (~0.7% precision) with reasonable near-surface sensitivity and global coverage, making it a strong candidate for methane assimilation analysis (Butz et al. 2011; Buchwitz et al. 2015). The potential bias in GOSAT XCH<sub>4</sub> retrieval has been addressed previously by evaluating the data against other types of measurements (Wunch et al. 2011; Inoue et al. 2016; Zhou et al. 2016; Oshio et al. 2020). Both PR and FP versions of GOSAT data used here are post-processed and validated against surface-based Fourier transform infrared (FTRS) methane column abundance from TCCON. It was shown that the difference between XCH<sub>4</sub> retrieval and TCCON correlates with the albedo,  $\alpha$ , at 1.6 nm in band 2 (Buchwitz et al. 2017).

In addition to the retrieval bias, GOSAT can still have biases relative to atmospheric chemical transport models. This bias most likely originates from the emissions as well as the limitation of global models to realistically simulate the methane in the stratosphere, particularly at higher latitudes (Alexe et al. 2015; Saad et al. 2016; Maasakkers et al. 2019). For example, Turner et al. (2015) showed that their GEOS-Chem simulation of GOSAT features a positive latitudinal bias, where the mean Model – Observation is adjusted based on a fit to a quadratic regression function. Cressot et al. (2014) obtained a linear regression fit as a function of air mass factor (AMF) to correct for the biases in GOSAT with their chemistry-transport model, LMDz-SACS. Following them, we parametrize the bias in GOSAT with respect to our unbiased CMAQ simulation. First, the model is mapped to observation space using our observation operator ( $\mathbf{H}$ ) described in Section 5.2.2. Next, GOSAT – CMAQ is obtained as a function of retrieval parameters. Accordingly, a separate linear regression fit of GOSAT – CMAQ is computed for each of latitude, air mass factor

(AMF), solar zenith angle  $\theta_s$ , and satellite viewing zenith angle,  $\theta_v$  (Figure 5.4). Except for the  $\theta_v$  with no specific correlation pattern, the largest correlation of GOSAT – CMAQ is found with respect to latitude ( $r = 0.43$ ), followed by  $\theta_s$  ( $r = 0.40$ ) and AMF ( $r = 0.33$ ). All three fits show similar behaviour, indicating that the discrepancy is smoothly increasing toward the larger values of those parameters (Figure 5.4a–c).



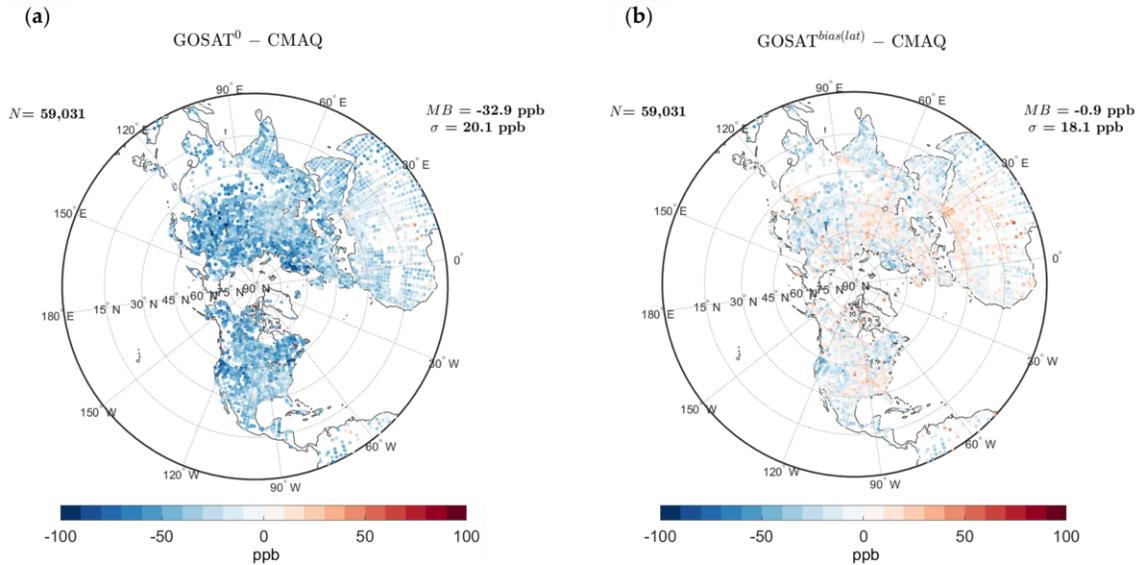
**Figure 5.4. Linear regression fit of the difference between GOSAT observation and H-CMAQ simulated XCH<sub>4</sub> as a function of (a) latitude, (b) air mass factor (AMF), (c) solar zenith angle,  $\theta_s$ , and (d) satellite viewing zenith angle,  $\theta_v$ . The colorbar indicates the number of GOSAT observations, with a higher observations density in red.**

Our latitude dependency of the bias agrees with Turner et al. (2015), but the AMF dependency disagrees with Cressot et al. (2014). Perhaps it is related to the fact that LMDz is a low-resolution general circulation model (GCM), whereas CMAQ and GEOS-Chem appear more as chemical transport models with higher resolution. The positive bias of CMAQ (Model > Observation) was found previously in other atmospheric chemistry models such as GEOS-Chem and addressed mainly as extratropical stratospheric bias due to the extra meridional stratospheric transport (Patra et al. 2011; Saad et al. 2016; Bader et al. 2017) as well as polar vortices (Zhang et al. 2021).

Figure 5.5 compares the Observation – Model before bias correction ( $\text{GOSAT}^0 - \text{CMAQ}$ ) and after bias correction with respect to latitude ( $\text{GOSAT}^{\text{bias}(\text{lat})} - \text{CMAQ}$ ). The hemispheric absolute mean bias ( $MB$ ) significantly decreased from 32.9 ppb to 0.9 ppb, while the residual standard deviation ( $\sigma$ ) turned out to be slightly smaller (20.1 ppb to 18.1 ppb) after bias correction. As expected from the correlation values in Figure 5.4, using AMF or  $\theta_s$  instead of latitude to remove the potential bias of GOSAT resulted in a less significant bias reduction because of a larger hemispheric  $\sigma$  and  $MB$ . Figure C.2 in Appendix C3 indicates a similar comparison using bias-corrected GOSAT data with respect to AMF and  $\theta_s$ .

In addition, in an attempt to maximize the information in the regression model, we adopted a simple multivariate linear regression (Seber and Lee 2012). However, we found that such an algorithm cannot considerably improve our prediction model (see Table C.2-Table C.4 in Appendix C3). This suggests that all three parameters are dependent and can be highly correlated to each other. Therefore, a bias-corrected GOSAT – CMAQ with only latitudinal adjustment could provide a reasonable bias correction to our data. After bias

correction, the model discrepancy with observations can be primarily random and could be attributed to the random emission errors, the correlation lengths  $(L_h, L_v)$ , and random errors in the observations  $(\mathcal{E}^o)$ , initial  $(\mathcal{E}^i)$ , and forward model  $(\mathcal{E}^q)$ , for which we provide a description in Section 5.4.4.



**Figure 5.5.** (a) Difference between observations and model before bias correction ( $\text{GOSAT}^0 - \text{CMAQ}$ ) and (b) after bias correction with respect to latitude ( $\text{GOSAT}^{\text{bias}(\text{lat})} - \text{CMAQ}$ ) over a month with the number of observations  $N = 59,031$ . A bias-corrected field of residuals calculated from the regression with respect to AMF and  $\theta_s$  is provided in Figure C.2.

### 5.4.3 Construction of Spatial Correlation Functions on the H-CMAQ Grid

The H-CMAQ grid is uniform on the projected plane. Nearly all spatial correlation models rely on an underlying uniform and isotropic grid representation. This is what most variational assimilation systems take into account, and further, assume either an infinite or periodic domain. Here, we will explain how to construct spatial error correlations on the H-CMAQ grid while representing an underlying homogeneous isotropic correlation function. The procedure consists in constructing a homogeneous isotropic and periodic

correlation function on the surface of a sphere that is independent of the grid, and then mapping it onto the H-CMAQ grid.

The construction of homogeneous isotropic correlation functions on the surface of a sphere has been developed by Gaspari and Cohn (1999). Background information on the method is also given in Menard (2000). A correlation function is a function of a pair of points  $(i, j)$ . Consider a sphere of radius  $a$ , with position vectors (from the centre of the sphere)  $\mathbf{r}_i, \mathbf{r}_j$  for the points  $i$  and  $j$ . We define a Cartesian coordinate system in  $\mathbb{R}^3$ ,  $(x^C, y^C, z^C)$ , where we use the superscript  $C$  to recall the Cartesian space. A point  $i$  (or  $j$ ) on the surface of the sphere that has (in a spherical coordinate system) a longitude  $\varphi_i$  and colatitude (or inclination)  $\theta_i$  is related to the Cartesian coordinates as follows

$$\begin{aligned} x_i^C &= a \cos \varphi_i \sin \theta_i \\ y_i^C &= a \sin \varphi_i \sin \theta_i, \\ z_i^C &= a \cos \theta_i \end{aligned} \quad (5.36)$$

and similarly for the grid point  $j$ . The chordal distance  $D_{ij}$  between the position vectors  $\mathbf{r}_i = (x_i^C, y_i^C, z_i^C)$  and  $\mathbf{r}_j = (x_j^C, y_j^C, z_j^C)$  is given by

$$D_{ij} = \|\mathbf{r}_i - \mathbf{r}_j\| = \sqrt{2} a \left\{ 1 - \left[ \sin \theta_i \sin \theta_j \cos(\varphi_i - \varphi_j) + \cos \theta_i \cos \theta_j \right] \right\}^{1/2}, \quad (5.37)$$

Using a chordal distance to define a “distance” for a correlation function on a sphere has the double advantage that a correlation model originally defined in  $\mathbb{R}$  can be used in  $\mathbb{R}^3$  to define a correlation on a sphere and has the property of being periodic (Gaspari and Cohn 1999).

Three models have been considered in this study: the Gaussian (or double exponential) model,

$$C_G(i, j) = \exp\left(-\frac{D_{ij}^2}{2L_c^2}\right), \quad (5.38)$$

the First-Order-Auto-Regressive (FOAR), or exponential model,

$$C_{FOAR}(i, j) = \exp\left(-\frac{D_{ij}}{L_c^*}\right) \text{ where } L_c^* = L_c / (0.5005), \quad (5.39)$$

and the Second-Order-Auto-Regressive (SOAR) model

$$C_{SOAR}(i, j) = \left(1 + \frac{D_{ij}}{L_c^{**}}\right) \exp\left(-\frac{D_{ij}}{L_c^{**}}\right) \text{ where } L_c^{**} = L_c / (1.3494). \quad (5.40)$$

Note that  $L_c$  denotes a correlation length, defined from the curvature at the origin, while  $L_c^*$  or  $L_c^{**}$  is a model parameter adapted from Ménard et al. (2016). We recommend always using  $L_c$ , as it is a physically meaningful quantity.

The second step consists of mapping the H-CMAQ grid onto the corresponding point on the surface of the sphere (see Figure C.1 in Appendix C1 for information on the polar stereographic projection used in H-CMAQ). As mentioned in Gaspari and Cohn (1999), if there exists a one-to-one transformation,

$$T(\varphi, \theta) = (x^p, y^p), \quad (5.41)$$

(see Equation (C.1)) then we can define a correlation function with respect to the  $(x^p, y^p)$  coordinates. The computation then goes as follows

$$\begin{Bmatrix} (x_i^p, y_i^p) \\ (x_j^p, y_j^p) \end{Bmatrix} \xrightarrow{T^{-1}} \begin{Bmatrix} (\varphi_i, \theta_i) \\ (\varphi_j, \theta_j) \end{Bmatrix} \rightarrow D_{ij} \rightarrow C(D_{ij}), \quad (5.42)$$

where  $C$  is one of the correlation models above (Equations (5.38)–(5.40)).

#### 5.4.4 Observation, Model and Initial Error Covariance Modelling

As in any assimilation system, the input error covariances need to be specified. In this section, we describe the modelling and assumptions of those error covariances. It is generally assumed that observation errors, model errors, and initial errors are uncorrelated, and furthermore, that the observation and model errors are serially (temporally) uncorrelated. These are the standard assumptions used to derive the Kalman filter. We should note that because of the dynamics and the interplay with the analysis, as the forecast error becomes realistic, it also becomes correlated in space and time and with past observation errors (this is a standard result in Kalman filtering theory; see, for example, Jazwinski (1970)).

Let us assume that the observation error  $\mathbf{R}$  and the model error  $\mathbf{Q}$  are both spatially uncorrelated. The observation error has two components: the measurement error ( $\varepsilon^m$ ) provided by the instrument team and the representativeness error ( $\varepsilon^r$ ) arising from the mismatch between the subgrid-scale represented in the observation and the gridded model quantity (Cohn 1997; Janjic et al. 2018). We assume that the observation error covariance is diagonal and takes the form of  $\mathbf{R} = (\varepsilon^o)^2 \mathbf{I}$ .

Many studies estimate the representativeness error as an additive source of error to the measurement error (Menard and Changs 2000; Menard et al. 2000; Heald et al. 2004; Berchet et al. 2013; Szenasi et al. 2021). However, we tune the overall observation error by considering a multiplicative correction factor applied to the measurement error as denoted in Equation (5.43). A multiplicative factor is consistent with other types of errors (i.e., modelling and initial error) used in this study. In fact, it is used in the majority of data

assimilation papers (Menard et al. 2000; Segers et al. 2005). Therefore, the representativeness error is included as part of the correction factor  $f^o$ .

$$\mathbf{R}' = (f^o \varepsilon^m)^2 \mathbf{I}, \quad (5.43)$$

where  $\mathbf{R}'$  denotes the observation error covariance after tuning. Note that initially, no representativeness error is accounted for  $\mathbf{R}'$ .

The model error ( $\varepsilon^q$ ) is assumed to be proportional to the analysis (Menard et al. 2000). We simply assume a uniform accumulation of model error in time, where it becomes almost equivalent to a particular time average of the analysis,  $\bar{X}^a(\mathbf{x}, t)$ , (i.e., monthly averaged analysis is used in this study). Model error covariance is considered as a diagonal matrix,  $\mathbf{Q} = (\varepsilon^q)^2 \mathbf{I}$  (Tremolet 2006; Stanevich et al. 2020), and a relative model error standard deviation,  $f^q$ , is defined for tuning  $\mathbf{Q}$ , which has the form

$$\mathbf{Q}' = (f^q X^a(\mathbf{x}, t))^2 \mathbf{I}, \quad (5.44)$$

where  $\mathbf{Q}'$  denotes the model error covariance after tuning. Note that model error variance,  $q$ , in Equation (5.9) is equivalent to the diagonal elements of  $\mathbf{Q}'$ .

We attempt to tune the initial concentration error ( $\varepsilon^i$ ) separately, although it is often integrated as part of the modelling error. The initial error covariance matrix is also assumed to be diagonal. We assume that the initial error before tuning is about 5% of the initial concentration (i.e.,  $\varepsilon^i = 0.05 X_0^f$ ) with the same spatial distribution. By conducting a similar analysis as in Section 5.4.2 on the first day of the simulation, we found that the mean standard deviation of the residual (Observation – Model) is relatively smaller ( $\sim 20 \text{ppb} \equiv 1.2\%$ ). The parametric form of the tuned initial error covariance is

$$\mathbf{P}_0' = (f^i \boldsymbol{\varepsilon}^i)^2 \mathbf{I}, \quad (5.45)$$

where  $\mathbf{P}_0'$  represents the initial error covariance after tuning.

## 5.5 Verification of the Basic Properties of the Assimilation System

### 5.5.1 One-Observation Experiment

Here, we conduct a one-observation experiment using a single simulated observation. It is a standard experiment to verify that the mechanics of the assimilation system is working and can also be used to verify the validity of some assumptions in the formulation (Gauthier et al. 1999; Lahoz and Schneider 2014). The one-observation experiment can also be used to compare different assimilation systems (Buchner et al. 2010). When a single observation is assimilated, the analysis increment has the same spread in space as the spatial correlation function—a property that can be derived directly from the analysis equation and Kalman gain.

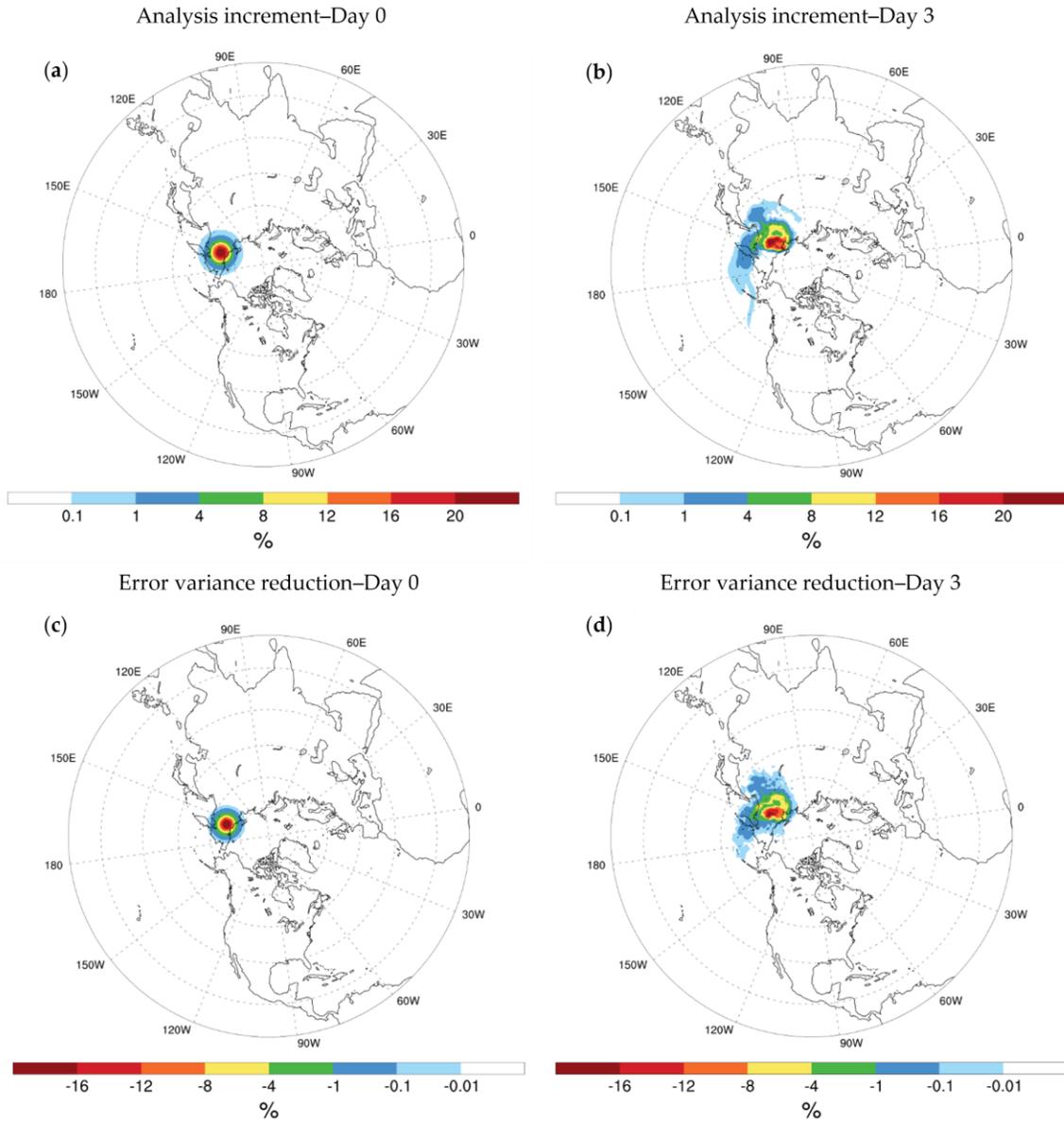
With the one-observation experiment, first, we will verify that our formulation of error covariances (in  $\mathbb{R}^3 \times \mathbb{R}^3$  and projected onto the polar stereographic grid of H-CMAQ) does yield a homogenous isotropic correlation. Secondly, we will show that our assumption to use only advection to propagate the error variance is adequate for assimilation on a time scale relevant to GOSAT assimilation. Errors in the wind, emissions, and any errors not accounted for by our simple advection of error variance, such as horizontal and vertical diffusion effect on error variance, will contribute to the model error.

The assimilation system is examined with a single simulated observation and arbitrary error parameters, including  $f^i = 0.05$ ,  $f^q = 0.015$ ,  $f^o = 0.5$ . Note that the initial concentration is equivalent to the bias-corrected concentration field (see Section 5.4.1), and

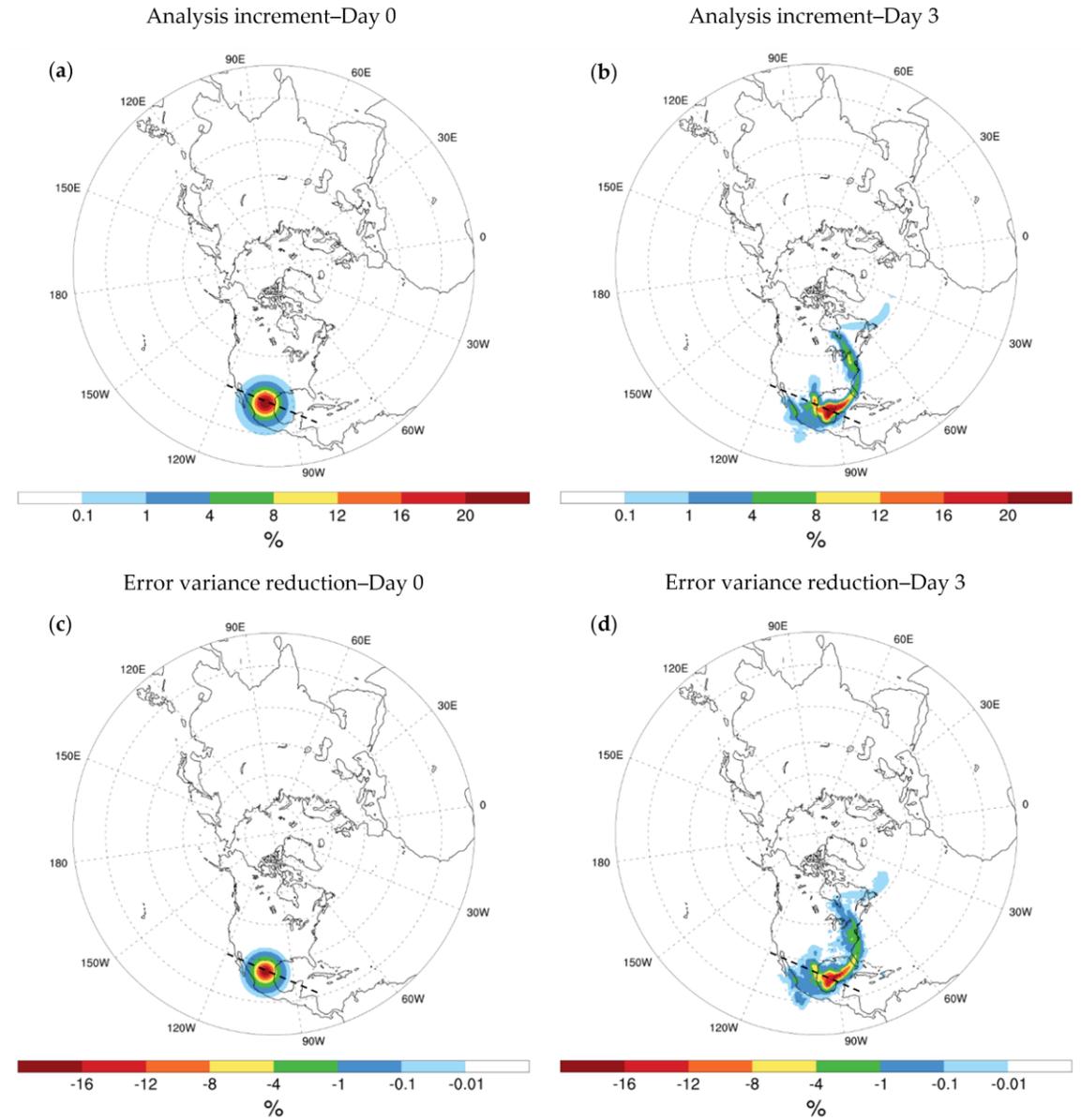
initial variance is obtained from Equation (5.45) where  $\varepsilon^i = 0.05 X_0^f$ . These quantities are kept the same for all the cases in this experiment. In this experiment, we consider the SOAR correlation model with a horizontal correlation length of 500 km and 10 model vertical layers (starting from the surface) of the sigma-pressure coordinate (referred to  $10 \sigma_l$  length scale). Figure 5.6 and Figure 5.7 show two single-observation experiments—one with an observation in the high latitude ( $71^\circ \text{ N}, 155^\circ \text{ E}$ ) and the other closer to the Equator ( $27^\circ \text{ N}, 105^\circ \text{ W}$ ). In both experiments, the assimilation started at Day 0 with a single synthetic observation. The synthetic observations represent column-averaged concentrations (Equation (5.1)) generated with the model while accounting for the GOSAT averaging kernels and a priori. To define the magnitude of the synthetic observations, we assume a multiplicative factor of 1.4 (40% higher) for the model column-averaged concentrations. The assimilation consists of propagating the analysis and its error variance for a duration of 3 days after the single observation is taken into account. We only examine 3 days, which corresponds to the GOSAT revisit time, that is, the time after which a new batch of observations becomes available for the same region.

For the observation in the high latitude (Figure 5.6a–d), the analysis increment and the error variance reduction are shown near the surface, whereas for the observation in the lower latitude (Figure 5.7a–d), they are demonstrated at about 600 hPa (~layer 23). By comparing Figure 5.6a with Figure 5.7a, we note that the error correlation, captured by the analysis increment, is indeed homogeneous and isotropic. The apparent increase in radius of the spatial spread in lower latitudes is due to the polar stereographic projection (we verified that the length scale is, in fact, the same). By comparing the propagated analysis increment with the propagated error variance reduction (i.e., Figure 5.6b,d, and Figure

5.7b,d), we found that the transport of the analysis increment and the reduction of error variance are quite similar in terms of their spread and spatial distribution.



**Figure 5.6.** The first one-observation experiment near the surface at higher latitude ( $71^{\circ}$  N,  $155^{\circ}$  E) to show (a) the analysis increment at Day 0 and (b) Day 3; and (c) the error variance reduction at Day 0, and (d) Day 3.



**Figure 5.7.** The second one-observation experiment at about 600 hPa and lower latitude ( $27^{\circ}$  N,  $105^{\circ}$  W) to show (a) the analysis increment at Day 0 and (b) Day 3; and (c) the error variance reduction at Day 0, and (d) Day 3. The dashed (black) line represents a cross-section along the vertical direction, where the vertical distribution of the analysis and its error variance are demonstrated in Figure 5.8.

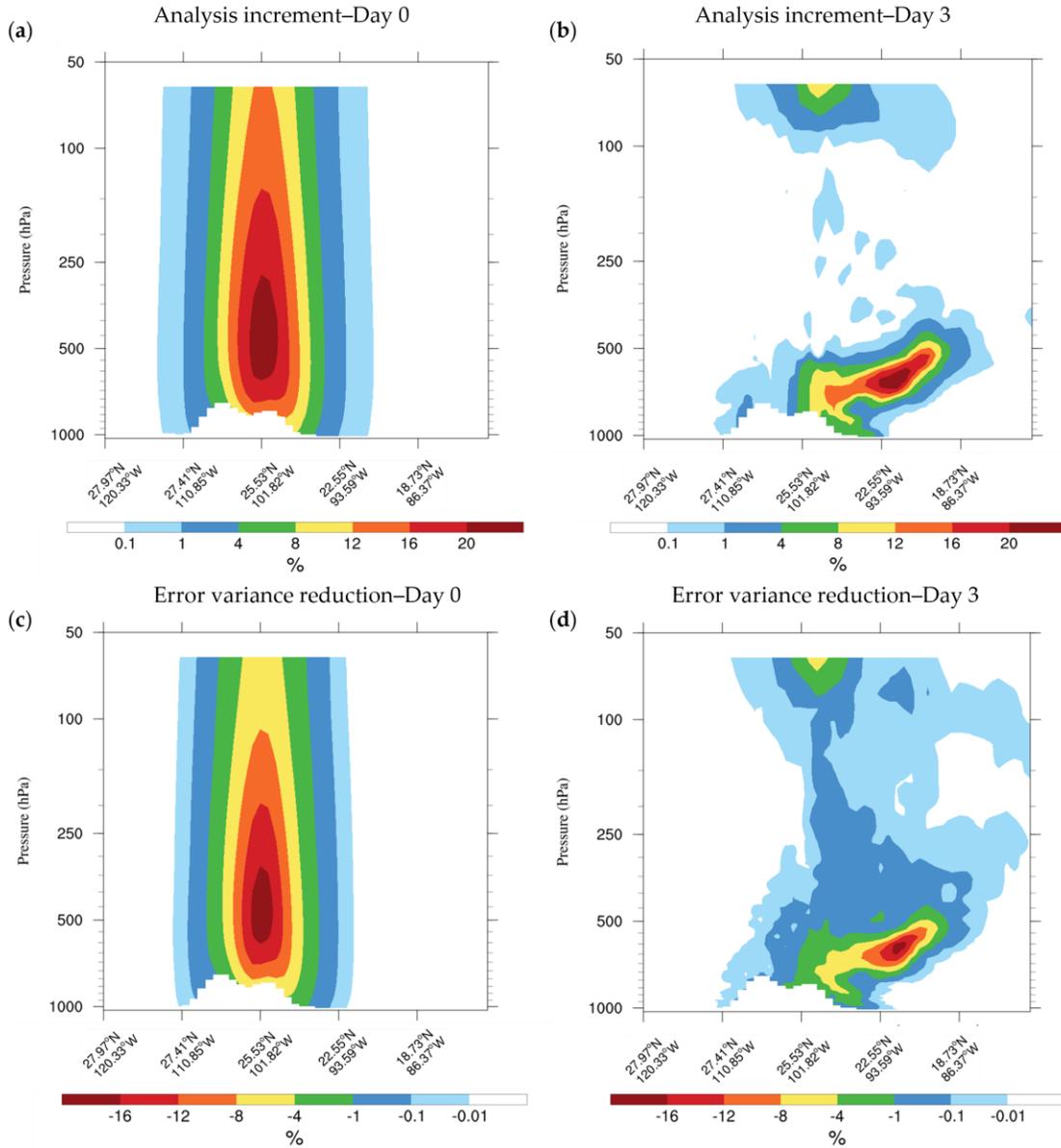
This suggests that the propagation of the error variance using advection only is a reasonable assumption to make, since it maintains a similar structure as the methane fields. Furthermore, we found a comparable weight of the analysis error variance at the start day (Day 0) and after 3 days; in particular, we note that the maximum reduction of error

variance has not changed more than 5% after 3 days (i.e.,  $\frac{\left(\left(\sigma_{(Day\ 3)}^a\right)^2 - \left(\sigma_{(Day\ 3)}^f\right)^2\right)_{\max}}{\left(\left(\sigma_{(Day\ 0)}^a\right)^2 - \left(\sigma_{(Day\ 0)}^f\right)^2\right)_{\max}} \sim 1$ ).

This result contrasts with the variance losses that are known to occur in Kalman filter and ensemble Kalman filter systems.

In another series of experiments, we examine the contribution of error correlation to address the vertical estimation properties, especially the quantities at the surface. To illustrate the propagation in the vertical direction, we obtain a vertical cross-section (Figure 5.7a-d) passing through the observation at Day 0, starting from (28° N, 120° W) to (15° N, 80° W). The observation is at the center of the analysis increment and center of the error variance reduction (Figure 5.7a,c), also called increment/reduction. Figure 5.8a,c illustrates the increment/reduction at the moment when a single observation is assimilated. Several factors, such as the averaging kernel, the model layer pressure weight ( $\omega$ ), and the correlation length scale based on the SOAR correlation model, define the increment/reduction patterns of the analysis and error variance. A relatively large correlation length ( $L_v = 10\sigma_l$ ) results in a stretched pattern of increment/reduction along the vertical direction. The vertical distribution also indicates that they are mainly influenced by  $\omega$  (Figure 5.9c), resulting in a maximum increment/reduction at mid-troposphere (~600 hPa, or 23<sup>rd</sup> layer). On Day 3, these quantities are shown on the same

plane. The largest increments/reductions are shifted toward the southeast of the initial point (Figure 5.7d), mainly due to advective transport.



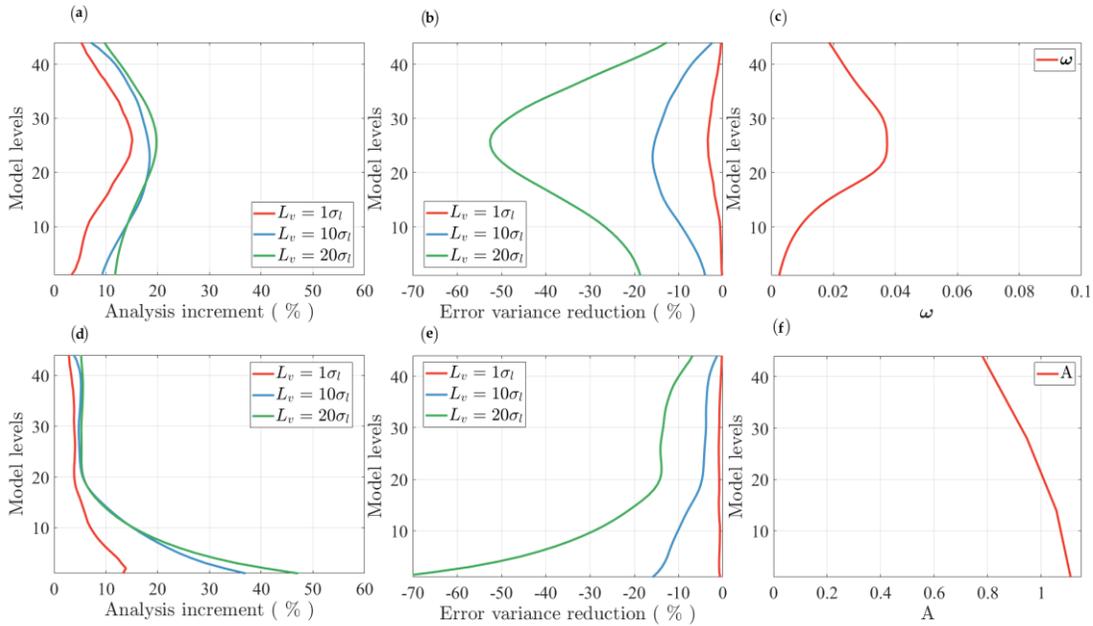
**Figure 5.8.** Vertical distribution of (a) analysis increment at Day 0 and (b) Day 3 along the cross-section (black dashed line) shown in Figure 5.7a,b; Vertical distribution of (c) error variance reduction at Day 0 and (d) Day 3 along the cross-section (black dashed line) shown in Figure 5.7c,d. The cross-section starts from (28° N, 120° W) to (15° N, 80° W).

The large distension is mainly influenced by the wind pattern, while the diffusion improves the smoothness of the field, especially in lower levels (Figure 5.6b,d). We should note that the higher increment/reduction to the lower troposphere and near-surface enhances the ability of the assimilation in constraining the surface quantities such as emissions within a short period of assimilation integration (i.e., 3 days).

Figure 5.8b,d indicate that there is reasonable consistency between the propagation of analysis increment and variance reduction in the vertical direction. It shows that the vertical effects from a single observation can persist for at least 3 days. Furthermore, a significant part of analysis increment and error variance reduction remains in the lower elevation with a downward shift of the maximum values. These increments/reductions are quite small between the mid to upper troposphere, except for another slight increase at the upper layers, which occurs likely due to a zero-flux assumption across the model top boundary (Byun and Schere 2006; CMAQv5.3 user's guide 2019). Hence, similar to the surface, part of the increments/reductions lingers near the upper layers.

We examine (Figure 5.9) the impact on the analysis increments and error variance reductions for different vertical correlation lengths ( $L_v = 1\sigma_l, 10\sigma_l, 20\sigma_l$ ). It indicates that the error variance reduction is significantly more sensitive than the analysis increment to the vertical correlation length scale. As mentioned earlier, the layer pressure weights normalize the adjustments and allocate more impact on the mid-layers. We remove this effect by considering uniform layer pressure weights, which results in emphasizing the influence of the averaging kernels on the vertical profile of the analysis increment and error variance reduction during the analysis step (Figure 5.9a,b,d,e). Thus, higher sensitivity of GOSAT retrieval to the surface associated with its average kernel (Figure 5.9f) results in a

similar effect on the analysis (Figure 5.9d). Likewise, the error variance, whose vertical sensitivity is more substantial than the analysis, exhibits a larger influence from the surface (Figure 5.9e). It implies that for observations with the highest retrieval sensitivity to the near-surface (e.g., GOSAT), integrating the reduction of error variance with the assimilation scheme enhances our ability to retrieve information from the surface.



**Figure 5.9.** (a) Analysis increment and (b) error variance reduction during assimilation step for different vertical correlation length scales (red:  $L_v = 1\sigma_l$ , blue:  $L_v = 10\sigma_l$ , and green:  $L_v = 20\sigma_l$ ); (c) Pressure weight ( $\omega$ ) at each model layer; (d) analysis increment and (e) error variance reduction where the pressure weight is considered uniform; (f) GOSAT column averaging kernel.

For a single-observation problem, one can theoretically show that the error variance reduction maintains a quadratic relationship with the analysis increments (Equation (5.23)). Therefore, besides the continuous estimate of the analysis and its uncertainty, the distinctive feature of PvKF in explicitly computing the error variance can be encouraging to infer surface quantities (i.e., emission inversion).

### 5.5.2 Timing (Computational Efficiency)

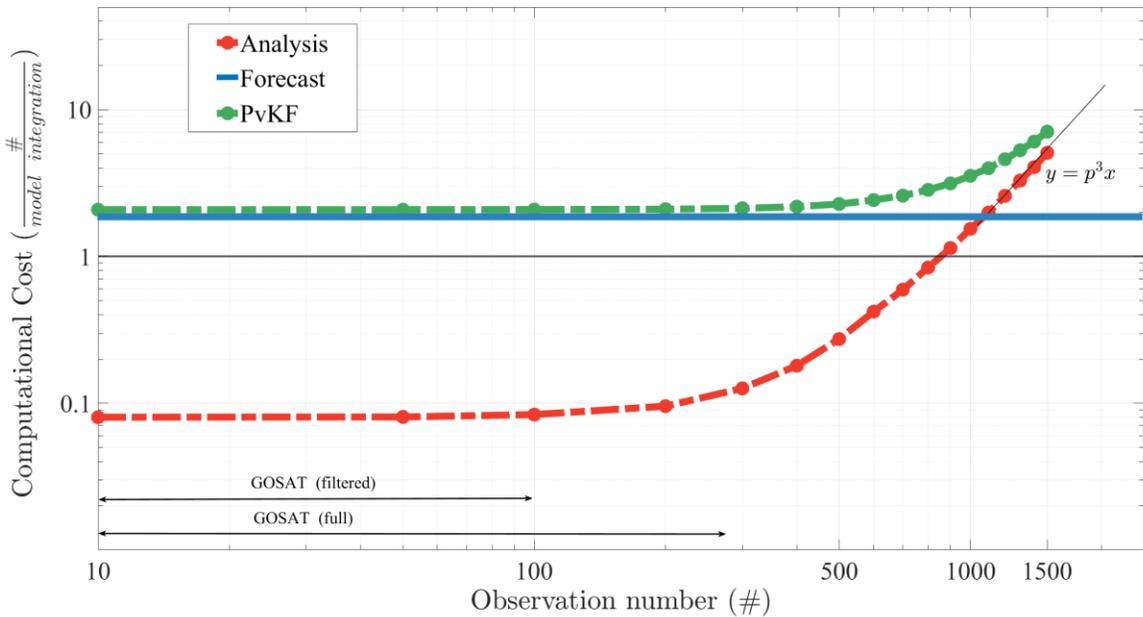
The computational cost of advanced assimilation schemes is often compared against the cost of one model integration. The cost of the forecast step in PvKF involves two model simulations: the forecast of concentration and the forecast of the error variance. Thus, the forecast step of PvKF requires a little less than twice the computational time of the H-CMAQ forecast simulation, since the chemistry and diffusion are turned off for the propagation of error variance. When observations and analyses are included, the total computational cost depends on how many observations are assimilated per time step. The computational time required to perform the analysis step increases rapidly as the number of observations increases, due to the inversion of the innovation covariance matrix (or solving the equivalent linear problem) and the computation of the analysis error variance. Note that, contrary to 4D-Var, the storage space of PvKF always remains the same for any length of integration.

The computational cost of assimilation with real GOSAT data (developed by SRON/KIT) should also take into account the number of actual quality-controlled observations. Obtaining quality-controlled observations consists of several steps. In addition to using the quality-controlled observations provided by the instrument team, we apply two additional filters on real GOSAT data before including them in the assimilation. First, we remove outliers whose departure from the global mean of the observations is three times larger than its standard deviation, and second, we conduct a thinning process on the observations with the aim of providing uncorrelated observation errors (Menard and Deshaies-Jacques 2018a) (see Section 6.3 for more details). Accordingly, the filtered GOSAT data holds below 100 retrievals per hour, which is about one-third of observations

before filtering (i.e., full GOSAT data). Note that the quality control steps are performed as pre-processing steps; hence, they do not impact the computational cost of assimilation.

In order to make an assessment of the computational cost of the PvKF algorithm, we investigate the assimilation of an arbitrary number of observations per hour. For this, we have an arbitrary number of simulated observations that imitate the behaviour and influence of real GOSAT observations. All the simulated observations are retrieval products and have averaging kernels and a priori columns similar to the retrieved XCH<sub>4</sub>, and are basically proxy retrievals (i.e., PR) of GOSAT observations developed by SRON/KIT. The advantage of this approach is that we can generate an arbitrary number of “retrieved” observations. Therefore, the simulated observations could be generated at other times than for the retrievals. Both model (H-CMAQ) and analysis are carried out for a one-hour simulation of size  $187 \times 187 \times 44$  in the computational domain with 108 km resolution. Figure 5.10 compares the computational time of the analysis step (dashed red line) as normalized to one model integration (solid black line). Contrary to the PvKF forecast (solid blue line), the analysis cost largely varies depending on the number of assimilated observations, attaining an equivalent cost to one model integration at about 850 observations. The total assimilation time of PvKF (dashed green line) is obtained by adding the analysis time with approximately twice the model running time. For the number of observations equivalent to the real GOSAT data (both filtered and full data), the analysis step is much shorter than the model (i.e.,  $\sim 0.1$  per model integration). In this case, matrix inversion of the innovation covariance matrix (i.e.,  $\mathbf{\Gamma}^{-1} = (\mathbf{HP}^f \mathbf{H}^T + \mathbf{R})^{-1}$  in Equations (5.17) and (5.20) uses Cholesky decomposition (Krishnamoorthy and Menon 2013) to solve either the inversion of the innovation covariance matrix or a system of linear

equations  $\Gamma b = d$ . Thus, our hourly assimilation scheme with real GOSAT data maintains approximately twice the H-CMAQ computational time. Nonetheless, for the higher number of measurements (e.g., >500) per hour generated synthetically with GOSAT, our computational time estimate increases exponentially to near  $O(p^3)$ , which is comparable to the order of a Cholesky decomposition to compute the matrix inversion (i.e., it is dominated by the analysis step computation).



**Figure 5.10.** 1-h computational time of the analysis (dashed red line) with respect to the number of observations within the PvKF assimilation. The solid black line shows the computational time of a 1-h H-CMAQ simulation with a known computational configuration. The solid blue line denotes the forecast step of PvKF that includes two forecasts, one for methane transport and the other for the advection of error variance. The dashed green line represents the total computational time expected for the PvKF.

As a remark, we can roughly compare the computational efficiency of the PvKF assimilation described above with other popular schemes such as 4D-Var and EnKF. For example, 4D-Var entails the adjoint of CMAQ where the computation of the backward

mode requires at least twice (and up to three times) the time as the forward simulation (Hakami et al. 2007). In addition, several forward-backward iterations are required until the algorithm's convergence. The adjoint computation also requires a relatively large (~10 times the model) amount of storage space (Zhao et al. 2020). The computational time of EnKF assimilation also highly depends on the number of ensembles carried out with the model that adopts it. In general, dozens of ensembles are expected to maintain a reliable assimilation (Houtekamer and Zhang 2016); for example, Peng et al. (2015) used 48 ensembles in their CFI-CMAQ. In summary, PvKF assimilation can achieve a high level of computational efficiency compared to the other typical assimilation schemes. We demonstrate the performance of this framework using real GOSAT observations in Chapter 6.

## **5.6 Summary and Conclusions**

We have designed an assimilation system based on the parametric (variance only) Kalman filter, or PvKF, with the hemispheric CMAQ model and GOSAT methane observations (the assimilation scheme can be used for any long-lived species). The scheme is capable of providing analysis together with its uncertainty (i.e., error variance) while being computationally cheaper than other popular data assimilation schemes such as 4D-Var and EnKF. The analysis is derived sequentially with a minimum hourly batch of observations that could maintain real-time assimilation. The uncertainty is obtained by advection of error variance using the H-CMAQ model with a predefined error correlation model. The assimilation system was tested with a single simulated observation experiment to verify basic properties, as well as the conservation of information and the

appropriateness of using advection of error variance. This scheme does not assume a perfect model, nor does it require the development of adjoint or ensemble simulations.

To conduct PvKF methane assimilation, we modified H-CMAQ to include methane transport, chemical reaction with OH, and emissions from both anthropogenic (EDGAR v6) and natural (WetCHARTs v1.0) source categories. Preparation of an unbiased initial field and addressing the bias in GOSAT with respect to independent surface measurements (GLOBALVIEWplus CH4-ObsPack v3.0 compiled by NOAA) is demonstrated. We found that a latitudinal correction provides a reasonable bias correction to our data. Thus, the model discrepancy with observations can be primarily attributed to the random errors in the model, observations, and emissions.

Results using simulated observations indicate that the expected behaviour of the analysis error variance and analysis increment, which is derived from the model, is fairly consistent, while the information content (i.e., total variance) is conserved. Note that this may not be the case for EnKF if inflation is not added. We also demonstrate that the effect of a single observation can persist within a period of the GOSAT revisiting time, which is about 3 days. In addition, it is shown that the vertical error correlation could assist in deducing quantities at or close to the surface.

We also discussed the computational cost of the PvKF assimilation against an arbitrary number of observations per time step. We found that the assimilation scheme with GOSAT maintains approximately twice the H-CMAQ computational time, while a larger number of observations per hour (e.g., >500), which is not the case of GOSAT, suggests an exponential increase in the computational time. Nonetheless, with a couple of thousands of observations per hour, the method still performs acceptably but is slower. We emphasize

that this assimilation system does not assume a perfect model, in agreement with the fact that the wind and emissions are not entirely known. Yet, PvKF is computationally advantageous compared to 4D-Var and EnKF.

The main limitation of this method is related to the lifetime of the species. PvKF is well-adapted to long-lived species, such as methane, and still applies to shorter lifetime chemical species, but with the caveat that a smaller fraction of the total forecast error variance is explained by the advection of error variance. In this case, the residual error variance (i.e., unexplained error variance) is captured by the stationary model error. Thus, for chemical species with a shorter lifetime, the PvKF more resembles an OI (Optimal Interpolation) scheme. Another limitation is that the framework's feasibility depends on the observation characteristics (e.g., observation number and density). The larger number of observations we assimilate, the more accurate analysis we may obtain. However, the assimilation scheme is limited to a certain number of observations due to the computational capacity of PvKF, as explained in Section 5.5.2. In addition, increasing the number of observations can result in a higher spatial density, which increases the error correlation in observation space. This contradicts the necessary condition to obtain an optimal PvKF analysis (see Sections 6.2 and 6.3). Therefore, the number of observations may limit both the efficiency and the quality of the assimilation.

In general, we found that the PvKF algorithm is sufficiently adaptable as a lightweight scheme for carrying long-lived tracers inside a chemistry-enabled atmospheric model. In Chapter 6, we will discuss this framework with the objective of obtaining optimal assimilation of GOSAT and deducing realistic statistics for transport, observations, and model parameters, including correlation lengths.

A potential application of this algorithm, which has not yet been pursued in this study, is to improve the inverse modelling of methane emissions on a highly resolved regional domain by integrating an accurate initial and the inflows of methane concentrations at the lateral boundaries of the regional model and their uncertainties.

## **Chapter 6: Assimilation of GOSAT Methane in the Hemispheric**

### **CMAQ; Part II: Results Using Optimal Error Statistics**

In Chapter 6, the parametric variance Kalman filter (PvKF) data assimilation designed in Chapter 5 is applied to GOSAT methane observations with the hemispheric version of CMAQ to obtain the methane field (i.e., optimized analysis) with its error variance. Although the Kalman filter computes error covariances, the optimality depends on how these covariances reflect the true error statistics. To achieve a more accurate representation, we optimize the global variance parameters, including correlation length scales and observation errors, based on a cross-validation cost function. The model and the initial error are then estimated according to the normalized variance matching diagnostic, also to maintain a stable analysis error variance over time. The assimilation results in April 2010 are validated against independent surface and aircraft observations. The statistics of the comparison of the model and analysis show a meaningful improvement against all four types of available observations. Having the advantage of continuous assimilation, we showed that the analysis also aims at pursuing the temporal variation of independent measurements, as opposed to the model. Finally, the performance of the PvKF assimilation in capturing the spatial structure of bias and uncertainty reduction across the Northern Hemisphere is examined, indicating the capability of analysis in addressing those biases originated, whether from inaccurate emissions or modelling error.

#### **6.1 Introduction**

In Chapter 5, we have developed a parametric variance Kalman filter (PvKF) data assimilation system of atmospheric methane using GOSAT observations and the hemispheric CMAQ (H-CMAQ) model. The formulation of the assimilation system and

its verification using synthetic observations were made in Section 5.5, demonstrating that PvKF maintains the information content and does not rely on a perfect model assumption. This scheme computes the assimilation error variance and compared with 4D-Var and ensemble Kalman filtering capable of computing also the error variance, it is computationally advantageous.

Furthermore, the method appears to be well-adapted for long-lived species such as methane and performs efficiently with a small number of observations, as is the case with GOSAT (i.e., < 300 retrieval/hour). In this Part II of the study, we employ the assimilation scheme to real GOSAT observations with the objective of producing high-quality (i.e., near-optimal) analysis by optimal estimation of the error covariance input parameters (e.g., model error variance, observation error variance, and background error correlation lengths). The high-quality analysis of PvKF offers the same spatiotemporal resolution as the model both for the optimal state estimate and for its uncertainty, expressed as an error variance.

GOSAT observations have been used for a decade to constrain methane emissions using a variety of inverse modelling techniques. Still, significant differences between several studies' results have been reported, even those using a similar dataset (Ganesan et al. 2019; Miller et al. 2019), implying that the inverse analyses, both on a regional and global scale, are faced with significant challenges. On the global scale, the challenge arises mainly from unaccounted uncertainties of all major sinks of methane, e.g., those resulting from OH oxidation, soil uptake, and stratospheric loss (Turner et al. 2017; Turner et al. 2018; Turner et al. 2019; Wang et al. 2019; Maasakkers et al. 2019; Saunio et al. 2020; Zhao et al. 2020b). On the regional scale, the challenge is primarily due to inaccurate lateral

boundaries and initial conditions, which are more influential than the unaccounted uncertainties at the global scale due to the short residence time of air (e.g., several weeks) in a limited regional domain (Jacob et al. 2016). However, it is assumed in inverse modelling studies such as Turner et al. (2015) and (Bergamaschi et al. 2018) that the model forecast of methane is perfect, and the observations and background error statistics are precisely known. Hence, it implies that the only source of error arises from inaccurate emissions. In a comparable study, Stanevich et al. (2020; 2021) account for model transport error under a weak-constraint 4D-Var inversion, which partly addresses the uncertainties due to lateral boundary inflow and the initial concentration. However, those uncertainties may not be associated with optimal error statistics, which influence the quality of the analysis.

In this chapter, rather than solely correcting the emissions, we consider emission errors as part of the modelling error. We recall that emissions are the model input (or parameter), and their errors can be considered separately or as part of the CMAQ model error. We use the GOSAT observations to estimate the methane concentrations using an assimilation scheme that does allow for (chemical transport) model (random) errors. The objective of data assimilation is to obtain the best representation of methane concentration. In addition, because the estimated variables and the observed variables are essentially the same (with the difference that the observation here is a vertically integrated quantity), diagnostics of the assimilation residuals are useful in determining error covariances such as the observation and the forecast model error covariance. These error covariances are essential inputs to the PvKF and for most data assimilation schemes.

Daley (1992a; 1992b; 1992c; 1992d) has shown that the performance of an assimilation system depends on an accurate estimation of the input error covariances. The theory of estimation of error covariances can be complex, and the procedures for doing so are limited. These procedures revolve around assumptions that are needed to make the problem tractable. Estimating each component of an error covariance matrix is an insurmountable task because there are simply not enough realizations or data that would permit independent estimation of each element of a covariance matrix. For that reason, covariance modelling, as discussed in Section 3.3, is essential. Nevertheless, within this framework, estimating only parameters of a covariance model (rather than the covariance matrix itself), which achieves a (near) optimal assimilation system raises the important question on how that can be established.

In meteorology, several techniques such as the National Meteorological Center (NMC) lagged-forecast method (Parrish and Derber 1992), the ensemble of data assimilation (Fisher 2003), the Hollingsworth–Lönnerberg technique (Hollingsworth and Lönnerberg 1986; Lönnerberg and Hollingsworth 1986), the maximum likelihood method (Dee and da Silva 1999; Dee et al. 1999), and the Desroziers et al. (2005) diagnostic have been used to estimate error covariances or their parameters. These statistical diagnostics are either based on innovations or assimilation residuals. They provide a reasonably accurate estimate of the error covariances (parameters) under the assumption that the underlying assimilation system is already nearly optimal (Desroziers et al. 2005; Menard 2016; Waller et al. 2016a; Tandeo et al. 2020). In meteorology, numerous observations are being used, and error statistics have been tuned, so that the assimilation system is already close to optimal. Accordingly, the estimation of the observation error of a new observation

type in meteorology can be assessed through the above techniques (see the discussion in Waller et al. (2016a) or Ménard (2016)). However, with methane assimilation or chemical data assimilation in general, we are rarely building upon an existing and well-proven assimilation system, but rather constructing one with little prior information. In this context, the optimality of the assimilation system is not granted, and needs to be established in addition to estimating the error covariance parameters.

Recently, a new estimation method has been introduced which does not rely on the optimality of the assimilation system. The method is based on cross-validation and innovation covariance consistency (Menard and Deshaies-Jacques 2018a; 2018b). First, it is shown that the cross-validation technique estimates the true analysis error variance without the assumption of an optimal analysis. Then, by varying the tunable covariance parameters to obtain the minimum analysis error variance (evaluated by cross-validation) while preserving the innovation covariance consistency, the necessary and sufficient conditions for the estimation of the true error covariances are met (Menard 2016). Thus, the analysis formed is nearly optimal (because the estimation is performed on parameters rather than the full covariance matrix), and the error statistics obtained are close to the true error statistics (Menard and Deshaies-Jacques 2018a).

In this Chapter 6, the cross-validation methodology which was developed for in situ observations (Menard and Deshaies-Jacques 2018a), has been extended to satellite observations and applied to estimate multiple covariance parameters. With the PvKF scheme, the estimation of the background error correlation, in particular correlation lengths, is quite important for obtaining an optimal analysis (Ménard and Chang 2000). However, we should note that finding the (near) optimal covariance parameters results in

additional assimilation runs that compound the global cost of assimilation by a factor of (roughly) 20. The technique (and results) of obtaining accurate covariance parameters is an important objective of this chapter. In a nutshell, we not only estimate the methane concentration field but also determine the most accurate input error covariance parameter values.

Finally, we note that because of the common aspects between inverse modelling and data assimilation schemes, the estimation of data assimilation error covariances could also be useful for the estimation of error covariances needed in inverse modelling schemes (although we have not attempted to demonstrate this in the current study).

Chapter 6 is organized as follows. First, we present in Section 6.2 the background of the theory of estimating covariance parameters, which includes the necessary and sufficient conditions and how they are linked with the method of cross-validation. In Section 6.3, we present the experimental setup and the preparation of GOSAT retrievals for the optimization framework, resulting in an estimate of the correlation lengths together with observation error variance. Section 6.4 discusses the estimation of the modelling error and initial error variance. In Section 6.5, we evaluate the analysis against several sets of independent observations. Section 6.6 first presents the spatial distribution of the analysis increment and analysis error variance and then discusses the temporal behaviour of the analysis. Finally, conclusions are drawn in Section 6.7.

## **6.2 Background on the Theory of Covariance Parameter Estimation**

In this section, we will review the theory of estimation of observation error covariance ( $\mathbf{R}$ ) and background error covariance in observation space ( $\mathbf{H}\mathbf{B}\mathbf{H}^T$ ) as developed in Menard and Deshaies-Jacques (2018) (Menard and Deshaies-Jacques 2018a;

2018b) and Menard (2016). By assuming a linear observation operator and uncorrelated observation and forecast errors, it was shown that the necessary and sufficient conditions to estimate the true error covariances (in observation space, i.e.,  $\mathbf{R}$  and  $\mathbf{HBH}^T$ ) are (Menard 2016)

1. Innovation covariance consistency, i.e.,

$$\tilde{\mathbf{\Gamma}} = \mathbf{\Gamma} \quad (6.1)$$

2. Optimality of the gain matrix, i.e.,

$$\tilde{\mathbf{K}} = \mathbf{K} \quad (6.2)$$

The first condition (Equation (6.1)) indicates that the sample covariance of the Observation – Model residuals (i.e.,  $\tilde{\mathbf{\Gamma}} = \mathbb{E}[(O-B)(O-B)^T]$ ) is equal to the innovation covariance computed in the assimilation algorithm (i.e.,  $\mathbf{\Gamma} = \mathbf{HBH}^T + \mathbf{R}$ ). The second condition (Equation (6.2)) expresses that the Kalman gain,  $\mathbf{K}$ , used in the assimilation algorithm is equal to the optimal Kalman gain,  $\tilde{\mathbf{K}}$ , that is the gain that uses the true observation error covariance and the true background error covariance. These conditions are particularly easy to interpret in a scalar problem. In a scalar problem, the observation error covariance is an error variance, and let its true value be denoted by  $\sigma_o^2$ . Similarly, the background error covariance is an error variance, and let its true value be denoted by  $\sigma_b^2$ . Equation (6.1) then says that the variance of the Observation minus Model,  $\text{var}(O-B)$ , is equal to the sum of the true observation and true background error variances, i.e.,  $\sigma_o^2 + \sigma_b^2$ . In the scalar problem, the Kalman gain depends only on the ratio of the observation error variance to the background error variance, not their values as such. Equation (6.2) then says that the Kalman gain used in the assimilation uses the ratio of the true error variances (not their values). Thus, if the sum of error variance is the sum of the

true error variances, and the ratio of error variances is equal to the ratio of the true error variances, then it implies that, individually, the observation and background error variances are equal to their true values. The conditions (1) and (2) are the generalization for error covariance matrices.

However, condition (2) is nontrivial to implement, as there is no measure of what the true Kalman gain is. Fortunately, there is an alternative formulation. It is known that the analysis error covariance for any (arbitrary) Kalman gain can be computed as

$$\mathbf{A} = (\mathbf{I} - \mathbf{K}\mathbf{H})\mathbf{B}^{true}(\mathbf{I} - \mathbf{K}\mathbf{H})^T + \mathbf{K}\mathbf{R}^{true}\mathbf{K}^T \quad (6.3)$$

where  $\mathbf{B}^{true}$  and  $\mathbf{R}^{true}$  are the true background error covariance and true observation error covariances, respectively. Minimizing the total analysis error variance, i.e.,  $tr(\mathbf{A})$ , with respect to the gain matrix, in fact, yield a gain matrix that is the true Kalman gain (Daley 1992d)

$$\arg \min_{\mathbf{K}} \{tr(\mathbf{A})\} = \hat{\mathbf{K}} = \mathbf{B}^{true}\mathbf{H}^T(\mathbf{H}\mathbf{B}^{true}\mathbf{H}^T + \mathbf{R})^{-1} = \mathbf{K}^{true} \quad (6.4)$$

Thus, we can replace condition (2) by the following condition

### 3. Optimality of the gain matrix, using

$$\arg \min_{\mathbf{K}} \{tr(\mathbf{A})\} = \mathbf{K}^{true} \quad (6.5)$$

The minimum can, in fact, be obtained using cross-validation (Menard and Deshaies-Jacques 2018a). We will detail the algorithm shortly, but first, a few comments are needed. The difficulty in the application of these conditions (i.e., Equations (6.1) and (6.5)) is that we never have enough data or realization to effectively verify each matrix element of these conditions (the  $\mathbf{K}$  matrix in the case of condition (3), and the full innovation covariance matrix with condition (1)). Because of this difficulty, we usually model the error

covariances with a few parameters, say a vector of parameters  $\boldsymbol{\alpha}$ , and then examine whether or not condition (3) and the trace of condition (1) are verified with “optimal parameters values”.

In our notation, we also distinguish the error covariances that are basically defined using the mathematical expectation in probability theory, from a statistical estimate of using a finite sample (e.g., one hundred or less). We will denote the mathematical expectation with  $E$  and sample mean as  $\langle \rangle$ . Thus, let us consider that the analysis error covariance,  $\mathbf{A}(\boldsymbol{\alpha})$ , depends on a number of covariance parameters,  $\boldsymbol{\alpha} = \{L_h, L_v, f^o, \dots\}$ , where  $L_h$  and  $L_v$  are the horizontal and vertical correlation lengths, and  $f^o$  is a multiplicative factor for the observation error variance (see Section 5.4.3 and 5.4.4). The idea then is to find the optimal values of parameters such that  $\arg \min_{\boldsymbol{\alpha}} \{tr(\mathbf{A}(\boldsymbol{\alpha}))\} \approx \mathbf{K}^{true}$ . Solving this problem using a limited number of parameters can be done using cross-validation (Menard and Deshaies-Jacques 2018a). There are two principal classes of cross-validation techniques. One is the leave-one-out (observations), and the other is the k-fold methodology. Here we consider the k-fold methodology as it is easier to apply in this context. In k-fold cross-validation, we separate the observations into  $k$  subsets of equal size. Here, we use a  $k$ -fold of 3. An analysis using  $k-1$  observation sets (called active observations) is compared with the remaining set of observations, called passive observations. By comparing the passive observations with the analysis interpolated at the passive observation sites, we construct the cross-validation cost function,

$$J_c = \langle (O - A)_c^2 \rangle, \quad (6.6)$$

where the ensemble also includes all permutations of the  $k$  subsets (Menard and Deshaies-Jacques 2018a; 2018b). It turns out that the cost function,  $J_c$ , is actually a measure of the analysis error variance (Marseille et al. 2016; Menard and Deshaies-Jacques 2018a). Indeed, assuming that the observation errors are spatially uncorrelated and uncorrelated with the forecast (or background) errors, it is established that

$$\mathbb{E}[(O-A)_c(O-A)_c^T] = \mathbf{R}_c + \mathbf{H}_c \mathbf{A} \mathbf{H}_c^T, \quad (6.7)$$

whether the analysis is optimal or not. In Equation (6.7), subscripts  $c$  denotes values estimated in the passive observation space that are independent observations.  $\mathbf{H}_c$  is the observation operator that interpolates the 3D field at the passive observation sites. By exploring the values in the parameter space, we can estimate the value of the cost function, Equation (6.6), for each parameter value, until we find the minimum of the cost function (Equation (6.6)). Thus, we argue that

$$\arg \min_{\boldsymbol{\alpha}} J_c(\boldsymbol{\alpha}) \Rightarrow \text{optimal } \mathbf{A}(\boldsymbol{\alpha}) \quad (6.8)$$

in the subspace spanned by the covariance parameters,  $\boldsymbol{\alpha}$  (Menard and Deshaies-Jacques 2018a). The essential part of this search also consists in maintaining the innovation covariance consistency in Equation (6.1) (in fact, the trace of it), so that we then get an estimate of the optimal parameters values that satisfies both conditions; condition (1) and condition (3). The modelled covariances with these optimal parameter values are then an estimate of the true error covariances.

### 6.3 Estimation of Correlation Lengths and Observation Error Variance

The cross-validation estimation technique (Menard and Deshaies-Jacques 2018a) was originally developed for in situ observations. Contrary to in situ observations, satellite

observation errors may contain significant spatial correlation. The error correlation mainly arises from the representativeness part of the error and inter-channel retrieval (Waller et al. 2016a), which may result in a degraded analysis obtained through cross-validation (Menard and Deshaies-Jacques 2018a). In practice, observation thinning (i.e., reducing the spatial density of the observations) or inflating the observation error variance are two standard procedures to deal with spatially correlated observation errors (Bormann and Bauer 2010; Bormann et al. 2010). Procedures based on variance inflation were employed to maintain a better consistency between the model and GOSAT (Maasakkers et al. 2019) using the residual error method of Heald et al. (2004) or between GOSAT and independent observations (Stanevich et al. 2021). However, in this study, we use observation thinning to alleviate the error correlation between observations. This is due to the fact that our main goal here is optimizing the error covariance parameters in a more objective manner (i.e., no optimality assumptions), rather than tuning the covariance parameters for better consistency between the model and observations. Note that this approach also aids PvKF in obtaining a higher computational efficiency due to lowering the number of assimilated observations (see Section 5.5.2). GOSAT SWIR retrievals, used in this study, are single-channel and considered sparse compared to other satellites such as AIRS and IASI with dense multichannel retrievals. Thus, the GOSAT SWIR observations retain smaller spatial correlated errors, making the practical solution of thinning more feasible. Note that the GOSAT observation covariance matrix is usually assumed to be diagonal (i.e., spatially uncorrelated errors) in methane source inversions due to a lack of better objective information (Lu et al. 2021; Maasakkers et al. 2021; Qu et al. 2021).

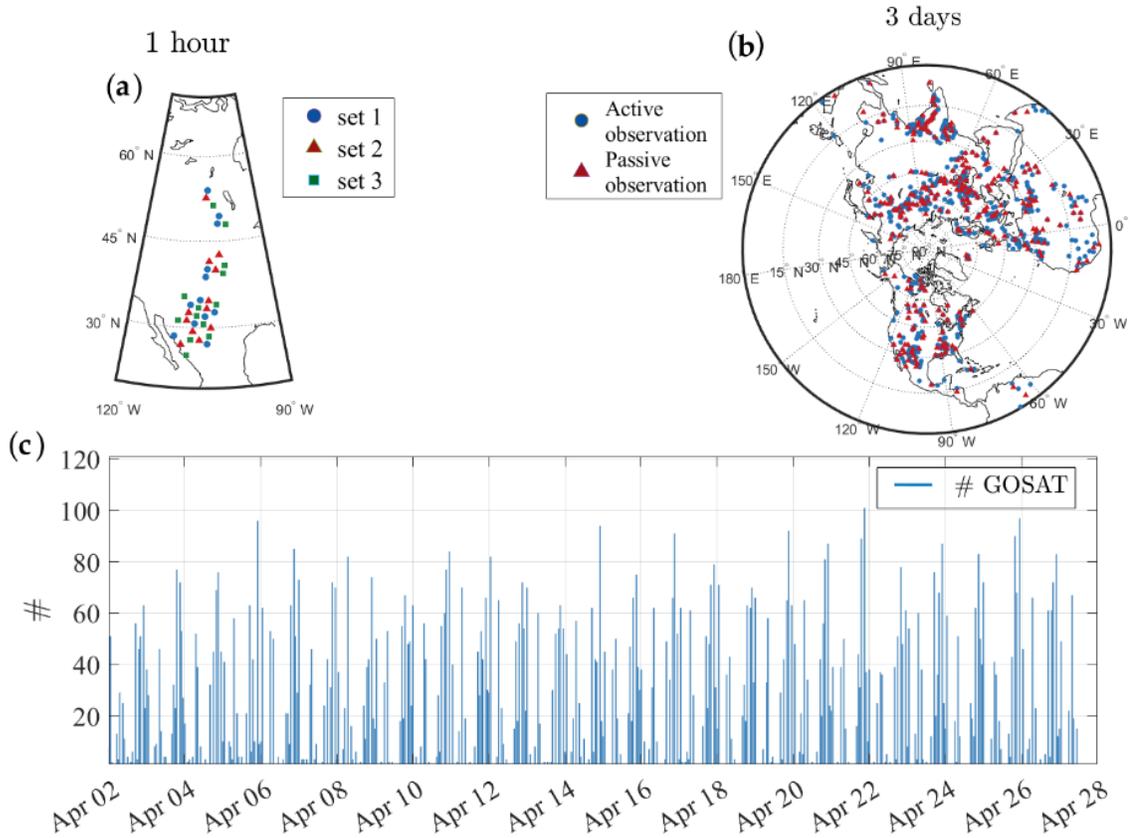
One objective of this section is to obtain the observation error variance. First, we conduct observation thinning to maintain nearly uncorrelated observation error as required by cross-validation. As it is shown for SEVIRI (Waller et al. 2016b), satellite nadir observations such as GOSAT could also represent spatially correlated errors of a few tens of kilometres. In this work, we simply assume a 10 km margin to filter any pair of GOSAT observations that are located in this range. We apply this filtering to the observations over the entire time window of assimilation, resulting in about 40% removal of observations after quality control. We argue that a small length (e.g., 1 km with about 15% removal) will not ensure obtaining uncorrelated (or near uncorrelated) observation errors, whereas a large marginal length (e.g., 100 km with about 85% removal) may lead to significant degradation of the analysis through filtering a great portion of observations. Furthermore, as we shall see later in this section, the thinned satellite observations can be used in cross-validation to obtain the optimal parameter values. The quality control first removes the outliers by filtering out observations whose departure from the global mean is three times larger than its standard deviation. The quality-controlled observations are then subjected to the thinning process as described above and separated into three sets of observations of equal numbers for 3-fold cross-validations. The selection of observations into three sets is made according to the order of retrieval time, resulting in a spatially random distribution of retrievals in each set (Figure 6.1a). The cross-validation is then applied by leaving one set out as a test set (i.e., passive observations) and using the remaining two sets to generate the training sets (i.e., active observations). Figure 6.1b shows the active and passive observations within three days or one revisit cycle of GOSAT. We recall that the analysis

is only produced using active observations. The total number of observations per hour shows the nine revisit cycles of GOSAT observations used in this study (Figure 6.1c).

Here we describe the experimental setup to estimate the covariance parameters. Three parameters, including  $L_h, L_v$ , and  $f^o$ , are considered for the optimization problem using cross-validation cost-function (Equation (6.4)). We recall that observation, model, and initial error covariance are considered uniform and uncorrelated (see Section 5.4.4 for details). The background error covariance adapts a homogeneous isotropic horizontal and vertical error correlation based on a second-order autoregressive (SOAR) correlation model (see Section 5.4.3). A series of hourly methane analyses are conducted for a period of two weeks (5 April to 18 April 2010), with 3 days spin-up of the assimilation system. This analysis is repeated to find the parameter optimum, altering  $L_h$  from 200 km to 600 km with a step size of 50 km,  $L_v$  from 0 to 13 vertical  $\sigma$  levels with  $1\sigma$  increment, and  $f^o$  from 0.1 to 1.2 with a 0.1 step size.

Simultaneous optimization of the three parameters as discretized above requires a total ensemble of assimilations of 1512, i.e.,  $9 \times 14 \times 12 = 1512$ , corresponding to the specified parameters,  $L_h \times L_v \times f^o$ , in the optimization. Perhaps, other optimization techniques could resolve this problem more efficiently (Wen and Yin 2013; Zhu et al. 2017). An iterative scheme is adapted instead, where  $L_h$  and  $L_v$  are estimated together while  $f^o$  is considered separately. With the initial values of  $L_h = 500$  km,  $L_v = 1\sigma$ , and  $f^o = 1$  taken as first guesses (Table 6.1, initial), the experiment is started for the estimation of  $f^o$  while  $L_h$  and  $L_v$  are kept fixed. This essentially corresponds to searching for the  $f^o$  that

minimizes the cross-validation cost function (Equation (6.6)) while the innovation covariance consistency (Equation (6.1)) is respected.



**Figure 6.1. Spatial and temporal distribution of GOSAT methane observations used in cross-validation. (a) 1-h GOSAT observations separated in three sets after thinning; (b) active observations (blue circles) versus passive observation (red triangles) over 3 days in the cross-validation framework to estimate the covariance parameters; (c) frequency of GOSAT observation per hour used in cross-validation.**

Note that the procedure is repeated for all three permutations of cross-validation subsets to ensure that all observations have been used for evaluation. The new estimate of  $f^o$  is obtained by averaging for the three verifying subsets (Table 6.1, itr 0 / step 1). The estimated  $f^o$  ( $f^o = 0.45$ ) is then used to estimate  $L_h$  and  $L_v$  together in the next step of zero iteration (Table 6.1, itr 0 / step 2). The procedure is repeated for successive iterations until

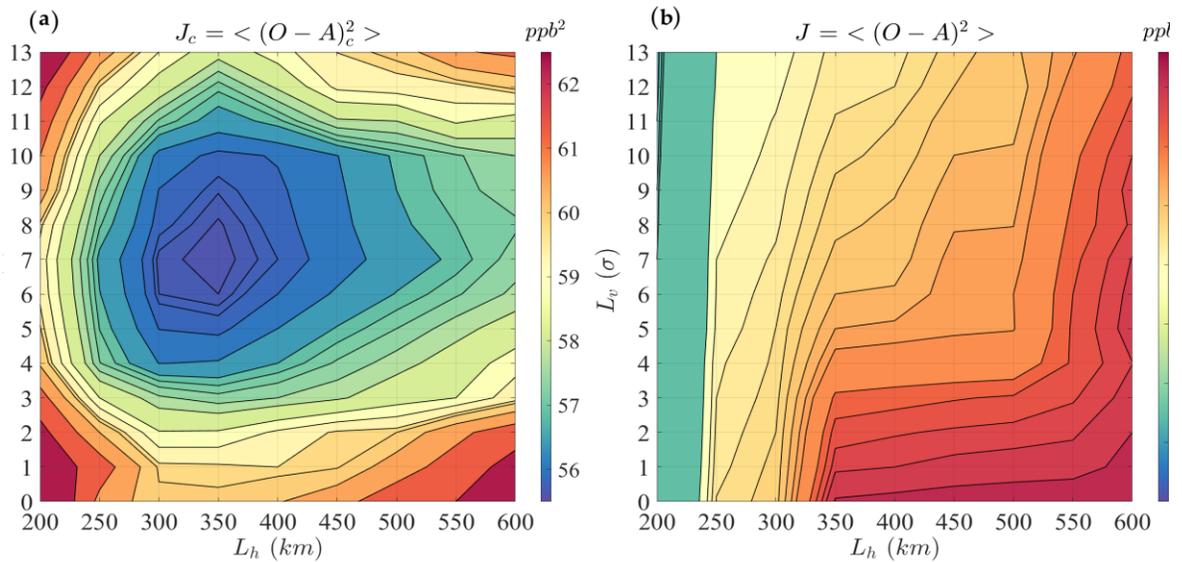
convergence. Nevertheless, it is established that a single iteration will be sufficient to reach convergence for a similar estimation method (Desroziers et al. 2009; Ménard et al. 2009; Menard and Deshaies-Jacques 2018a; 2018b). We also confirm this for our estimation problem as shown in the first iteration (itr 1) in Table 6.1 (itr 1/step 1 and itr 1/step 2). Thus, the number of computations declines to  $2 \times (9 \times 14 + 12) = 276$  for the same number of parameters and the same step size of each parameter.

**Table 6.1.** Comparison of the error variance parameters ( $L_h, L_v$ , and  $f^o$ ) along with the cross-validation cost function generated with passive observations at different stages of iterative optimization, including the initial step, zero iteration (itr 0) and first iteration (itr 1). Each iteration contains two steps, one for optimizing  $f^o$  (step 1) and another for optimizing  $L_h$  and  $L_v$  together.

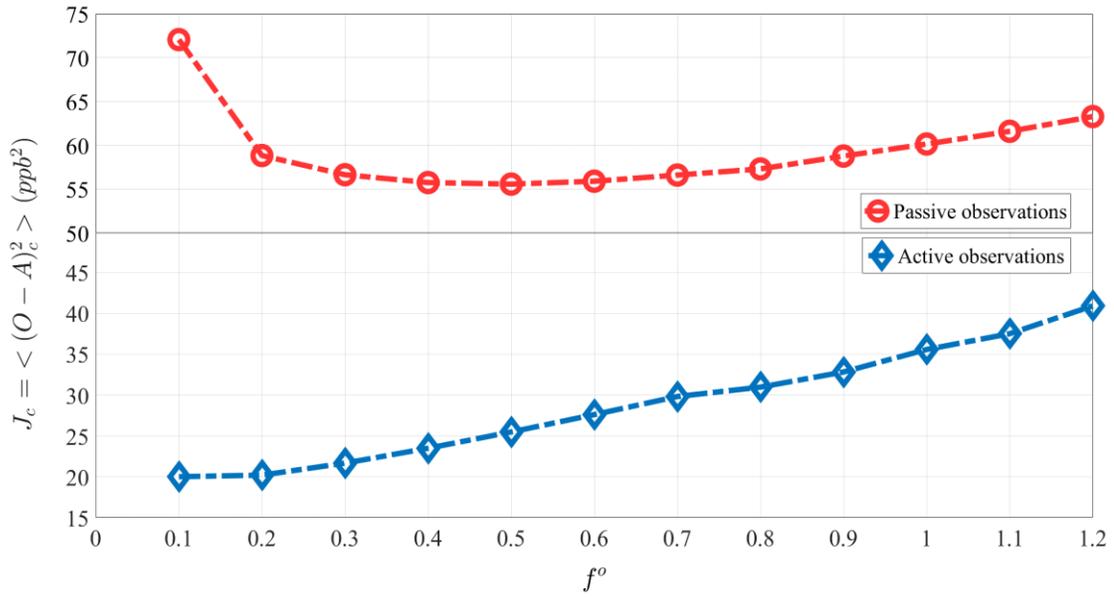
<b>itr/Step</b> \ <b>Parameters</b>	$L_h$	$L_v$	$f^o$	$J_c = \langle (O - A)_c^2 \rangle$
initial	500 km	$1\sigma$	1	62.4 ppb <sup>2</sup>
itr 0/step 1	500 km	$1\sigma$	0.45	60.5 ppb <sup>2</sup>
itr 0/step 2	350 km	$7\sigma$	0.45	55.9 ppb <sup>2</sup>
itr 1/step 1	350 km	$7\sigma$	0.5	55.6 ppb <sup>2</sup>
itr 1/step 2	350 km	$7\sigma$	0.5	55.6 ppb <sup>2</sup>

Figure 6.2a and Figure 6.3 (red curve) illustrate the estimation of  $L_h, L_v$ , and  $f^o$ , after the first iteration that corresponds to the minimum value with the specified optimization resolution in the cross-validation (i.e., passive observation) space. A cost function based on the active observations,  $J = \langle (O - A)^2 \rangle$ , is constructed to compare with the cross-validation results. Figure 6.2b and Figure 6.3 (blue curve) represent the similar parameter estimation procedure described above, but over the active observations (i.e., training set). In Figure 6.3, the evaluation against active observations (blue curve) always shows smaller variances than the evaluation against passive observations (red curve). It implies that the

use of active observations overestimates the performance (a well-known property of cross-validation optimization (Menard and Deshaies-Jacques 2018a)) of the analysis. In addition, the use of passive observations in cross-validation results in the existence of a minimum variance cost function, consistent with the finding of Ménard and Deshaies-Jacques (2018b) (Menard and Deshaies-Jacques 2018a).



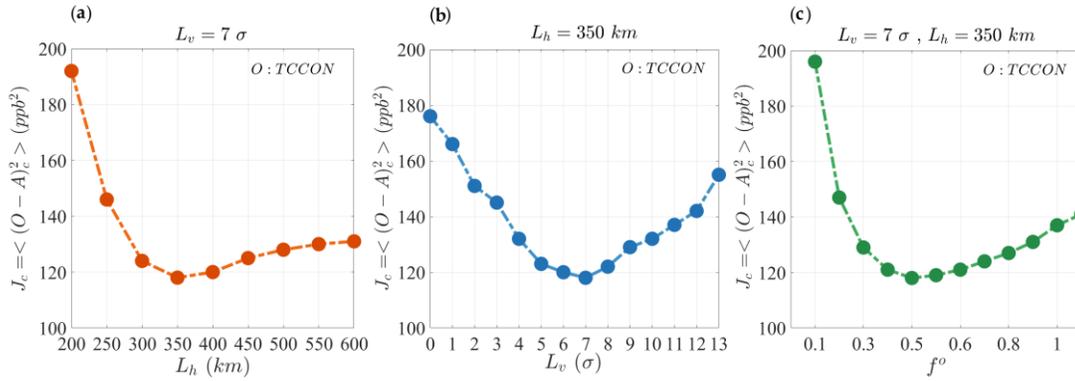
**Figure 6.2.** Estimation of the horizontal length scale ( $L_h$ , x-axis) and the vertical length scale ( $L_v$ , y-axis) together through (a) a cross-validation cost function of passive observations  $J_c = \langle (O - A)_c^2 \rangle$  and with (b) a cross-validation cost function of active observations  $J = \langle (O - A)^2 \rangle$ .



**Figure 6.3.** Estimation of observation error variance parameter,  $f^o$ , using a cross-validation cost function of passive observations (red curve) and using a cross-validation cost function of active observations (blue curve).

To verify that the cross-validation method using 10 km thinning of satellite observations yields the optimal parameter, we have conducted the same analysis as shown above, but this time against independent Total Carbon Column Observing Network (TCCON) observations. Note that TCCON is often considered as a reliable observation data set. However, their coverage is rather limited in space but continuous in time (seven sites only; see Section 5.5 for detail of TCCON observations). Figure 6.4 shows the comparison of analysis ( $A$ ) using active GOSAT observations against TCCON ( $O$ ), which is an independent source of observation. The cross-validation cost function,  $\langle (O - A)^2 \rangle$ , is drawn against the three parameters ( $L_h$  in the left,  $L_v$  in the middle, and  $f^o$  in the right panel of Figure 6.4). To save on the computation, we perform the estimation on each parameter individually. Figure 6.4 shows that the optimal parameter values obtained

against TCCON are the same as those estimated through cross-validation (using GOSAT passive observations). Hence, despite the fact that cross-validation was developed for in situ observations, this result indicates that satellite observation thinning can indeed yield the optimal values. This example shows that the applicability of the cross-validation method (Menard and Deshaies-Jacques 2018a; 2018b) to the satellite observations is valid.



**Figure 6.4.** Estimation of the (a) horizontal correlation length,  $L_h$ , (b) vertical correlation length,  $L_v$ , and (c) observation error covariance parameter,  $f^o$ , using independent TCCON observations with the cross-validation cost function,  $J = \langle (O - A)^2 \rangle$ .

## 6.4 Estimation of Model Error and Initial Error Variance Using Innovation

### Variance Consistency

By conducting the estimation presented in Section 6.3 on a different time window (e.g., a different two weeks period or a shorter/longer period), we found that the optimal estimation of the three parameters,  $L_h, L_v$ , and  $f^o$ , is relatively insensitive to the time window of estimation. The model and initial error variances, on the other hand, (or their covariance parameter counterparts,  $f^q$  and  $f^i$ ) are excluded from our covariance parameter vector,  $\alpha$ , in the cross-validation optimization due to their time-varying influence (i.e., accumulating or decaying behaviour). In other words, having a different optimization time

window results in a different set of optimized  $f^q$  and  $f^i$ . Thus, instead of including  $f^q$  and  $f^i$  in the time-insensitive parameter vector ( $\boldsymbol{\alpha}$ ), we tuned them separately using the variance matching diagnostic technique defined in Section 6.2 (Equation (6.1)). A description of the form of initial and model error covariance is presented in Section 5.4.4. First, we characterize a normalized variance matching diagnostic that offers a diagnostic independent of units and independent of the number of observations. The covariance matching diagnostic, reformulated as the total variance matching, can be written as

$$tr\left(\mathbb{E}[(O-B)(O-B)^T]\right) - tr(\boldsymbol{\Gamma}) = tr(\tilde{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}). \quad (6.9)$$

The normalized variance matching diagnostic is then obtained by dividing the right-hand side of Equation (6.9) with the number of observations,  $p$ , and  $\text{var}(\boldsymbol{\Gamma})$  representing the variance at a specific observation location or the mean of all observations. Thus, we have

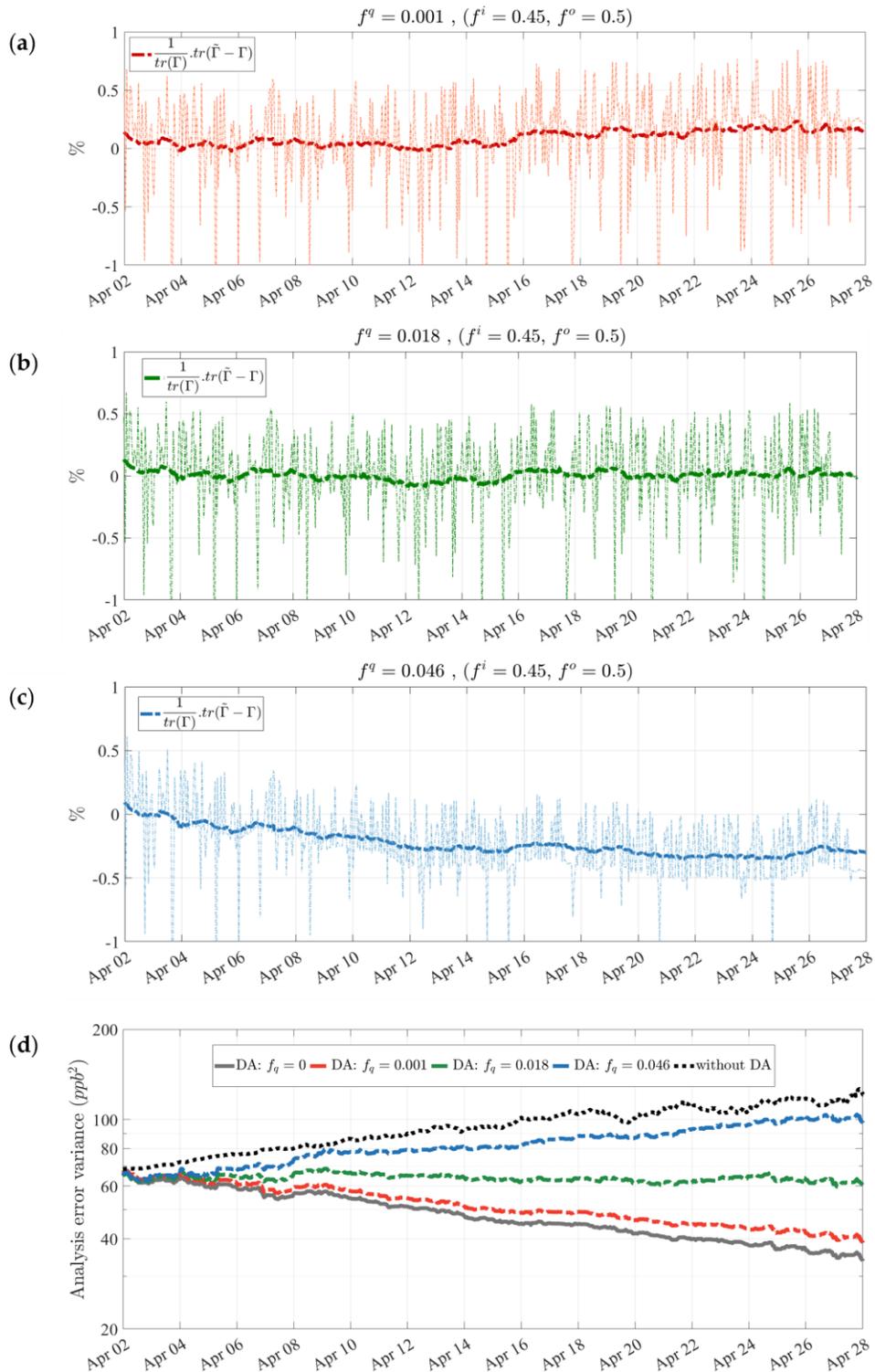
$$\frac{1}{\text{var}(\boldsymbol{\Gamma})} \frac{1}{p} tr(\tilde{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}) = \frac{1}{tr(\boldsymbol{\Gamma})} tr(\tilde{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}). \quad (6.10)$$

Writing the diagnostic in this fashion removes the dependency of the diagnostic to the absolute statistics, so that it retains a similar form to other diagnostics such as  $\chi^2$  (Menard et al. 2000). Therefore, an appropriate value of model error in our estimation system is obtained when it satisfies the near-zero normalized variance matching with continuous stability over time. Figure 6.5 illustrates the influence of the model error on this diagnostic. A small  $f^q$  (Figure 6.5a) leads to an increasing pattern ( $\tilde{\boldsymbol{\Gamma}} > \boldsymbol{\Gamma}$ ) over time, whereas a large  $f^q$  (Figure 6.5c) results in a decrease ( $\tilde{\boldsymbol{\Gamma}} < \boldsymbol{\Gamma}$ ); both of them tend to disagree with the innovation variance consistency over time. A proper value of  $f^q = 0.018$  (Figure 6.5b) can be found with trial and error that fulfills the innovation variance

consistency. Note that the initial error variance parameter is kept constant  $f^i = 0.45$  for all cases, while it is shown for its best estimate (see Figure 6.6).

Furthermore, we verify the influence of the model error on the analysis error variance in Figure 6.5d. Both cases of “without DA” and perfect model or ( $f^q = 0$ ) produce too small or too large analysis error variance as time proceeds, causing degraded assimilation results with overestimation or underestimation implications. Therefore, the presence of a certain amount of model error is crucial. As found in the diagnostic above, we infer that the proposed model error parameter ( $f^q = 0.018$ ) also maintains a stable analysis error variance over time.

We repeat the same diagnostic experiment to obtain an appropriate  $f^i$ . We note that different values of  $f^q$  are considered initially for this test, and the best value,  $f^q = 0.018$ , which maintains a stable diagnostic, is shown in Figure 6.6. Figure 6.6 suggests that  $f^i = 0.45$  meets the condition of innovation variance consistency, although it may not reach the final error variance within two weeks to show asymptotic stability of the Kalman filter estimation (see Jazwinski (1970), Theorem 7.4). We infer that the thinned GOSAT data used in this study may produce an analysis that induces a lack of observability (e.g., see Jazwinski (1970) for the definition of observability in section 7.5 therein). Nonetheless, we have all the parameters that lead to an optimal analysis and to innovation variance consistency. Those estimated parameters are close to the true error variance parameters.



**Figure 6.5. Normalized innovation variance consistency diagnostic for a (a) low ( $f^q = 0.001$ ), (b) proper ( $f^q = 0.018$ ), and (c) high ( $f^q = 0.046$ ) model error parameter value. (d) indicates the effect of model error on analysis error variance over time.**

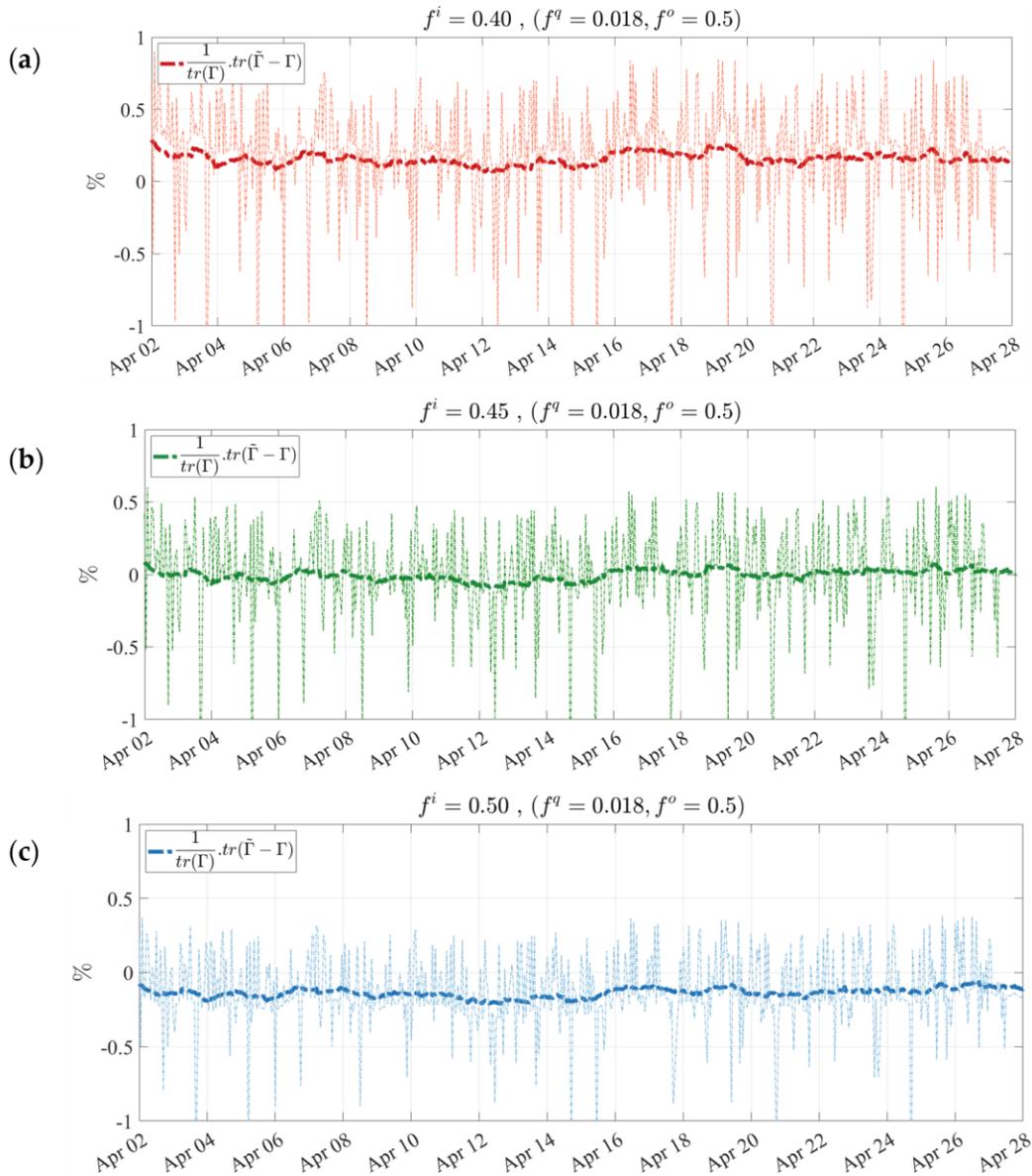


Figure 6.6. Normalized innovation variance consistency diagnostic for a (a) low ( $f^i = 0.40$ ), (b) proper ( $f^i = 0.45$ ), and (c) high ( $f^i = 0.50$ ) initial error parameter value.

### 6.5 Evaluation against Independent Observations

The optimal analysis can be obtained as a result of successful assimilation. This relies not only on the assimilation processes but specifically on the precise characterization of the input error parameters. The observation error covariance ( $\mathbf{R}$ ), the model error covariance ( $\mathbf{Q}$ ), and in some cases, the initial error covariance ( $\mathbf{P}_0$ ), besides the error

correlation length scales, are among those determined in the PvKF assimilation. Hence, the optimality of the analysis depends on how precisely the Kalman filter error covariances ( $\mathbf{P}^f, \mathbf{P}^a$ ) reflect the true estimation of error covariance parameters. We recall the definition and the formulation of the input error covariances in Section 5.4.4

The assimilation results shown in Sections 6.5 and 6.6 are produced with thinned GOSAT data that pass all quality control flags (e.g., retrieval flag, outliers), maintaining about 35%- 50% of all GOSAT retrievals for assimilation. Accordingly, the desired assimilation involves globally optimized (or quasi-optimal) error parameters, including  $f^o, f^i, f^q, L_h$  and  $L_v$ . After finding these optimized error parameters, as discussed in Sections 6.3 and 6.4, independent measurements from multiple sources are used to evaluate the assimilation performance. Figure 6.7 demonstrates the location/pathways of all surface/aircraft available measurements collected for the duration of our assimilation results in April 2010. Below, we provide a brief description alongside the preparation of these independent observations.

TCCON is a ground-based FTS network that provides a time series of  $\text{CH}_4$  column-averaged abundance worldwide. TCCON data has been used to compare with model simulations and satellite data primarily for validation purposes (Yoshida et al. 2013; Scheepmaker et al. 2015; Zhou et al. 2016; Liang et al. 2017; Wunch et al. 2019; Stanevich et al. 2021; Zhang et al. 2021). We used the GGG2014 version of TCCON  $\text{XCH}_4$  data from seven sites (red circles in Figure 6.7), including Park falls (Wennberg et al. 2017), Orleans (Warneke et al. 2017), Lamont (Wennberg et al. 2016), Bremen (Notholt et al. 2019), Sodankyla (Kivi et al. 2014), Izana (Blumenstock et al. 2017), and Bialystok (Deutscher et al. 2019), available at <https://tccodata.org/2014>. TCCON is calibrated using aircraft

profiles to maintain better than 0.5% accuracy of XCH<sub>4</sub> retrievals (Wunch et al. 2017). Prior to the comparison with the model or analysis, we convolve CMAQ with TCCON average kernels and the corresponding a priori profiles, for which the procedure is described in Wunch et al. (2010) or through the website <https://tcon-wiki.caltech.edu/Main/AuxiliaryData>. Similar to the aircraft measurements described later in this section, TCCON retrievals are assumed to provide an independent evaluation for our assimilation results. This is due to the fact that TCCON is neither used in generating the model input nor in calibrating the GOSAT data with CMAQ. We recall that GOSAT data are bias-corrected against only surface measurements of GLOBALVIEWplus CH<sub>4</sub> ObsPack v3.0 as described in Sections 5.4.1 and 5.4.2. Therefore, any influence of TCCON calibration on GOSAT data prior to this study will be entirely alleviated. Note that we employed the same latitudinal correction on TCCON as we used for GOSAT to ensure a consistent comparison with independent observations

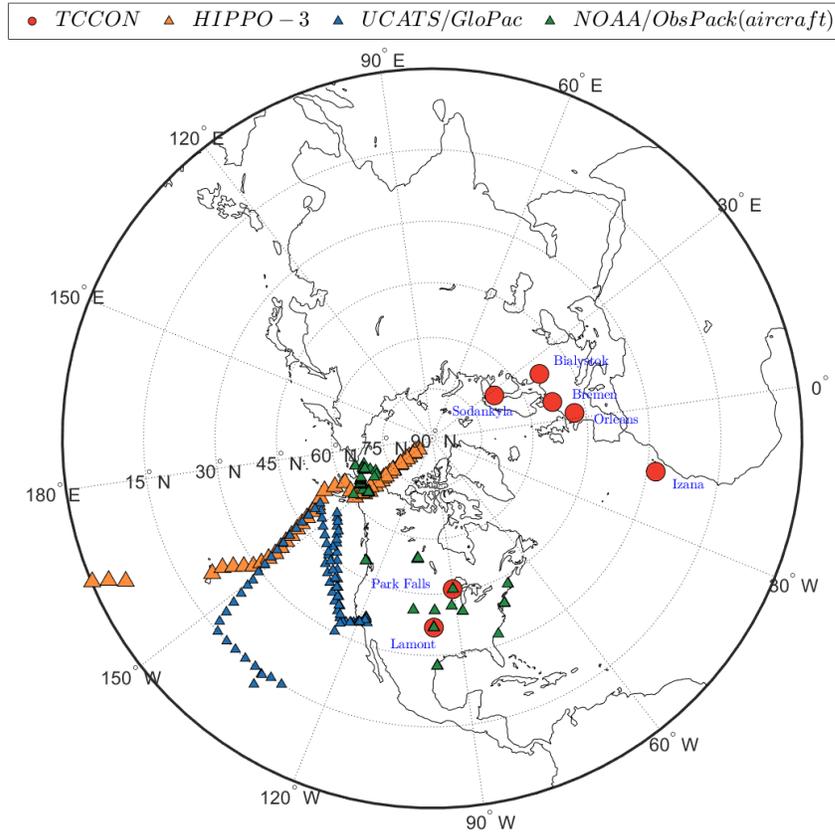
HIAPER Pole-to-Pole Observations (HIPPO-3) are among the five HIPPO aircraft missions that provide methane concentration measurements (Wofsy et al. 2011) from surface to 14 km across the Pacific Ocean between 20 March and 20 April 2010. The measurement is conducted every second using a quantum cascade laser spectrometer (QCLS) with an accuracy of 1 ppb. Methane measurements, meteorology and flight tracking data are accessible at: [https://www.eol.ucar.edu/field\\_projects/hippo](https://www.eol.ucar.edu/field_projects/hippo). For the North Hemisphere, the measurement data are available on 10, 13 and 15 April 2010, as shown by orange triangles in Figure 6.7. To compare with HIPPO-3, we interpolate the model concentration at the specific location, height, and time of the measurement. The significantly shorter time scale of the measurements (i.e., 1 s) compared to the model

simulation time step (i.e., 1 h) allows a larger variation, resulting in a degraded comparison with the model or analysis at the observation space. Therefore, with the aim of meaningful comparison, we exclude those HIPPO-3 data that depart by more than three standard deviations from the average measurement over one minute or 60 consecutive measurements. This results in removing 58% of the total HIPPO-3 data as the outliers for this comparison.

Global Hawk Pacific (GloPac) is an aircraft mission operated by NASA in April 2010 primarily designed to validate monitoring satellite missions and to measure trace gases in the upper troposphere and lower stratosphere. UAS Chromatograph for Atmospheric Trace Species (UCATS) was integrated into GloPac to measure methane alongside N<sub>2</sub>O, SF<sub>6</sub> and several other trace gases. It offers an overall precision of 0.5% for methane (Hints et al. 2021). All the data are available at <https://espoarchive.nasa.gov/archive/browse/glopac>, where the methane measurements were recorded on 7 and 13 April 2010 with a time resolution of 140 s and are shown in Figure 6.7 by blue triangles. Note that we use methane measurements without filtering, and the model and analysis are computed at every measurement.

GLOBALVIEWplus ObsPack v3.0 data product (Schuldt et al. 2021) published via NOAA Global Monitoring Laboratory provides high accuracy measurements of methane concentration from a variety of sampling platforms, including surface, tower, aircraft, and shipboard measurements. Since the surface flask and tower data are used in the calibration of GOSAT (see Sections 5.4.1 and 5.4.2), we only use ObsPack aircraft data to ensure an unbiased comparison with the analysis. Accordingly, only daily above 800 m from the dataset of ObsPack available at (<https://gml.noaa.gov/ccgg/obspack/>) is used to represent

the aircraft measurements, which are collected from multiple aircraft campaigns (Schuldt et al. 2021). These measurements are demonstrated by green triangles in Figure 6.7.



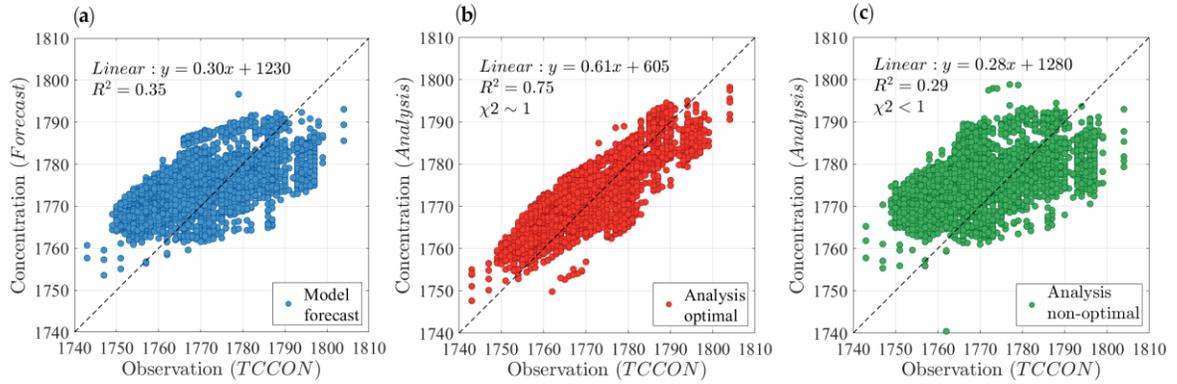
**Figure 6.7. Methane measurement data used for the evaluation of the assimilation system in April 2010: TCCON stations (red circles) at seven sites; HIPPO-3 aircraft measurement pathways (orange triangles) on 10, 13 and 15 April 2010; UCATS/GloPac aircraft measurement pathways (blue triangles) on 7 and 13 April 2010; and NOAA/Obspack aircraft measurements (green triangles).**

Before evaluating the performance of the optimal analysis against independent observations, we highlight the differences between an optimal and nonoptimal analysis. Accordingly, under similar conditions (e.g., model configuration, inputs), the optimal analysis is compared against another analysis produced with an arbitrary, yet commonly used, non-optimal set of parameters. The globally optimized parameters in April 2010, as described in Sections 6.3 and 6.4, consist of  $f^o = 0.5$ ,  $f^i = 0.45$ ,  $f^q = 0.018$ ,  $L_h = 350$  km,  $L_v$

$= 7\sigma$ , whereas the arbitrary parameters include  $f^o = 1.2$ ,  $f^i = 0.45$ ,  $f^q = 0$ ,  $L_h = 600$  km,  $L_v = 1\sigma$ , for the nonoptimal case. Figure 6.8 demonstrates the comparison of model forecast without data assimilation (blue circles), analysis with optimal parameters (red circles), and analysis with nonoptimal parameters (green circles) against independent TCCON observations. The results indicate that both  $R^2$  and the regression slope are the smallest for the nonoptimal analysis while the mean bias is the largest. It suggests that the nonoptimal analysis generated with a set of arbitrary, but typically used, parameters could result in an agreement even worse than the model against independent measurements. This behaviour was also found for the analysis of  $PM_{2.5}$  with surface observations (Menard and Deshaies-Jacques 2018b). On the contrary, the optimal covariance parameters produce an analysis that maintains a significantly higher consistency with TCCON observations than the model without assimilation. Similar results are obtained with other types of independent observations and a different set of arbitrary (nonoptimal) parameters (see Figure Figure D.1-Figure D.6 in Appendix D). Thus, it could be inferred that the estimation of error covariance parameters that correspond to the optimal solution is essential for PvKF assimilation and most data assimilation schemes that also depend on those input error covariance parameters (e.g., 4D-Var, EnKF).

The GOSAT methane assimilation with optimized parameters is further examined by comparing it with the model forecast against all four types of independent measurements, including TCCON (Figure 6.9a), HIPPO-3 (Figure 6.9b), NOAA/ObsPack aircraft (Figure 6.9c), and UCATS/GloPac (Figure 6.9d) as shown in Figure 6.6. In Figure 6.9, the blue circles (model) represent a pure forecast simulation of H-CMAQ between 2 to 28 April 2010, whereas the red circles (analysis) denote GOSAT assimilation with

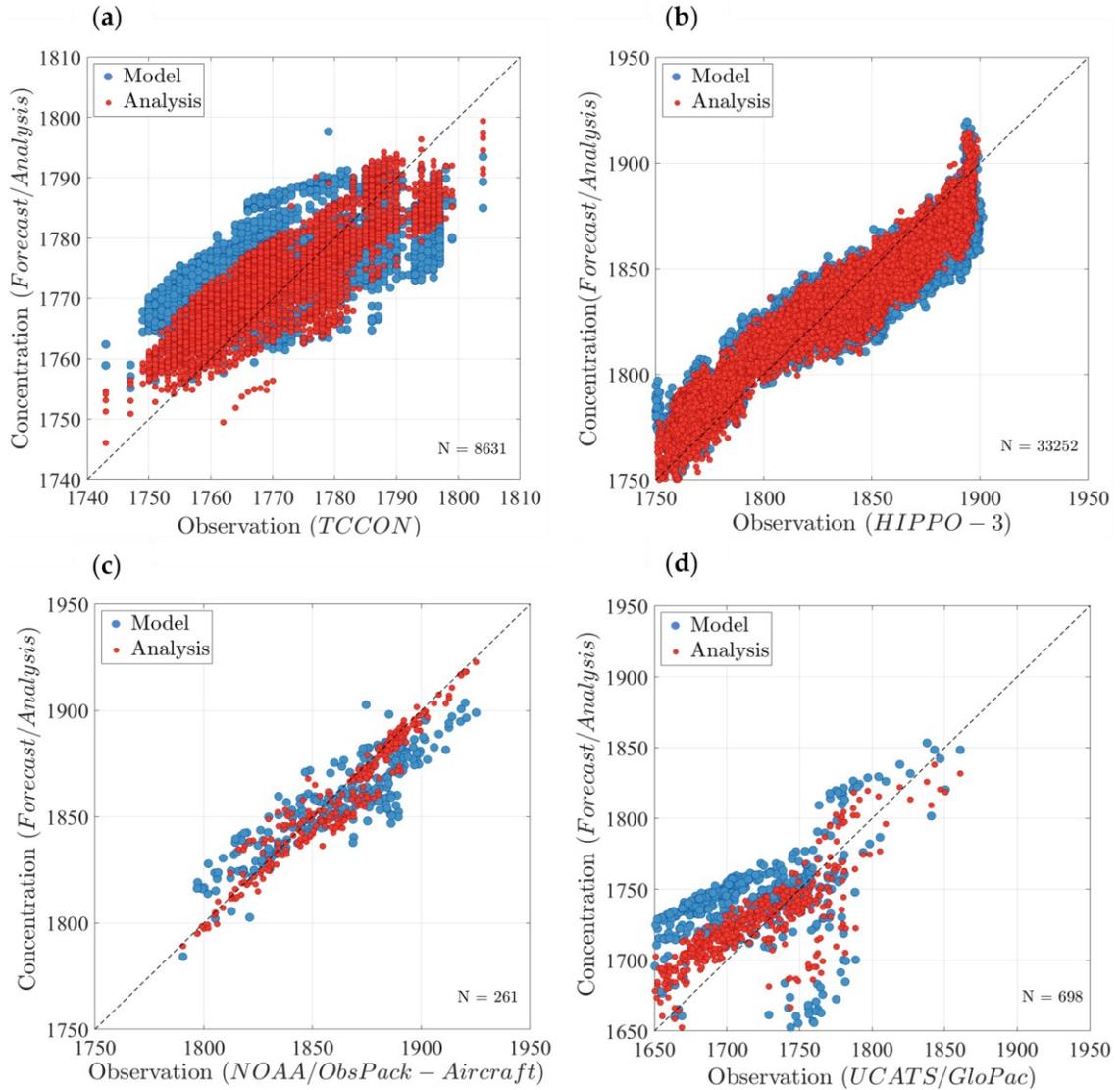
optimized parameters, both of which are sampled at the same measurement location and time and using the same model inputs and configuration.



**Figure 6.8.** (a) Comparison of model forecast (blue circles), (b) analysis with optimal parameters (red circles), and (c) analysis with nonoptimal parameters (green circles) against independent TCCON observations. Optimal parameters of the analysis include  $f^o = 0.5$ ,  $f^i = 0.45$ ,  $f^q = 0.018$ ,  $L_h = 350$  km,  $L_v = 7\sigma$  and the nonoptimal parameters are assumed to be  $f^o = 1.2$ ,  $f^i = 0.45$ ,  $f^q = 0$ ,  $L_h = 600$  km,  $L_v = 1\sigma$ .

The analysis shows an overall and consistent improvement in agreement with all four types of independent measurements, even though the level of improvement is not the same for all of them. Furthermore, among those four observations, analysis at TCCON and NOAA/ObsPac aircraft, which are located over land, shows a relatively better agreement due to higher correlation and a narrower spread of data points. This implies that constructing the analysis using land-only GOSAT data with high retrieval quality reflects a better consistency with independent land measurements. Nevertheless, the analysis maintains a better agreement than the model with HIPPO-3 and UCATS/GloPac, located mainly over the ocean where no GOSAT observations were used. It can be inferred that the evolution of error variance using an advection scheme enhances the assimilation capabilities to maintain a relatively higher consistency with independent observations

within a short period of time, even far from the observations used to generate the analysis. Comparison of the model (M) and analysis (A) statistics for each type of measurement are listed in Table 6.2 and emphasize the benefit of using the assimilation with optimal parameters. The largest improvement of the correlation ( $R^2$ ) along with a reduction of mean bias ( $MB$ ) and standard deviation ( $\sigma$ ) is observed at TCCON locations ( $\Delta R^2 = +39.9\%$  or  $R^2_A/R^2_M = 2.1$ ), followed by NOAA/ObsPack aircraft measurements ( $\Delta R^2 = +13.9\%$  or  $R^2_A/R^2_M = 1.18$ ). These two observations both occur over land and differ in their vertical sampling, where TCCON measures the total column, and NOAA/ObsPack provides point measurements from the mid-troposphere up to the lower stratosphere (Schuldt et al. 2021). Using TCCON, the analysis is capable of detecting the total methane bias, whether due to incorrect surface emissions or modelling errors of unknown origin. In contrast, the bias of the analysis relative to NOAA/ObsPack aircraft can isolate the impact of modelling errors on the analysis since little of the emissions signal reaches aircraft measuring altitude (i.e., several thousand kilometers) within a short period of time (i.e., several weeks). The positive bias of the forecast model with respect to TCCON contrasted with its negative bias relative to NOAA/ObsPack suggests that an excessive amount of surface emissions or high tropospheric methane background is incipient in H-CMAQ, while methane abundance is underestimated in the upper troposphere and lower stratosphere.



**Figure 6.9.** Comparison between the model forecast (blue circles) and analysis with GOSAT assimilation (red circles) sampled at four types of independent measurements, including (a) TCCON, (b) HIPPO-3, (c) NOAA/ObsPack aircraft, and (d) UCATS/GloPac. N denotes the number of observations of each type.

Knowing that the analysis increments are driven by emissions at or near the surface or from modelling errors, we can attempt to qualitatively identify the significance of this influence from each one of these sources. We recall that evaluation with TCCON could show a combined emissions and modelling error effect, while NOAA/ObsPack only

corresponds more closely to the bias originating from the modelling error. The analysis increments statistics at TCCON and NOAA/ObsPack show a 39.9% and 13.9% increase of  $R^2$ , respectively. The mean bias also denotes an improvement of 3.51 ppb with TCCON, and 1.32 ppb with NOAA/ObsPack. It could be inferred that the global influence of the surface emissions is more significant than the modelling error due to a significant improvement of both  $R^2$  and  $MB$  against TCCON rather than NOAA/ObsPack. It is worth mentioning that the model is also less capable of predicting methane concentration over land than over the ocean, mainly due to incorrect model emissions generated from the land surface. This explains the lowest model correlation against TCCON with total column measurements (Table 6.2) compared to the other three observations at higher altitudes and mainly over the ocean. Accordingly, the analysis correlation against TCCON has remained lower than the other three observations; however, we remark the largest correlation improvement in the analysis against TCCON. We discuss later in Section 6.6 the spatial distribution of the analysis increment and error variance reduction. Note that our analysis evaluation with TCCON stations shows a comparable result to the weak-constraint 4D-Var in the global GEOS-Chem data assimilation employed by Stanevich et al. (2021)—Table 1 therein.

Comparing the model and analysis at HIPPO-3 and UCATS/GloPac, both taken over the Pacific Ocean, shows a reduction of 66% and 41% in  $MB$  alongside a 3.7% and 6.8% increase of  $R^2$ , respectively (Table 6.2). The forecast model sampled at HIPPO-3 locations exhibits a negative bias relative to the observations, representing a combined tropospheric and lower stratospheric bias between 0.9 km and 13 km altitude. Addressing the origin of this bias is not a trivial task; however, our analysis successfully removed a

significant portion of that. Given an unbiased initial condition and the small impact of emissions over the HIPPO-3 domain, we relate this bias to modelling error originating primarily from the oxidation of methane with OH in the troposphere (Turner et al. 2019; Zhao et al. 2020a), methane stratospheric transport (Zhang et al. 2021), and methane oxidation with chlorine radical over the Oceans (Sherwen et al. 2016). The model and analysis evaluation at UCATS/GloPac aircraft measurement locations reveals a large positive bias that contradicts the earlier comparison results with HIPPO-3 and NOAA/ObsPac aircraft measurements. This disagreement likely arises from the limited vertical structure of H-CMAQ extended up to 50 hPa (~20 km), which cannot entirely and adequately cover all the measurements. In fact, GloPac aircraft is mainly flying over the lower stratosphere (Naftel 2009; Hintsa et al. 2021), for which quite a few measurements are recorded beyond or at the few top layers of H-CMAQ. Therefore, in addition to the sources of modelling error mentioned earlier, inter/extrapolation of model concentration at these measurements' height adds up to the positive bias, which is unrealistic. Nonetheless, our analysis could remove a noticeable part of this bias, whether from a physical model bias (e.g., part of the stratospheric bias) or an artificial bias (e.g., numerical inter/extrapolation).

Several studies provide in-depth investigations of the possible causes of methane model bias in a global atmospheric model (e.g., GEOS-Chem), such as stratospheric bias, weakening of vertical transport due to a coarse model resolution (Stanevich et al. 2020), OH burden in the chemistry model. Besides those, H-CMAQ could suffer from bias due to inaccurate initial conditions, insufficient lower top of the model boundary conditions, and interhemispheric exchange of methane.

**Table 6.2. Evaluation of model forecast (M) and analysis (A) of GOSAT assimilation against TCCON, HIPPO-3, UCATS/GloPac, and NOAA/ObsPack measurements. Comparison between mean bias ( $MB$ ), standard deviation  $\sigma$ , coefficient of determination  $R^2$ , and the linear regression line, using all measurements of each type of observation in April 2010**

Observations	Type	$MB$ (ppb)		$\sigma$ (ppb)		$R^2$		Regression Line	
		M	A	M	A	M	A	M	A
TCCON	Total column	6.09	2.58	8.39	5.45	35.4	75.3	$0.3x + 1229$	$0.61x + 605$
HIPPO-3	Aircraft	-6.50	-2.21	15.29	12.95	92.1	95.8	$0.68x + 573$	$0.76x + 429$
UCATS/GloPac	Aircraft	59.1	34.8	82.3	59.2	90.6	97.4	$0.55x + 774$	$0.67x + 559$
NOAA/ObsPack	Aircraft	-4.61	-3.29	14.22	7.66	79.1	93.0	$0.64x + 654$	$0.91x + 148$

Uncovering the origins of the model bias and distinguishing it from surface emissions bias requires an exhaustive forward and/or adjoint sensitivity analysis besides identifying all the possible sources of uncertainty, which is out of the scope of this study. Still, the discrepancy between the optimal analysis and the model (i.e., analysis increment) together with the uncertainty reduction can reveal useful information to address the key sources of bias and uncertainty associated with methane simulation. We illustrate these in Section 6.6, both in the H-CMAQ and GOSAT observation space.

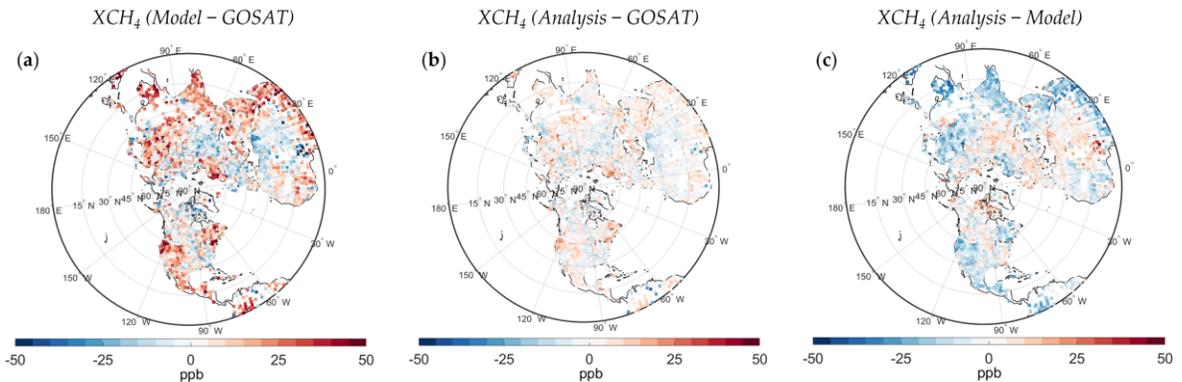
## 6.6 Characteristics of Analysis and Error Variances

This section examines the performance of our analysis with optimal parameters in capturing the spatial structure of bias and uncertainty reduction across the domain. First, we compare one month of analysis against the model forecast in the observation space, where model and analysis are sampled at GOSAT times and locations with an identical observation operator. A difference between model/analysis and GOSAT is shown in Figure

6.10. Comparing Figure 6.10a,b indicates that a significant portion of biases in the model from different parts of the world, including a positive bias at eastern China, India, eastern parts of Russia, South Asia, tropical and eastern Africa, northern Europe, eastern Canada, and the western U.S. are substantially removed by the analysis. At the same time, the analysis also resolves the negative bias in areas such as northern Africa, eastern Europe, and the southeastern U.S. Accordingly, the analysis demonstrates an 80% and 65% adjustment over the areas with the largest positive and negative model bias, respectively. A direct comparison between the analysis and the model is also shown in Figure 6.10c, indicating the consistency of the analysis with GOSAT retrievals to correct the spatial variation of the biases.

The methane spatial bias in many studies is primarily attributed to its prior emissions allocation and is resolved using inverse modelling analysis (Maasakkers et al. 2019; Janardanan et al. 2020; Lu et al. 2021; Maasakkers et al. 2021; Zhang et al. 2021). The bias can also arise from the limitation of the chemical transport model to realistically simulate atmospheric methane, also known as the bias due to model error (Saad et al. 2016; Yu et al. 2018; Stanevich et al. 2020; Stanevich et al. 2021). The optimal analysis described in this study corrects for these biases and assists in identifying the origins of these biases. Continuous real-time estimation of assimilated methane concentration alongside its error variance reduction enables a direct and reliable comparison with the model forecast, which reveals valuable information. It is common to consider over a short period of time (e.g., one month) that the analysis increment at the surface reflects the correction needed for improving the prior surface emissions while the increments at the higher elevations (e.g., <500 hPa) is due to the combined effect originated from model error and inaccurate

emissions. Furthermore, the analysis error variance reduction computed here by PvKF represents the analysis uncertainty, thus is an indicator of assimilation accuracy with respect to the true state.



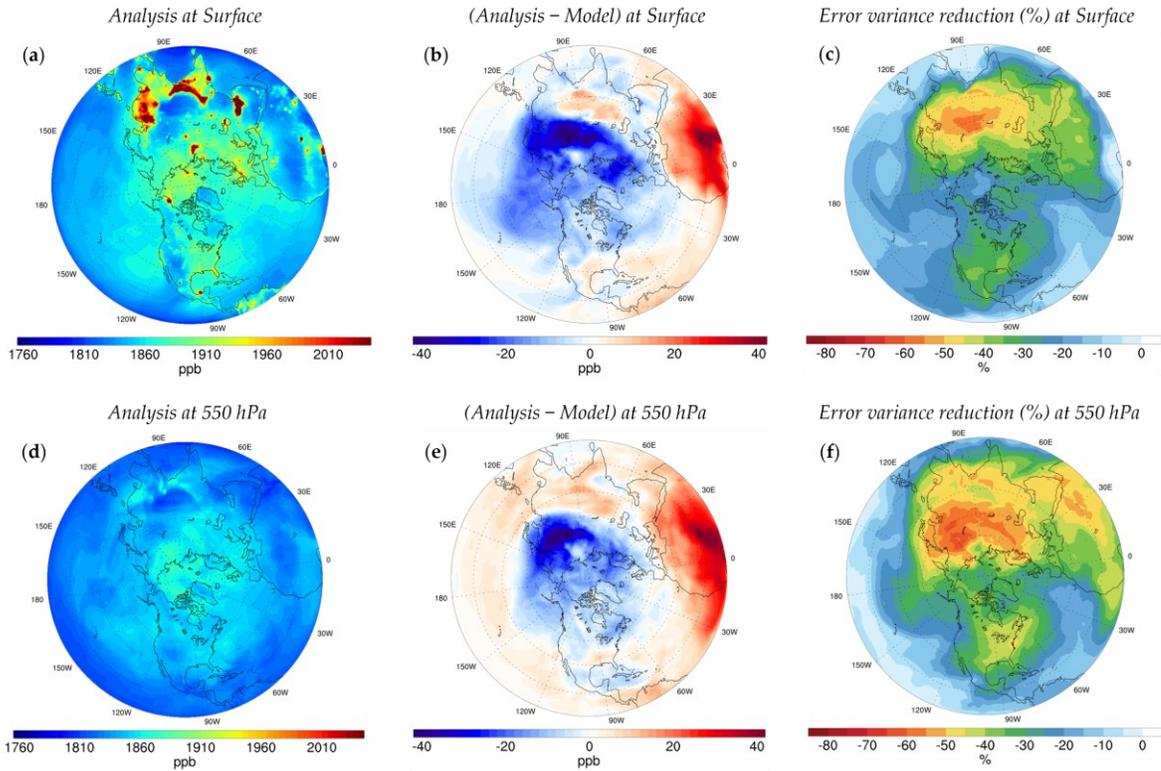
**Figure 6.10.** Differences between (a) model and GOSAT, (b) analysis and GOSAT, and (c) analysis and model in the observation space in April 2010.

The monthly average of the analysis, analysis increments (i.e., Analysis – Model), and the analysis error variance reduction at the surface and 550 hPa are shown in Figure 6.11. We found that the largest downward correction occurs over Russia, while the largest upward correction is over tropical Africa (Figure 6.11b). These results are consistent overall with recent methane inverse studies; for example, Wang et al. (2019) considered a 200% downwards adjustment to the prior emissions over Russia from EDGAR. Zhang et al. (2021) estimated Russia’s anthropogenic emissions to be about 59% of the prior value, mainly attributed to the oil-and-gas sector. In the same study, a large upward correction in the prior inventory is also reported over East and tropical Africa, mainly due to the growth of livestock and wetland emissions. Our analysis at the surface also suggests a positive bias over India, Japan, eastern China, Europe, high latitudes, western U.S. and all over Canada, whereas a comparatively negative bias over the South and Southeastern U.S., northern regions of South America, and Central Asia. Note that several factors, such as the

discrepancy between emission inventories, the atmospheric model, and the temporal variations, can explain slight differences between studies. In addition, the correction near the Equator in our hemispheric assimilation could be too diffusive to constrain the spatial pattern of the bias on regions next to the Equator; hence, all plots in Figure 6.10 and Figure 6.11 are shown above 5 °N.

Besides those regions with maximum and minimum values of the analysis increment at the surface, the analysis increment at 550 hPa (Figure 6.11e) suggests a similar scale and pattern of correction in most regions over North America and Africa. However, there is a noticeable discrepancy over the lower and mid-latitude Pacific Ocean, East Asia, India, and the Middle East, where the analysis suggests a positive correction in the mid-troposphere rather than the negative correction captured at the surface. As suggested by Stanevich et al. (2021), we consider the vertical dipole structures in our analysis increment as the cause of modelling error in the bias, whereas a monotonic correction represents the impact of the inaccurate emissions on the bias. Another region of interest is the eastern part of the Atlantic Ocean, where the analysis increment is significantly larger at 550 hPa than the surface, yet with the same correction sign (Figure 6.11e). That possibly indicates the presence of model error over those regions, where convection does not lift up enough methane to the mid and upper troposphere in H-CMAQ, partly due to a persistent high-pressure and low-pressure system in the African continent. Such a pattern is also observed in previous global studies with different chemistry transport models (Arellano et al. 2006; Dyer et al. 2017; Stanevich et al. 2021).

As a distinctive feature, the analysis error variance is explicitly computed by the PvKF assimilation. It represents the estimation uncertainty with respect to the true state, thus inferring how reliable the assimilation results (i.e., 3D optimal analysis) are.



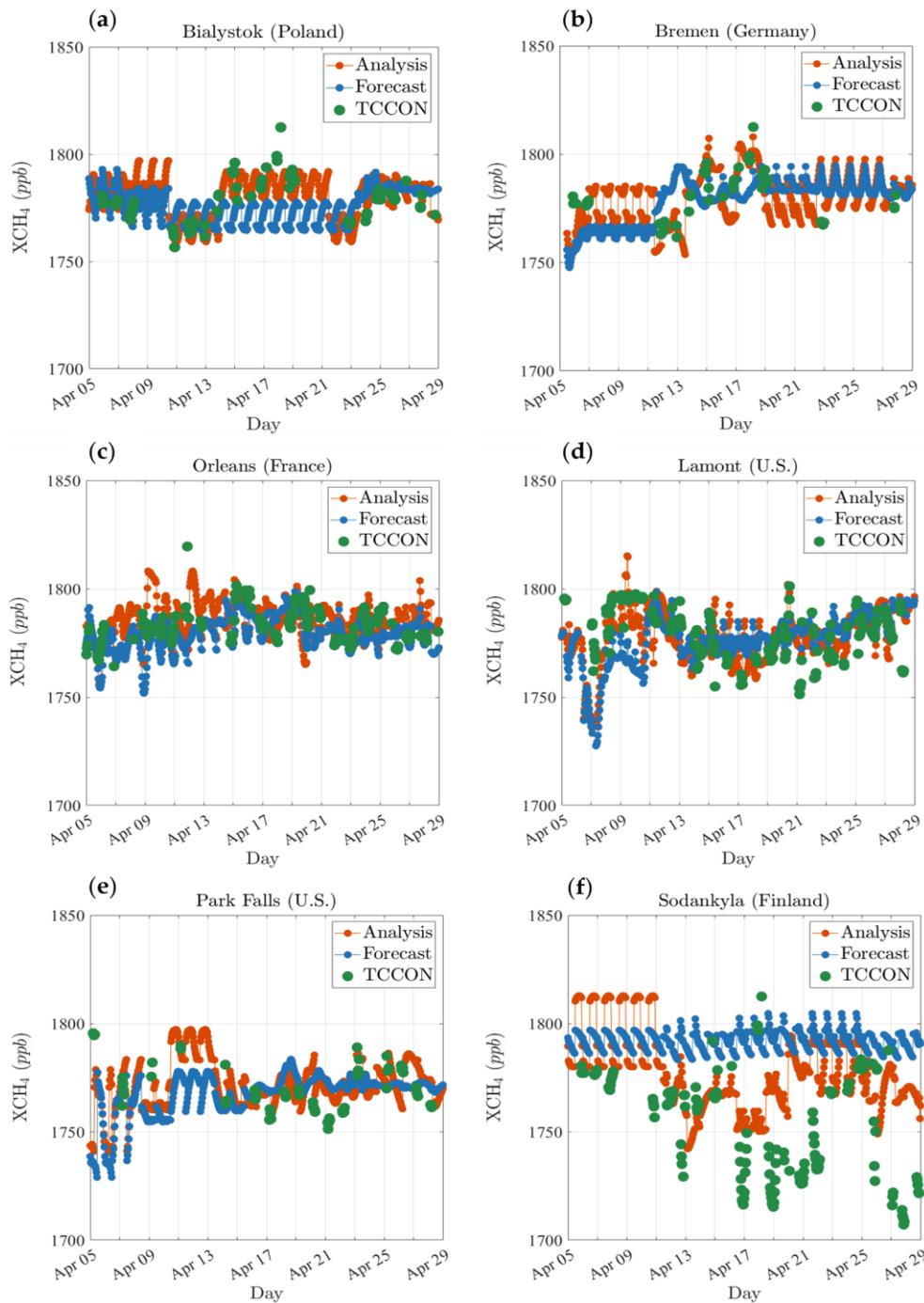
**Figure 6.11.** (a) Monthly analysis at the surface and (d) at 550 hPa; (b) analysis increment at the surface and (e) at 550 hPa; (c) error variance reduction at the surface and (f) at 550 hPa.

The spatial distribution of the analysis increment is mainly influenced by the innovation (Observation – Model), whereas the distribution of the analysis error variance depends largely on the observation density and its error. This discrepancy arises from the fact that the analysis of the second moment (i.e., error covariances) is decoupled from the first moment (i.e., mean state) in linear Kalman filtering estimation theory (Kalman 1960). In the upper troposphere (i.e., 550 hPa), the spatial pattern of error variance reduction (Figure 6.11f) is more consistent with the analysis increment (Figure 6.11e), such that a larger reduction of error variances occurs over the same regions with larger analysis increment,

suggesting that those corrections are fairly reliable. In contrast, the spatial distribution of the error variance reduction at the surface (Figure 6.11c) is not always following the analysis increment spatial pattern (Figure 6.11b). For instance, the analysis over North/East Africa shows the largest increase while the error variance reduction is relatively small, indicating that the analysis correction over those regions is less reliable. It could be implicitly inferred that conducting emissions inversion over these areas would entail higher uncertainty in corrected emissions.

Since the analysis provides us with an hourly real-time assimilated methane with GOSAT, it is interesting to examine its time series and compare it with equivalent model values. Figure 6.12 shows the analysis and the model values projected to the six TCCON observation sites (red circles in Figure 6.7), using the TCCON averaging kernels and a priori. In the time series plots, the green dots denote the TCCON observations, while the red and blue dots represent the analysis and model values, respectively. Note that TCCON averaging kernels are fairly smooth and vary only slightly with solar zenith angles and pressure (Wunch et al. 2010), so that we averaged those values of two consecutive measurements in order to provide a continuous model and analysis over time.

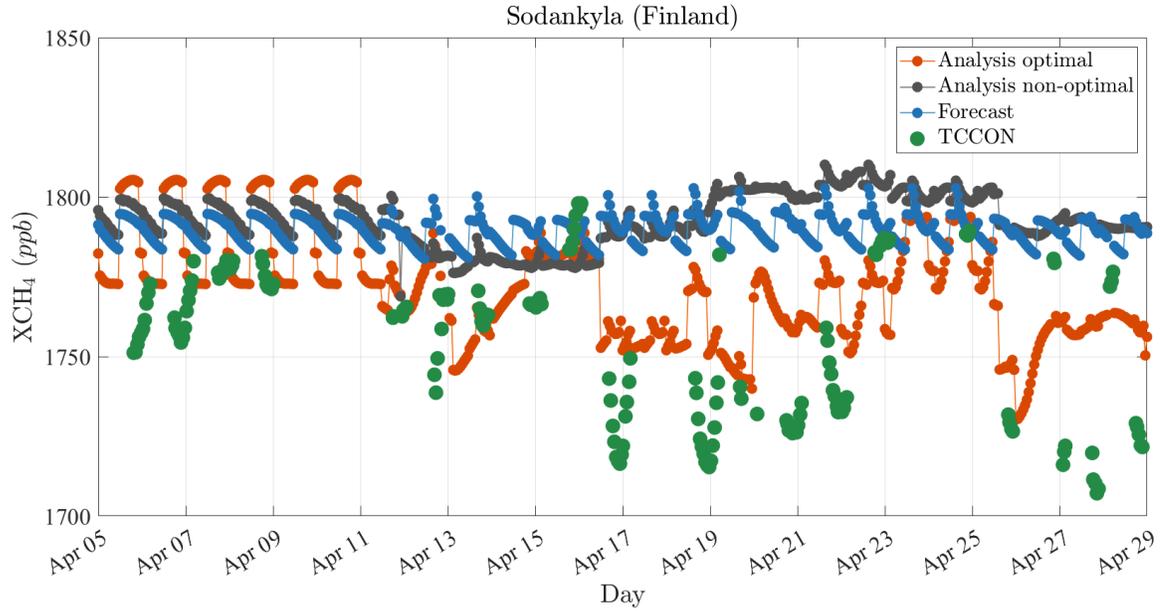
We showed earlier (Figure 6.9) that the analysis with optimal parameters maintains a significantly better agreement with TCCON. The spatial distribution of the optimal analysis is also consistent with GOSAT retrievals (Figure 6.10) and agrees with the recent findings of methane inversion studies.



**Figure 6.12.** Comparison of the time series of the analysis (red dots) and the model (blue dots) against TCCON observations (green dots). Six TCCON observation sites (Figure 6.7—red circles) include (a) Bialystok (53.23 °N, 23.03 °E), (b) Bremen (53.10 °N, 8.85 °E), (c) Orleans (47.97 °N, 2.11 °E), (d) Lamont (36.60 °N, 97.48 °W), (e) Park Falls (45.95 °N, 90.27 °W), (f) Sodankylä (67.37 °N, 26.63 °E).

The times series plots (Figure 6.12) indicate that the analysis also follows the temporal variation of independent measurements, whereas the model forecast is almost insensitive to these variations. For example, at the Sodankylä observation site (Figure 6.12f), TCCON shows a consistent temporal pattern measuring between 1710 to 1770 ppb from 13 to 29 April 2010. The model shows no sign of agreement with TCCON over the same period; however, the analysis attempts to correct the model toward these independent measurements depending on the feedback provided by GOSAT observation for methane assimilation.

We should note that the analyses in Figure 6.12 are optimized with the error parameters obtained in Sections 6.3 and 6.4. Similar to the evaluation of optimal and nonoptimal analysis with TCCON (Figure 6.8), we compare the time series of those two cases (using the same set of optimal and nonoptimal error covariance parameters) at Sodankylä. As shown in Figure 6.13, an analysis obtained with a set of nonoptimal, yet commonly used, error covariance parameters (black dots) can result in an even worse temporal agreement with TCCON (See Figure D.7 and Figure D.8 for more examples). Therefore, we conclude once more that obtaining the optimal error covariance parameters is essential to have a reliable analysis that improves the methane representation both spatially and temporally.



**Figure 6.13.** Comparison of model forecast (blue dots), analysis with optimal parameters (red dots), analysis with nonoptimal parameters (black dots) against Sodankylä (67.37 °N, 26.63 °E) TCCON observations (green dots). Optimal parameters of the analysis include  $f^o = 0.5$ ,  $f^i = 0.45$ ,  $f^q = 0.018$ ,  $L_h = 350$  km,  $L_v = 7\sigma$  and the nonoptimal parameters are assumed to be  $f^o = 1.2$ ,  $f^i = 0.45$ ,  $f^q = 0$ ,  $L_h = 600$  km,  $L_v = 1\sigma$ .

## 6.7 Conclusions and Summary

We employed the PvKF data assimilation scheme developed in Chapter 6 to the GOSAT observations in the hemispheric CMAQ domain to produce a high-quality (i.e., near-optimal) analysis of atmospheric methane concentration. Although most assimilation schemes such as PvKF are derived from minimum variance estimation theory, it does not mean that the analysis produced using real observations will be the best one. The input error covariances (e.g., observation error covariance, model error variance and correlation lengths) also need to be realistic. In this chapter, we used several techniques to obtain accurate error covariances and show, using independent observations and cross-validation, that these realistic error covariances indeed achieve better analyses. For example,  $R^2$  of

independent TCCON observations with analysis is 0.29 when using inaccurate error covariances and increases to 0.75 with optimal error covariances.

In this work, the cross-validation methodology developed originally for in situ observations (Menard and Deshaies-Jacques 2018a) has been extended for satellite observations using observation thinning. In fact, we have verified that thinned satellite observations (10 km apart) give the same optimal parameter values as those obtained with independent TCCON observations. Our procedure is not to estimate all elements of covariance matrices (since it requires an extremely large number of data or realizations), but rather a few key global parameters of these covariances. We have estimated the horizontal and vertical correlation lengths of the background error covariance and the observation error variance. It is important to note that the cross-validation estimation technique is the desired approach since it does not assume that the analysis is (already) optimal. Thus, the covariance parameter optimization results in realistic error covariances and near-optimal analyses. We recall that optimal analyses and innovation covariance consistency are the necessary and sufficient conditions to estimate true observation and background error covariances (Menard 2016). We also found that these error parameters are nearly insensitive to the optimization time window, so that they are valid for the entire month of assimilation of this study. On the contrary, the cross-validation is not properly applicable to estimate the model and initial error variances due to a noticeable time-varying influence on the optimization window. We demonstrate a normalized variance matching diagnostic approach to tune these parameters while maintaining a stable analysis error variance. Finally, we note that the optimization approach of this study is applicable to any

other assimilation experiment but may result in different correlation length scales with different observation densities.

Once the optimal analysis using GOSAT is obtained, we evaluate the analyses against different types of independent observations (TCCON, NOAA/ObsPack, UCATS/GloPac, HIPPO-3). We show its superiority against analysis produced using reasonable but arbitrary values of covariance parameters. Furthermore, the comparison between the model (with no assimilation) and optimal analysis against TCCON and NOAA/ObsPack suggest an overestimation of surface methane (most likely due to emissions) and an underestimation of the upper-tropospheric model methane. Remarkably, our PvKF assimilation is capable of correcting both of these biases regardless of their origin. The correction's statistics due to assimilation are more significant with respect to TCCON, which is a total column measurement, than with respect to NOAA/ObsPack, which only samples the upper troposphere. This suggests that the assimilation of GOSAT makes the larger correction near the surface, where presumably there are larger errors than in the upper troposphere. The assimilation of GOSAT also seems capable of addressing the known model problem in the horizontal direction. For instance, our analysis increment suggests the largest negative correction over Russia, whereas the largest positive correction over tropical Africa. This is coherent with the recent methane inverse studies (Stanevich et al. 2021; Zhang et al. 2021). Furthermore, we have a large reduction of error variance over those regions, suggesting that those corrections are reliable. In addition, the PvKF assimilation that offers a time-continuous series of analyses enables us to investigate the temporal behaviour of the model biases. Our results show a better temporal agreement between analysis and TCCON than between model (no assimilation) and TCCON. In

summary, it appears that the optimal assimilation of GOSAT is able to correct for the vertical, horizontal, and temporal model errors, as revealed by comparison with independent observations.

Because this assimilation system provides the state estimate and its error, it is suitable to couple it with another estimation problem such as a limited domain data assimilation, source inversion, etc. In theory, the PvKF formalism also seems to be applicable to the joint assimilation-inversion (state-source estimation) problem, but we need to demonstrate that in a realistic context, which will be considered in a future manuscript. This method could also be applied to chemical species with a shorter lifetime, knowing that a smaller fraction of the total forecast error variance is explained by the advection of error variance. In this case, more (unexpected) error variance is carried out with a stationary model error, which emphasizes a more sophisticated design of the model error covariance. Therefore, a shorter lifetime of a chemical species may induce a weaker performance of the PvKF assimilation and make it behave closer to an Optimal Interpolation (OI). Another application of the PvKF assimilation that could be investigated in future studies is to use the analysis over the oceans to develop the bias correction of GOSAT over ocean observations.

## **Chapter 7: Use of Assimilation Analysis in 4D-Var Source Inversion: Observing System Simulation Experiments (OSSEs) with GOSAT Methane and Hemispheric CMAQ**

We explore how novel parametric variance Kalman filter (PvKF) assimilation is used in conjunction with a four-dimensional variational (4D-Var) inversion system to improve methane source estimation. PvKF assimilation is a cost-efficient scheme that provides an optimal analysis field and its uncertainties (error variance) while allowing for the propagation of error variance in a dynamically coherent manner. The main objective of this study is to evaluate the contribution of PvKF optimal analysis and its error covariance evolution on the performance of a 4D-Var inversion, particularly when the state concentrations are not perfectly known. Our experiments are conducted using the hemispheric version of CMAQ. We perform a set of observing system simulation experiments (OSSEs) with PvKF assimilation along with the CMAQ adjoint model and compare four different types of inversion cost functions to determine the assimilation impact on recovering true emissions. For practical purposes, inversions neither assume a perfect model nor a true initial state. The effect of the initial field of analysis, forecast of analysis error covariance, and model transport error is individually examined through modified 4D-Var cost functions. Different perturbations of prior emissions from four main source categories, including agriculture, energy, waste, and wetlands, are considered. Our results show that using PvKF analysis instead of the model forecast to initialize the inversion, besides eliminating the need for a fairly extended model spin-up, improves posterior emissions estimate (~35% reduction in NMB) across the domain. Furthermore, propagation of analysis error covariance using PvKF formulation tends to retain the effect

of model correlation structures in the observation space and, thus, results in lowering the variance of posterior emissions in most cases (~50% reduction in NME). Sectoral perturbation results also suggest that accounting for the analysis error covariances and model transport errors both improve the emission estimates of sectors spread over a larger area (e.g., agriculture). Whereas the impact of including the (initial) analysis field is more substantial for constraining local or point sources (e.g., energy). Finally, we found that using the optimal analysis, contrary to the error covariances in the inversion, can significantly reduce the computational cost (~ one-third of a conventional 4D-Var).

## 7.1 Introduction

Methane ( $\text{CH}_4$ ) is a critical atmospheric component in both climate and air quality contexts (Staniaszek et al. 2022). Over the past decade, a continuous increase in global methane concentrations has drawn urgent attention to quantifying and reducing methane emissions (Saunois et al. 2020; Dlugokencky 2022; Nisbet et al. 2022; Worden et al. 2022). Methane emissions are derived from either the bottom-up or top-down estimation methods. Despite containing a substantial amount of data, bottom-up inventories suffer from two weaknesses: (i) significant uncertainties due to inaccurate or missing information and (ii) not being constrained by atmospheric observations to retain a closed-form global budget (Brasseur and Jacob 2017; Saunois et al. 2020; Minx et al. 2021). Those eventually can hinder progress toward methane mitigation policies (Ganesan et al. 2019). Top-down inventories, also known as atmospheric inversion, are less prone to issues of the bottom-up method due to incorporating information from observations. In particular, observations and a chemical transport model (CTM) are used together to make corrections to the prior bottom-up emissions. Providing adequate information from the observation network along

with an accurate CTM maintains a top-down estimation that is less dependent on the prior, resulting in estimated emissions with lower uncertainties. Satellite observations, due to their higher density and global spatial coverage, have been used extensively over the past decades to infer methane emissions on different scales (Palmer et al. 2021; Jacob et al. 2022).

Atmospheric inversions play a key role in evaluating and improving bottom-up estimation but carry their own limitations for constraining methane emissions. For example, global or regional inversions are usually incapable of resolving emissions of small source sectors, such as non-wetland emissions, due to their minimal sensitivity to atmospheric observations (Dlugokencky et al. 2011; Saunio et al. 2020). Inversions also depend on the choice of the prior emissions, particularly over areas where the observation constraints are limited, or observations are not precisely determined (Bergamaschi et al. 2018; Maasakkers et al. 2021). In addition, significant uncertainties in the inversion are attributed to satellite measurements. Previous studies indicated a large discrepancy between satellite retrievals (e.g., from Scanning Imaging Absorption Spectrometer for Atmospheric Chartography, SCIAMACHY) and in situ measurements (Frankenberg et al. 2011; Wecht et al. 2014; Houweling et al. 2017). Furthermore, inversions rely on a CTM to translate the emissions signal into the observation space. Although advancements in remote sensing and emissions inventories every year provide us with more accurate satellite measurements and more reliable prior emissions, potential errors in CTM still exist and may reflect on the inversion results (Prather et al. 2008; Locatelli et al. 2015).

Many earlier methane inverse modelling studies assume that the CTM is perfect (Kopacz et al. 2010; Turner et al. 2018; Janardanan et al. 2020; Lu et al. 2022). However,

it has been shown that running an inversion with different CTMs can exert a tangible discrepancy in emissions estimates (Locatelli et al. 2013; Locatelli et al. 2015). Transport errors such as those originating from meteorological fields, model advection parameterization, and model spatial and temporal resolutions are identified as crucial aspects contributing to the CTM's error, particularly on a short timescale (Locatelli et al. 2015; Saad et al. 2016; Stanevich et al. 2020; 2021). One promising way to address those errors is to simultaneously estimate methane emissions and model transport error, such as through weak-constraint 4D-Var (Tremolet 2006; 2007; Stanevich et al. 2021). However, besides substantial computational cost, such a method applied for methane estimation does not account for the entire model error in the state as they may not necessarily originate due to transport (Bousserez et al. 2016; Zhang et al. 2018). For example, errors in the chemistry, initial, and boundary conditions are among those that are not addressed through methane weak-constraint 4D-Var.

A reliable emissions inversion result depends on maintaining a precise and realistic state estimation, whether it is used as the initial state (or background) or boundary conditions in a limited domain. One way to fulfill that requirement is to jointly estimate the state and emissions (Elbern et al. 2007). A joint estimation typically entails a substantial extra computational cost mainly due to resolving posterior errors for both concentrations and emissions. Furthermore, the procedure may not lead to convergence of emissions (Wecht et al. 2014). This is likely due to the fact that the impact of the initial or boundary conditions on methane concentrations is much larger than the emissions, although with less variability. From an estimation point of view, the emissions signal in the inversion is masked by a larger influence of the (inflow or initial) concentrations. Thus, a joint

estimation usually entails reducing the state uncertainties to a level lower than the emissions signal, which is hard to achieve.

Chemical data assimilation may be used to separately estimate the model state concentrations and its error statistics. It is a promising method that is also used to deal with the issue of the state in inverse modelling. Previous studies assumed that the initial (or boundary) concentrations provided by data assimilation are perfectly known; hence, potential uncertainties from the state cannot be taken into account for performing an inversion (Basu et al. 2013; Deng et al. 2014). On the other hand, obtaining the state and its error statistics using conventional data assimilation approaches, such as 4D-Var and EnKF assimilations, is a task with high computational cost (Skachko et al. 2014; Skachko et al. 2016; Pannekoucke et al. 2021; Voshtani et al. 2022a). We have used a newly developed assimilation system known as parametric variance Kalman filter (PvKF) that provides a computationally efficient way of estimating methane state along with its error variance (Voshtani et al. 2022a; 2022b).

In this chapter, we frame a new source estimation system that links PvKF assimilation to a 4D-Var inversion technique, with the objective of examining the effect of the state estimation on constraining emissions. In this framework, not only can we change the first guess concentrations to the cost function, but also explicitly include the estimate of error covariances of concentrations for performing an inversion. PvKF formulation allows for dynamically propagating those errors while not relying on a perfect model transport assumption (Voshtani et al. 2022a). This enables our modified inversion formulation to account for model spatial correlation structures during the inversion, which is typically missed in methane inversions studies since only diagonal observational error

covariances are considered (Turner et al. 2015; Bousserez et al. 2016; Bousserez and Henze 2018; Turner et al. 2019; Maasakkers et al. 2021; Stanevich et al. 2021; Yu et al. 2021; Zhang et al. 2021).

We conduct observing system simulation experiments (OSSEs) using the hemispheric Community Multiscale Air Quality (CMAQ) (Byun & Schere, 2006) model and Greenhouse Gases Observing Satellite (GOSAT) observations to optimize monthly methane emissions. Our inversions include three major anthropogenic and one natural category of methane emissions. Using different initial conditions and different formulations of a 4D-Var cost function, we are able to determine three individual effects on the source inversion, including (i) the effect of the optimal initial analysis field, (ii) the model propagated analysis error covariance, and (iii) the approximated transport error. We perform different perturbations of prior emissions, aiming to address the limitation of a typical 4D-Var inversion that relies on perfect state assumptions (i.e., initial and CTM) as well as a diagonal observation error covariance. In addition, the impacts of those cost function configurations on the inversion of individual source sectors are further demonstrated in our analyses.

The organization of this chapter is as follows. Section 7.2 overviews the GOSAT observation data, the CMAQ chemical transport model, and the prior methane emissions. In Section 7.3, we first present an overview of the PvKF assimilation and the standard 4D-Var inversion; we then describe the formulation and assumptions for linking PvKF to 4D-Var. Section 7.4 details the OSSE experiments. In Sections 7.5 and 7.6, we discuss our OSSE results and provide implications for our proposed source estimation system. Conclusions drawn from this study are provided in Section 7.7.

## 7.2 Background

Every assimilation and inversion system (including PvKF assimilation and 4D-Var inversion herein) requires two categories of information: (i) a chemical transport model derived from a set of inputs (or parameters) to predict the evolution of the model state (i.e., concentrations) and (ii) a set of observations that monitor the change of concentrations. In the assimilation context, the model is combined with observations to obtain an improved state prediction (or methane concentrations), while in the inversion, the model is typically used to provide a linkage between model inputs (e.g., emissions) and observed concentrations, aiming to better constrain those inputs. Here, observations from GOSAT satellite retrievals together with the hemispheric CMAQ model, are used in both PvKF assimilation and 4D-Var inversion.

Many 4D-Var methane inversions in the past assumed that the initial state concentrations were perfectly known (Bergamaschi et al. 2013; Cressot et al. 2014; Deng et al. 2014; Alexe et al. 2015; Bergamaschi et al. 2018); in addition, a perfect CTM to simulate the methane in the atmosphere is considered. In both cases, the uncertainty of the state is assumed to be negligible and thus does not play a role in the inversion process. Although for an extended period of model integration (e.g., more than six months), those assumptions might be acceptable due to the homogeneity of the background methane concentrations (Turner et al. 2015), the errors in the state, such as transport error, for a short period of inversion (e.g., one month) are considerable and thus exert a large impact on source estimation (Stanevich et al. 2021). In order to proceed to more highly resolved emissions inversions, an evolution in methodology is required that accounts for errors in the state.

In this study, we seek to examine the capability of an alternative 4D-Var inversion formalism, which not only depends on the initial field but also relies on the model-propagated uncertainties and the approximated model transport errors. The method section (Section 7.3) demonstrates how the PvKF assimilation framework is linked to a 4D-Var inversion system in order to address those state characteristics for source estimation. Before that, in this section, an overview of GOSAT observations, the CMAQ model and its prior emissions used in both PvKF and 4D-Var systems are provided as follows.

### **7.2.1 Satellite Observations**

GOSAT, launched in January 2009 by the Japanese Space Agency (JAXA) (Kuze et al. 2009), is in a Sun-synchronous orbit at an altitude of 666 km with a 3-day revisit time. The primary goal of GOSAT is to monitor the abundance of greenhouse gas, including atmospheric methane, globally. Because of the instrument's global coverage, reasonable spatiotemporal resolution, and acceptable near-surface sensitivities, the assimilation of methane and inverse modelling of its sources and sinks using GOSAT observations is desirable. GOSAT retrievals provide a column-average dry-mole fraction of methane that corresponds to the methane average volume mixing ratio (VMR) of a partial column atmosphere. Methane VMRs and corresponding standard deviations are obtained by performing a retrieval algorithm on the radiance spectrum. We use GOSAT proxy products from the retrieval algorithm developed at the Netherlands Institute for Space Research (SRON) and Karlsruhe Institute for Technology (KIT) (Butz et al. 2011), available through the ESA GHG-CCI initiative, <https://climate.esa.int/en/projects/ghgs/> (Buchwitz et al. 2017).

Since this study only performs OSSE experiments (see Section 7.4), we do not use actual GOSAT retrievals (or VMR) but simulated ones for methane inversion. Although simulated observations depend on the model forecast, they require supplementary products of the retrievals such as column-average kernels and vertical pressure weights to compute the equivalent value to VMR at observations time and location (Butz et al. 2011; Buchwitz et al. 2017). That supplementary information, however, is related to the actual observations. Thus, for the consistency of the information content (or the number of retrievals) between actual and simulated observations, we only account for those data locations where corresponding actual observations met the quality control. Our quality control of GOSAT consists of removing outliers whose departure from the global mean of the methane observations is three times greater than the standard deviation. As a result, in total, 11,489 simulated GOSAT observations over land are used for the inversion between 1 to 30 April 2010, and 6,173 land-only actual GOSAT observations are provided for the assimilation between 16 to 31 March 2010.

### **7.2.2 Chemical Transport Model and Methane Emissions**

The forecast of the PvKF assimilation and the forward model of the 4D-Var inversion system both rely on a CTM, for which the CMAQ model is used here. CMAQ is a regional air quality model developed by the U.S. Environmental Protection Agency (EPA) (Byun and Schere 2006). As we seek to estimate methane across the Northern Hemisphere in this study, we use the hemispheric version of CMAQ, which has  $187 \times 187$  grid cells horizontally with a 108 km grid spacing and 44 vertical layers from the surface to the model top at 50 hPa. Contrary to the regional model, hemispheric CMAQ provides an extended and finer vertical resolution above the boundary layer to better support long-

distance transport suitable for long-lived species (Mathur et al. 2017). For our hemispheric modelling of methane, the initial conditions are obtained from global modelling and measurements data, which is demonstrated in Voshtani et al. (2022a) as based on Olsen et al. (2013). The same process computes the lateral boundary while it is assumed to be time-invariant. Following Voshtani et al. (2022b), we also exclude a buffer zone below 5° N to minimize the influence of the boundary conditions on domain concentrations for a month of assimilation/inversion.

The primary sink process of methane is oxidation with hydroxyl radical through the chemical reaction  $\text{CH}_4 + \text{OH} \rightarrow \text{CH}_3 + \text{H}_2\text{O}$ . In the hemispheric simulation, we consider including reactive methane in the gas-phase chemistry of CMAQ v5, which is based on the CB05 chemical mechanism (CMAQ tutorials 2021). In the PvKF assimilation, the propagation of error variance is treated as a chemically inert tracer in CMAQ and follows an advection-only transport scheme. The detailed configuration of reactive methane and error variance evolution with CMAQ is illustrated in Voshtani et al. (2022a). To perform a 4D-Var inversion, besides the CMAQ model, we integrate the adjoint of CMAQ (Hakami et al. 2007; Zhao et al. 2020) based on the same version and chemical mechanism used in CMAQ. The adjoint of CMAQ has been validated previously and used in various inversion studies (Turner et al. 2015b; Chen et al. 2021).

Methane emissions implemented in CMAQ are generally derived from bottom-up inventories of two categories: anthropogenic sources (~60%) and natural sources (~40%). 4D-Var inversions usually require a first guess estimate of emissions (also known as prior emissions), which is provided by the bottom-up inventories. The Emission Database for Global Atmospheric Research (EDGAR) (Crippa et al. 2020) inventory is frequently used

to provide prior anthropogenic methane emissions at  $0.1^\circ \times 0.1^\circ$  spatial and monthly-yearly temporal resolution for the inversion (Turner et al. 2015; Wang et al. 2019; Janardanan et al. 2020). We use monthly emissions from EDGAR v6 inventory as the prior emissions (and true emissions for our OSSEs, see Section 7.4), which consist of 23 subsectors (see Table 7.1) (Crippa et al. 2021). Wetlands are the primary natural source consisting of more than 85% of the total natural emissions globally. Monthly wetlands emissions data from WetCHARTs v3.0 with the full ensemble mean (Bloom et al. 2017) is used and mapped into the domain using a uniform temporal profile.

**Table 7.1. Daily methane emissions in four main sectors and their subsets. Anthropogenic emissions are based on EDGAR v6, and natural wetland emissions are derived from WetCHARTs v3.0 with the full ensemble mean.**

Anthropogenic			Natural
Agriculture [ $mg_{CH_4} d^{-1} m^{-2}$ ]	Energy [ $mg_{CH_4} d^{-1} m^{-2}$ ]	Waste [ $mg_{CH_4} d^{-1} m^{-2}$ ]	Wetland [ $mg_{CH_4} d^{-1} m^{-2}$ ]
Agriculture Soil [1.913] Agriculture waste burning [0.684] Enteric fermentation <sup>a</sup> [51.195] Manure management [14.947]	Aviation <sup>b</sup> (all types) [0.042]		Wetland [87.823]
	Chemical process [0.112]		
	Combustion manufacturing [0.417]		
	Energy for building [1.790]		
	Fossil fuel fire [0.086]	Solid waste incineration [0.919]	
	Coal [15.954]	Solid waste <sup>c</sup> [34.034]	
	Gas [23.666]	Water waste handling [8.653]	
	Oil [3.799]		
	Iron-steel production [0.021]		
	Oil refineries [0.546]		
	Power industry [0.352]		
	Off-road [0.017]		
	Road transportation [0.717]		
Shipping [0.030]			

<sup>a</sup> Enteric fermentation and manure management represent the emissions of "livestock" used in other similar inversion studies. <sup>b</sup> all types of aviation refer to three subsets in EDGAR v6, including aviation climb descent, aviation cruise, and aviation landing takeoff. <sup>c</sup> Solid waste is equivalent to landfills in similar studies.

We process methane emissions from anthropogenic and natural sources using Sparse Matrix Operator Kernel Emissions (SMOKE v3.6) (UNC 2017) to provide hourly

gridded methane emissions into the hemispheric CMAQ model. Note that, for the proof of concept of the new inversion formulation proposed in this study (Section 7.3.3), in our OSSE experiments, we only consider four main sectors of emissions to be optimized within the inversion framework (Table 7.1). To fulfil this, we merge the 23 anthropogenic subsectors into three main categories (i.e., source sectors), namely agriculture, energy, and waste, based on methodological guidelines from IPCC (IPCC2013). Wetland is the fourth sector considered in our inversion.

### **7.3 Methodology**

Our approach is composed of two main parts. First, we perform an assimilation of CH<sub>4</sub> observations in order to provide a first guess concentration field along with the error variance of the concentrations. Then, the second part executes a source inversion using a 4D-Var cost function with a modified error covariance weight that accounts for both observation errors and model-propagated errors. An illustration of how those two parts perform is given in Figure 7.1, and further details will be provided later in this section.

#### **7.3.1 PvKF Assimilation**

The assimilation scheme performed here is a simplified form of the Kalman filter (not of an ensemble Kalman Filter). This algorithm is based on a parametric variance Kalman filter (PvKF), where the correlations are assumed to be homogeneous and isotropic, and the dynamics of the error variance are approximated by using only advection. The idea of evolving only error variance with an advection scheme emerged in Kalman filtering by Cohn (1993), and the first practical atmospheric implementation was demonstrated by Menard et al. (2000) and Menard and Changs (2000). The design and implementation of the PvKF assimilation with the CMAQ model using GOSAT methane

observations have been detailed in Voshtani et al. (2022a; 2022b). One main objective of developing the scheme was to reduce the assimilation cost while avoiding the challenges in the previous assimilation approaches, such as EnKF and 4D-Var. In fact, the assimilation requires only two model integrations, one for the state and the other for the error variance. It was shown that the method, despite its simplicity, is well-adapted for the assimilation of long-lived species (e.g., methane) without loss of variance. In the following, we briefly review the PvKF assimilation and its underlying assumptions.

The error covariance matrix,  $\mathbf{P}$ , is derived from an error covariance function of the following form

$$P(\mathbf{r}, \mathbf{r}', t) = \sigma(\mathbf{r}, t) C(\mathbf{r}, \mathbf{r}', t) \sigma(\mathbf{r}', t) \quad (7.1)$$

between a pair of model gridpoints  $(\mathbf{r}, \mathbf{r}')$ .  $\sigma$  is the standard deviation of model forecast error at a grid point  $\mathbf{r}$  at time  $t$ , and the model forecast error variance is propagated according to the advection equation.  $C(\mathbf{r}, \mathbf{r}', t)$  denotes the correlation function based on the second-order autoregressive (SOAR) correlation model, according to Voshtani et al. (2022a). Note that the error correlation function is never stored as a matrix but is computed during the assimilation as we need the correlation between a pair of points, and thus there is no requirement to store a large matrix in the model state-space. The algorithm consists of two steps:

- Analysis step

$$c_t^a = c_t^f + \mathbf{K}_t (y_t^o - \mathbf{H}_t^o c_t^f), \quad (7.2)$$

$$\mathbf{P}_t^a = \mathbf{P}_t^f - \mathbf{K}_t \mathbf{H}_t^o \mathbf{P}_t^f = \mathbf{P}_t^f - (\mathbf{H}_t^o \mathbf{P}_t^f)^T (\mathbf{H}_t^o \mathbf{P}_t^f \mathbf{H}_t^{oT} + \mathbf{R})^{-1} (\mathbf{H}_t^o \mathbf{P}_t^f), \quad (7.3)$$

$$\left. \begin{aligned} H_{\mathbf{r}}[P_t^f(\mathbf{r}, \mathbf{r}', t)] &\Leftrightarrow \mathbf{H}_t^o \mathbf{P}_t^f \\ H_{\mathbf{r}'}[P_t^f(\mathbf{r}, \mathbf{r}', t)] &\Leftrightarrow \mathbf{P}_t^f \mathbf{H}_t^{oT} \\ H_{\mathbf{r}}[H_{\mathbf{r}'}(P_t^f(\mathbf{r}, \mathbf{r}', t))] &\Leftrightarrow \mathbf{H}_t^o \mathbf{P}_t^f \mathbf{H}_t^{oT} \end{aligned} \right\} \Rightarrow \mathbf{K}_t = (\mathbf{H}_t^o \mathbf{P}_t^f)^T (\mathbf{H}_t^o \mathbf{P}_t^f \mathbf{H}_t^{oT} + \mathbf{R})^{-1}, \quad (7.4)$$

- Forecast step

$$c_{t+1}^f = M_t c_t^a, \quad (7.5)$$

$$(\sigma_{t+1}^f)^2 = M_t^* (\sigma_t^a)^2 + q. \quad (7.6)$$

$c_t^f$  and  $c_t^a$  are the forecast and analysis concentrations, respectively, and  $\mathbf{P}_t^f$  and  $\mathbf{P}_t^a$  are the corresponding error covariances.  $H_{\mathbf{r}}$  and  $H_{\mathbf{r}'}$  represent observation operator, acting on gridpoint  $\mathbf{r}$  and  $\mathbf{r}'$ , and  $\mathbf{K}$  is the Kalman gain matrix.  $M$  denotes the CMAQ forecast of the state,  $M^*$  represents the forecast based on the CMAQ advection scheme, and  $q$  is the model error variance. Those forecast error parameters are then used in Equation (7.1) to update the analysis step.

The initial and model error covariances are assumed to be diagonal and have the form of  $\mathbf{P}_0 = (\varepsilon^i)^2 \mathbf{I}$  and  $\mathbf{Q} = q \mathbf{I}$  respectively. According to the method of innovation variance consistency proposed in Voshtani et al. (2022b),  $\varepsilon^i$  is obtained as 2.25% of the initial concentrations field, and the accumulative model error variance,  $q$ , is 0.66% relative to the hourly analysis field. The observation errors are proportional to the measurement errors ( $\varepsilon^o = f^o \varepsilon^m$ ), and the estimated factor is  $f^o = 0.5$  based on the cross-validation technique demonstrated in Voshtani et al. (2022b). Accordingly, the observation error covariance is considered diagonal and has the form of  $\mathbf{R} = (\varepsilon^o)^2 \mathbf{I}$ . Furthermore, background error correlation, based on the SOAR correlation function, has the form

$$C = C_{SOAR} = \left(1 + \frac{D}{L}\right) \exp\left(-\frac{D}{L}\right), \quad (7.7)$$

where  $D$  is a chordal distance between the position vector of a pair of grid points on the surface of the sphere, and  $L$  denotes the correlation length scales. Based on the cross-validation method, the estimated horizontal and vertical length scales, as used in this study, are  $L_h = 350$  km and  $L_v = 7\sigma_l$ . Note that  $\sigma_l$  denotes the model vertical layer, starting from the surface, based on the sigma-pressure coordinate.

### 7.3.2 Source Inversion Procedure (4D-Var)

The 4D-Var inversion performed here integrates the same type of observations as used in PvKF assimilation (GOSAT column methane) but for a different period of time to avoid double-counting. The inversion relies on a CTM and its adjoint model to optimize methane emissions in the Northern Hemisphere. To optimize surface methane emissions in a formal 4D-Var inversion, we seek to minimize the cost function in the form of

$$J(x) = \frac{1}{2} \gamma (x - x_b)^T \mathbf{B}^{-1} (x - x_b) + \sum_{t=0}^n \frac{1}{2} (y_t^o - H_t^o c_t)^T \mathbf{R}_t^{-1} (y_t^o - H_t^o c_t) \quad (7.8)$$

where  $x = \log(e/e_b)$  denote the emissions scaling factor with a logarithmic form,  $x_b$  represents the corresponding scaling factor of baseline emissions ( $e_b$ ),  $n$  is the number of the hourly timesteps,  $c_t$  denotes the model state (i.e., a 3D field of methane concentrations) at the time  $t$ ,  $y_t^o$  is the methane observations for the timestep  $t$ , and  $H_t^o$  represents the observation operator that maps the model state into the observation space.  $\mathbf{B}$  and  $\mathbf{R}$  denote matrices of the prior emissions and observations, respectively, and  $\gamma$  is a regularization parameter.

Based on Equation (7.8), we correct for monthly mean methane emission ( $e$ ) at  $108 \times 108$  km resolution using variational optimization, which is an iterative procedure

(Hakami et al. 2005; Sandu et al. 2005; Elbern et al. 2007; Bousserez et al. 2016; Stanevich et al. 2021). In fact, the gradients of the cost function with respect to the methane emissions are computed at each hour using the CMAQ adjoint model. Those gradients are used to obtain emission-weighted monthly mean sensitivities of emissions scaling factors in each grid cell. At the end of the iteration, the sensitivities are supplied to a quasi-Newton limited memory optimization routine, L-BFGS (Byrd et al. 1995), for which we use the "optimr" package in R. Using the new scaling factors provided by L-BFGS, updated emissions are obtained and used in the next iteration. We assume that the convergence of this iterative procedure occurs once the reduction of the cost function in consecutive iterations is less than 1%.

A conventional method in inverse modelling to balance the weight of the prior constraint in the cost function is to use a global regularization parameter  $\gamma$  (Hansen 1999; Hakami et al. 2005; Chen et al. 2021). It is also interpreted as a rough compensation for the missing objective information in quantifying error correlations in  $\mathbf{R}$  (Lu et al. 2022). The estimation of the regularization parameter is performed by inspection and by conducting a series of 4D-Var inversions while choosing the parameter  $\gamma$  that minimizes the cost function  $J$  (Equation (7.8)) among all optimal experimented estimations. Accordingly, we found the value of  $\gamma = 900$  as our best estimate (see Figure E2 in Appendix E2 for the details). For matrix  $\mathbf{B}$ , we have adopted a simple approach that the emissions errors are assumed uncorrelated in space and between sectors, making  $\mathbf{B}$  a diagonal matrix. In fact, we give the same error weight geographically and for all sectors, resulting in an emissions-weighted matrix with 100% uncertainty in each grid cell.

The evolution of the model state using hemispheric CMAQ can be shown by the operator  $M$  that computes the state at a future time, given the initial state  $c_i$  and emissions scaling factor  $x$ .

$$c_t = M(c_i, x), \quad (7.9)$$

The model forecast equivalent value to the observation ( $y_t^f$ ) thus, can be denoted as

$$y_t^f = H_t^o c_t = H(c_i, x) \quad (7.10)$$

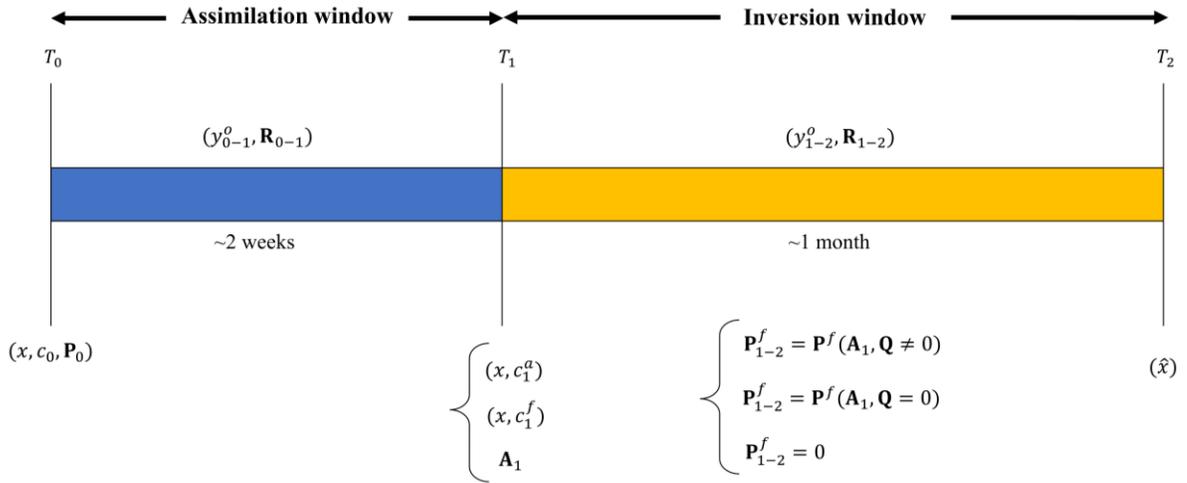
where  $H$  represents a sensitivity operator (in the context of inverse modelling, it is also called an observation operator of inversion) that acts on the initial concentrations and emissions scaling factor to provide the model equivalent value at the location of the observations. Following Equation (7.10), it is typically assumed in a 4D-Var inversion that the state of the system is perfect, meaning that there is no error involved in  $H$  (neither in the model,  $M$ , nor the initial state,  $c_i$ ). Equation (7.8), as used in this study, is re-written in the form

$$J(x) = \frac{1}{2} \gamma (x - x_b)^T \mathbf{B}^{-1} (x - x_b) + \sum_{t=0}^n \frac{1}{2} (y_t^o - H_t(c_i, x))^T \mathbf{R}_t^{-1} (y_t^o - H_t(c_i, x)) \quad (7.11)$$

### 7.3.3 Using PvKF Assimilation Analysis in 4D-Var Inversion: the Formulation

A common assumption made in 4D-Var source inversion is that the uncertainty in the initial state is negligible compared to the accumulated effect of emissions uncertainty over a long time period, which is also typically used for methane inversion. With such an assumption, there is no dependency on the initial state uncertainty in the 4D-Var cost function (Equation (7.11)). This is equivalent to considering the initial concentrations close to the true estimate (Basu et al. 2013; Cressot et al. 2014; Deng et al. 2014; Alexe et al. 2015; Bergamaschi et al. 2015; Bergamaschi et al. 2018).

Here in this chapter, we employ a different approach by making the initial state uncertainty as small as possible by assimilation of observations prior to the 4D-Var inversion window. Furthermore, by conducting PvKF assimilation, we will know what the state assimilated uncertainty is at the beginning of the 4D-Var window. Figure 7.1 shows the assimilation window (the blue bar or  $T_{0-1}=(T_0, T_1)$ ) alongside the inversion window (the yellow bar or  $T_{1-2}=(T_1, T_2)$ ) and how they are connected to each other. The PvKF assimilation starts with the initial condition ( $c_0$ ), derived from the global modelling and measurements data (see Section 7.3.1) and computes the error variance at the end of the assimilation time window  $T_1$ . Note that as part of the assimilation, an estimation of the modelling error variance (growth rate), observation error variance, and background correlation length are obtained (see Section 7.3.1 for details). These estimated error statistics will be an informative source for the subsequent OSSEs. At the end of the PvKF



**Figure 7.1. Sketch of the assimilation window followed by an inversion window, used for estimation of methane emissions scaling factor,  $x$ .**

assimilation window ( $T_1$ ) the analysis field ( $c_1^a$ ) and the analysis error covariance ( $\mathbf{A}_1$ ) are obtained and thus will be used for the inversion (Figure 7.1). In the context of an

OSSE (Section 7.4), we need to generate observations from the true state. Since the PvKF optimal analysis can be considered as the closest estimate of the true state, we thus have

$$y_{1-2}^f = H(c_1^a, x^t) + \varepsilon_{1-2}^f \quad (7.12)$$

$$y_{1-2}^o = y_{1-2}^f + \varepsilon^o \quad (7.13)$$

where  $y_{1-2}^f$  is the model forecast mapped on observations time and locations using the observation operator  $H^o$ ;  $H$  includes the transport and the observation operator,  $H^o$ , (see Section 7.3.2 for details).  $\varepsilon_{1-2}^f$  represents the model transported analysis error up to the current forecast time, and  $\varepsilon^o$  is the observation error used to construct  $\mathbf{R}$ . The associated forecast error covariance for the inversion window is denoted as  $\mathbf{P}_{1-2}^f = \mathbf{P}^f(\mathbf{A}_1, \mathbf{Q})$ . In order to obtain the forecast error covariance, we use the predicted forecast error variance (Equation (7.6)), and we have

$$\mathbf{P}_i^f = \Sigma_i^f \mathbf{C} \Sigma_i^f \quad (7.14)$$

where  $\Sigma_i^f$  is the diagonal matrix of forecast error standard deviation ( $\sigma_i^f$ ), and  $\mathbf{C}$  is the matrix of error correlations. The forecast error variance is computed according to Equation (7.6), and correlation is derived from Equation (7.7).

Now, let us define a new form of the cost function for performing our inversion. In this case, other than the observation errors that affect the innovation (i.e., Observations – Model; see Equation (7.8)), we have to account for the propagated analysis error in the inversion window. The analysis error at the time  $T_1$  depends on observations in the window  $T_{0-1}$ . Since the observation errors are assumed to be temporally uncorrelated, the analysis error  $\varepsilon_1^a$ , is uncorrelated with the future observations used in the inversion window ( $T_{1-2}$

). Hence, the innovation with respect to the analysis ( $y_i^o - H_i(c_1^a, x)$ ) should have a weight of the form  $H^o \mathbf{P}_i^f (\mathbf{A}_1, \mathbf{Q}) H^{oT} + \mathbf{R}$ , which is a sum of two terms (indicating that their contributions are uncorrelated). Therefore, the modified form of the cost function that is appropriate for our two-part scheme has the form

$$J(x) = \frac{1}{2} \gamma (x - x_b)^T \mathbf{B}^{-1} (x - x_b) + \sum_{i=0}^n \frac{1}{2} (y_i^o - H_i(c_1^a, x))^T (H^o \mathbf{P}_i^f (\mathbf{A}_1, \mathbf{Q}) H^{oT} + \mathbf{R}_i)^{-1} (y_i^o - H_i(c_1^a, x)) \quad (7.15)$$

### 7.3.4 Numerical Aspects of Matrix Inversion

We use GOSAT methane observations for a period of a month, which contains 11,489 observations after all quality control and bias removal (Voshtani et al. 2022a). It is typically assumed in methane inversion studies that the observation errors of this type are uncorrelated (or with insignificant correlations) both in space and time, resulting in a diagonal observation error covariance ( $\mathbf{R}$ ) (Turner et al. 2015; Bousserez et al. 2016; Turner et al. 2018; Stanevich et al. 2021; Zhang et al. 2021). Contrary to the observation error covariance, the forecast error covariance in observation space ( $H^o \mathbf{P}^f H^{oT}$ ) is not diagonal (Equation (7.15)). From a physical point of view, since the forecast error is propagated in time and space, its error covariance mapped into a one-month-long observation space is no longer a diagonal matrix (in observation space). To make this more clear, the elements of the  $11,489 \times 11,489$  covariance matrix (i.e.,  $H^o \mathbf{P}^f H^{oT} + \mathbf{R}$ ) are, in fact, ordered by observation ID numbers (not by time and not with space). Therefore, the forecast error correlates with different ID numbers (or observation space), resulting in a covariance matrix,  $H^o \mathbf{P}^f H^{oT} + \mathbf{R}$ , that is not a diagonal matrix for a month-long data

window. In contrast, if we would only account for observation errors that are spatially and temporally uncorrelated, it would lead to a diagonal matrix in the observation space.

Since the number of observations is significant, as is the case here, inverting the covariance matrix,  $H^o \mathbf{P}^f H^{oT} + \mathbf{R}$ , becomes problematic. For regular matrices of that size (i.e., nondiagonal, non-sparse), the matrix inversion is not only computationally expensive but could lead to numerical issues (Strang and Borre 1997).

In data assimilation, a traditional approach, known as the data selection procedure (Cohn et al. 1998; Houtekamer and Mitchell 2001; Lahoz and Schneider 2014) is used to avoid inverting large covariance matrices in observation space,  $H^o \mathbf{P}^f H^{oT} + \mathbf{R}$ . The data selection procedure involves partitioning the entire set of observations into smaller sets, known as batches of observations, that are mutually uncorrelated (also referred to as observation packets by Rodgers (2000)). In this case, the sizeable nondiagonal observation error covariance matrix transforms into a block diagonal matrix (with a reasonable block dimension), where each block represents the corresponding batch of observations (Figure 7.2).

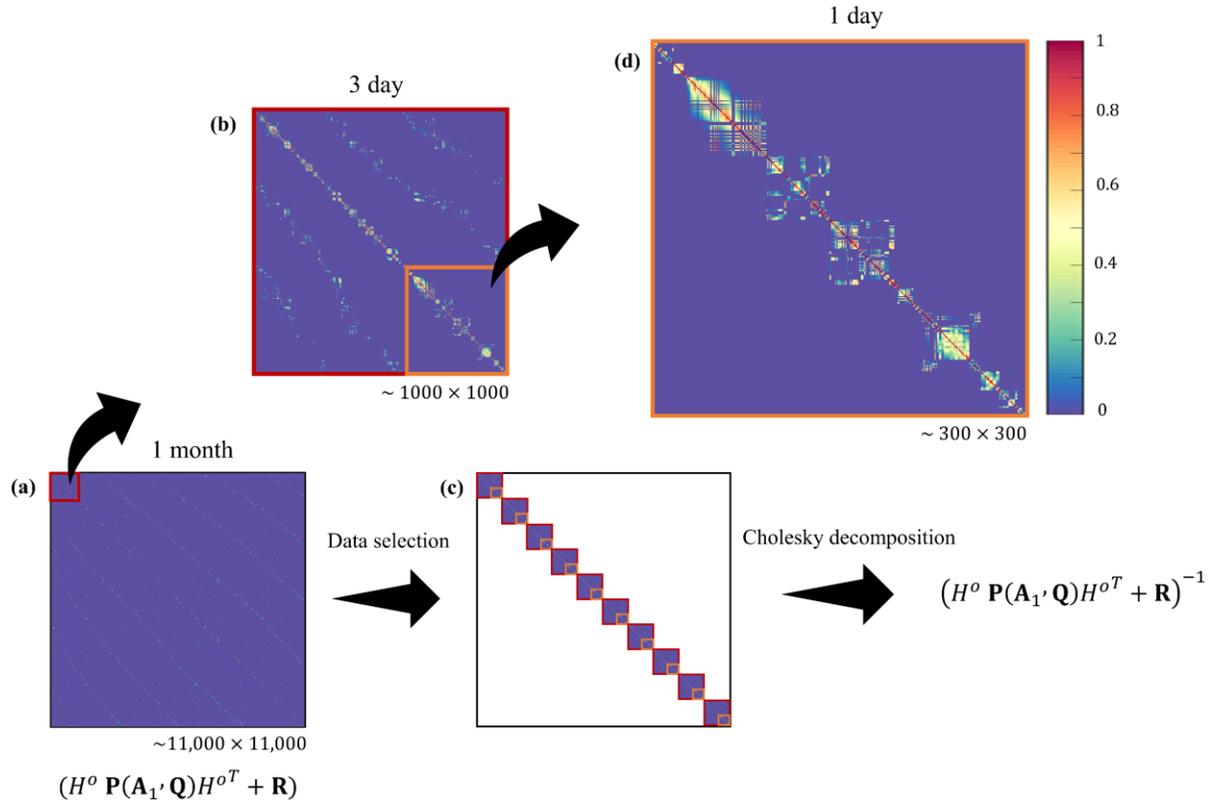
In data selection, in general, we divide the model domain into  $N$  regions and perform the analysis for each region. By limiting the number of observations ( $p$ ) that influence analysis in a given region (e.g.,  $p < 1000$ ), the size of the error covariance matrix to invert will be reduced (i.e.,  $p \times p$ ) and thus be manageable. The influence of observations on a region is determined by the correlation length scale. For example, any observation that is farther than five times the correlation length will not contribute to the analysis equation of the corresponding region. Thus, the number of regions to consider depends on the correlation length scale and the maximum number of observations that we can process in

matrix inversion. This procedure is typical for optimal interpolation (OI) (Lahoz and Schneider 2014). Note that some variants of this scheme (not discussed here) are also designed for the ensemble Kalman filter (Houtekamer and Mitchell 2001; Migliorini 2013).

In the case of GOSAT satellite observations, each observation in space is considered with its own time (i.e., satellite retrieval time). Thus, it is more appropriate to partition the GOSAT observations according to their retrieval times, meaning that the error correlations between two observations not only depend on their geographical distance but also on their time difference. In this study, we simply assume that there are no correlations after three days between two locations of observations. As shown in Figure 7.2, we conduct a data selection procedure by considering a 3-day batch of GOSAT observations equivalent to the satellite revisit time. In this case, the number of observations within a batch remains as low as about 1000. Note that for a larger number of observations ( $p \gtrsim 1500$  herein), the condition number of the covariance matrix increases rapidly, resulting in a non-full rank matrix.

Figure 7.2a displays the full error covariance matrix for a period of a month (and of a size  $\sim 11,000 \times \sim 11,000$ ). Ignoring the small covariances between the 3-day batches results in a block diagonal matrix (Figure 7.2c). Each block is represented in Figure 7.2b, where we perform matrix inversion using the Cholesky decomposition (Krishnamoorthy and Menon 2013) for solving a system of linear equations of the same size as in that block. In Figure 7.2d, the covariance elements are shown within a day and are normalized relative to the values in the main diagonal. The correlation structures (off-diagonal elements) in this figure retain values that are as substantial as the main diagonal. Figure 7.2b indicates that the sub-diagonal elements (i.e., non-zero values that run parallel to the main diagonal)

are from observations that occur one day (for the closest sub-diagonal line) or two days (for the second sub-diagonal line) later, but at slightly different locations. It implies that observations in a subsequent day are being translated into space according to the 3-day revisit cycle of GOSAT, but because of accounting for correlations using  $H^o \mathbf{P}^f H^{oT}$ , they appear as non-zero correlations, although  $\mathbf{R}$  remains diagonal.



**Figure 7.2.** (a) one-month non-diagonal error covariance,  $H^o \mathbf{P}^f H^{oT} + \mathbf{R}$ , is applied to the data selection procedure with (b) 3 days batch of observations to form (c) a block diagonal matrix of the same size. (d) shows the non-diagonal covariance matrix in one day.

#### 7.4 Description of the OSSE Experiments

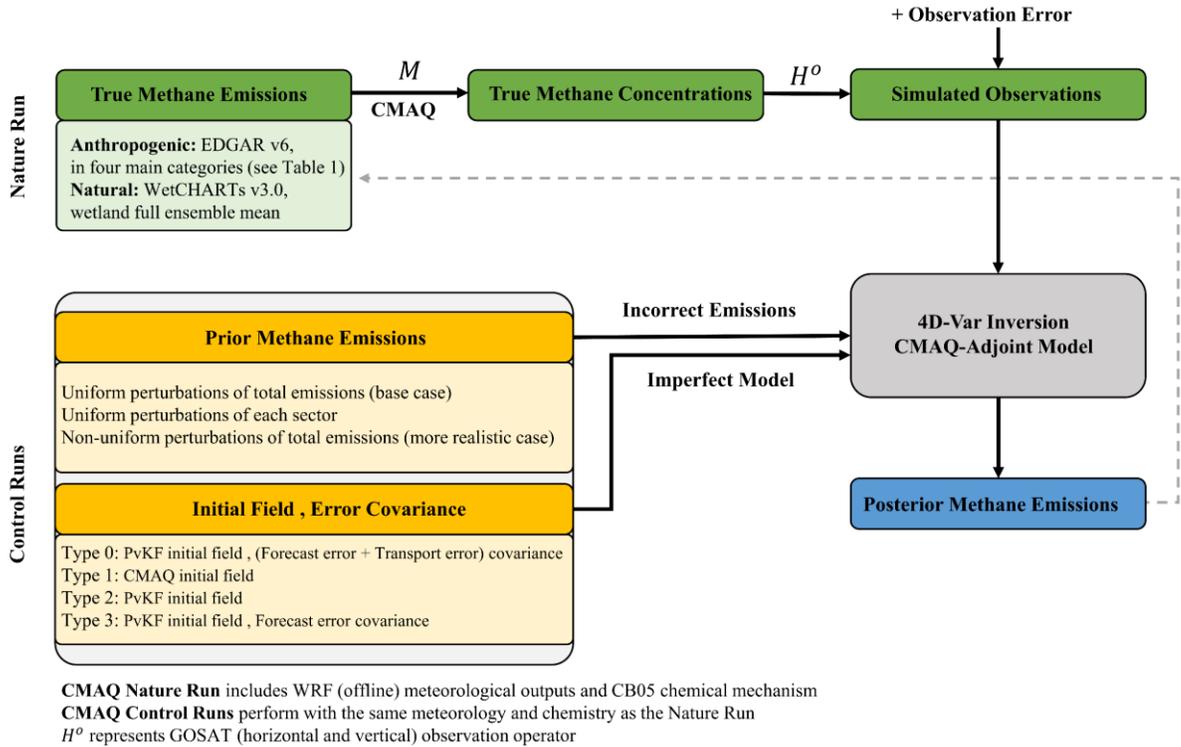
OSSE is a standard method to evaluate the ability of atmospheric inversion or assimilation systems without using the actual observation data (Lahoz and Schneider 2014). Our OSSE experiments are designed with synthetic GOSAT observations, aiming

to verify how optimal state analysis and its model error propagation may improve a 4D-Var inversion for constraining methane emissions. Accordingly, we test different effects from the state estimation on the source inversion (see Section 7.3.3, Figure 7.1). To distinguish those effects, we evaluate the OSSE using different 4D-Var cost functions, across several forms of emissions perturbation, and we discuss various aspects of those effects. The basic description of the OSSE setup and emissions perturbations are provided in Section 7.4.1, followed by explanations of the cost function variations in Section 7.4.2.

#### **7.4.1 Perturbation Tests**

Figure 7.3 shows the generic design of an OSSE framework. The structure of every OSSE system consists of two main parts: A nature run and control runs (Brasseur and Jacob 2017). Through nature run, the CMAQ model forecasts produce the synthetic (true) concentrations field to be sampled by the observation operator, providing simulated observations. The nature run is derived by the initial analysis field of concentrations at time  $T_1$ , meteorological fields from WRF output, and anthropogenic and natural methane emissions that are taken to be the same as provided by corresponding inventories (see Section 7.2.2). It is assumed that both inputs and the model CMAQ in the nature run are deterministic, comparable to a mean estimate of a stochastic process. On the other hand, the simulated observations are not assumed to be perfect but include GOSAT observation errors. Hence, the simulated observations are imperfect only due to observation errors. The synthetic (i.e., simulated) observations generated through a nature run are used under a controlled environment to perform different inversion runs. In fact, those are conducted within control runs (also known as perturbed runs), for which different forms of initialization, error statistics, and emissions perturbations are configured for the inversion

window (Figure 7.3). Therefore, control runs only cover the inversion window but include the effects of the assimilation window by providing the initial field and/or initial error covariances at time  $T_1$ .



**Figure 7.3. Flowchart of the OSSE framework for optimizing methane emissions.**

Besides four types of cost functions that will be described in Section 7.4.2, we consider three forms of perturbations to produce different prior methane emissions in the control runs. Those perturbations reflect uniform and variable biases in the prior methane emissions and cover both total and sectoral emissions aspects. Note that in all our control runs, we assume that the initial state and the CTM are imperfect, but we use the same meteorological fields, chemical reactions, and boundary conditions as the nature run. Hence, those processes are considered perfect, and their effects (potential errors) are not investigated for the objective of our OSSE experiments in this study.

The main goal of our OSSE experiments is to test the ability of our proposed inversion cost function (Equation 7.15) to reproduce true methane emissions (comparison between optimized and true emissions in Figure 7.3). In addition, by exploring three other variations in the cost function (see Section 7.4.2), we aim to address the limitation of a typical 4D-Var inversion that relies on perfect state assumptions due to the initial field and the CTM (meteorology and chemistry are excluded). The approximation of diagonal observation error covariance is also evaluated in our OSSE experiments.

#### 7.4.2 Experimenting with Different Cost Functions

Table 7.2 summarizes four variations in the cost function used to determine the different impacts of the state assimilation when it is linked to the inversion system. We recall that for this study, the inversions neither rely on a perfectly known initial state nor a perfect forward model. We start with Type 0 as the basis for cost function variations proposed in this study (see Section 7.3.3, Equation (7.15)). In this cost function, we account for the entire information provided by PvKF assimilation, including the initial analysis field ( $c_1^a$ ) and its analysis error covariance ( $\mathbf{A}_1$ ). In addition, according to the PvKF formulation for propagating errors using the advection scheme (see Section 7.3.1), the forecast of the analysis error covariance and the estimated model transport error ( $\mathbf{P}_t^f(\mathbf{A}_1, \mathbf{Q})$ ) are integrated into the second term of the cost function (Equation (7.16)). In Types 1-3 (Table 7.2), we consider other forms of cost functions where portions of the connection between the assimilation and inversion are removed. For Type 1, we neither consider the analysis field nor the propagation of the error covariances (i.e.,  $\mathbf{P}_t^f(\mathbf{A}_1, \mathbf{Q}) = 0$ ) during the inversion. In fact, the inversion is independent of the state assimilation and is initialized by the model forecast, which is assumed to be perfectly known. In this case, the

inversion also only relies on the observation error covariance ( $\mathbf{R}_t$ ) in the cost function (Equation (7.17)). Note that the Type 1 cost function is frequently used in different inversion studies (Cressot et al. 2014; Turner et al. 2015; Wang et al. 2019). Type 2 inversion is similar to Type 1, except that it begins with the initial analysis field rather than the forecast initial concentrations (Equation (7.18)). Type 2 cost function is also commonly used in the literature (Basu et al. 2013; Deng et al. 2014; Basu et al. 2022). Type 3 inversion not only accounts for the initial analysis field but also considers the propagation of its error covariance during the inversion, yet without the effect of model transport error (i.e.,  $\mathbf{P}_t^f(\mathbf{A}_1)$ ) (Equation (7.19)). Note that we keep  $\gamma$ ,  $x_b$ , and  $\mathbf{B}$  the same between all these cost functions for consistency of the evaluations.

**Table 7.2. Cost functions for different formulations of 4D-Var inversion.  $\mathbf{P}_t$  is the model propagated (forecast) of error covariance in the model space,  $\mathbf{A}_1$  is the analysis error covariance at the initial time of inversion, and  $\mathbf{Q}$  is the model transport error covariance that is estimated independently (see Section 7.3.1).  $c_1^f$  and  $c_1^a$  represent the initial field of concentrations produced from the CMAQ model and Pvkf assimilation, respectively.**

Type #	Cost function
Type 0:	$J_0(x) = \frac{1}{2}\gamma(x - x_b)^T \mathbf{B}^{-1}(x - x_b) + \sum_{t=0}^n \frac{1}{2}(y_t^o - H_t(c_1^a, x))^T (H^o \mathbf{P}_t(\mathbf{A}_1, \mathbf{Q})H^{oT} + \mathbf{R}_t)^{-1}(y_t^o - H_t(c_1^a, x)) \quad (7.16)$
Type 1:	$J_1(x) = \frac{1}{2}\gamma(x - x_b)^T \mathbf{B}^{-1}(x - x_b) + \sum_{t=0}^n \frac{1}{2}(y_t^o - H_t(c_1^f, x))^T (\mathbf{R}_t)^{-1}(y_t^o - H_t(c_1^f, x)) \quad (7.17)$
Type 2:	$J_2(x) = \frac{1}{2}\gamma(x - x_b)^T \mathbf{B}^{-1}(x - x_b) + \sum_{t=0}^n \frac{1}{2}(y_t^o - H_t(c_1^a, x))^T (\mathbf{R}_t)^{-1}(y_t^o - H_t(c_1^a, x)) \quad (7.18)$
Type 3:	$J_3(x) = \frac{1}{2}\gamma(x - x_b)^T \mathbf{B}^{-1}(x - x_b) + \sum_{t=0}^n \frac{1}{2}(y_t^o - H_t(c_1^a, x))^T (H^o \mathbf{P}_t(\mathbf{A}_1)H^{oT} + \mathbf{R}_t)^{-1}(y_t^o - H_t(c_1^a, x)) \quad (7.19)$

By comparing all the cost functions provided in Table 7.2, we can distinguish the different effects of linking assimilation with inversion on the inversion results. Accordingly, a comparison between Type 1 and Type 2 cost functions shows the influence of only the initial field on the inversion result. By comparing Type 2 and Type 3 cost functions, we can isolate the effect of considering the uncertainties in the model concentrations that originates from the initial state (i.e., model-propagated initial error covariance). Finally, if we compare Type 3 with Type 0 cost function, the influence of model transport error  $\mathbf{Q}$  (already estimated independently) on the inversion can be extracted. In fact, the forecast of model error covariances for these two types have the form:

$$\text{Type 0: } (H^o \mathbf{P}^f(\mathbf{A}_1, \mathbf{Q}) H^{oT})_{t+1} = (H^o M \mathbf{P}^f(\mathbf{A}_1) M^T H^{oT})_t + H^o \mathbf{Q} H^{oT}, \quad (7.20)$$

$$\text{Type 3: } (H^o \mathbf{P}^f(\mathbf{A}_1) H^{oT})_{t+1} = (H^o M \mathbf{P}^f(\mathbf{A}_1) M^T H^{oT})_t + \underline{H^o \mathbf{Q} H^{oT}}. \quad (7.21)$$

## 7.5 Results and Discussions

The results of the OSSEs contain a month of modified 4D-Var inversion in April 2010, preceded by two weeks of PvKF assimilation (or model forecast when the assimilation is turned off). The posterior (i.e., optimized) emissions estimate involves monthly mean methane emissions (or, in particular, emissions scaling factors). The following subsections (Sections 7.5.1-7.5.3) are each dedicated to a particular form of emissions perturbations. We recall that all the input and configurations, except those described in Figure 7.3, which impose the discrepancy between the OSSE cases, are the same for all experiments (e.g., meteorological field, regularization parameter  $\gamma$ ).

### 7.5.1 Base case Uniform Perturbation

The first type of perturbation within our OSSE control runs (Figure 7.3) provides prior emissions that include a uniform bias in all sectors. In this case, the total true

emissions are uniformly scaled up by 50%, which is a common perturbation method also used in previous methane inversion studies with OSSEs (Bousserez et al. 2016; Sheng et al. 2018; Turner et al. 2018; Yu et al. 2021). For this type, in fact, it is assumed that the prior emissions are strongly in line with the spatial distributions of the true emissions, yet with different levels of magnitude. Although this perturbation method implicitly considers a low level of uncertainty for the spatial allocation of emissions, it can be taken as the base case, particularly to evaluate the ability of the inversion method and underlying assumptions for reproducing true emissions (Zhang et al. 2018; Yu et al. 2021; Wu et al. 2022).

Figure 7.4 compares the posterior emissions from the four inversions with different cost function variations, given a prior estimate with +50% uniform perturbation of the true emissions. The spatial distribution of the differences between prior and true emissions ( $\Delta e^{prior}$ ) and between all four posteriors and true emissions ( $\Delta e^{posterior}$ ) are shown in Figure 7.4a and Figure 7.4b-e, respectively. Type 0 inversion (Figure 4b) corresponding to what is proposed in this study accounts for both optimal PvKF analysis ( $c_1^a$ ) and model propagated (forecast) of analysis error covariance during inversion ( $\mathbf{P}_t^f(\mathbf{A}_1, \mathbf{Q})$ ). Posterior emissions show reasonable overall consistency with the true emissions, particularly for the larger and more local (or point) sources. Now, we remove all the dependency on the assimilation such that a perfect model forecast ( $c_1^f, \mathbf{P} = 0$ ) is linked to the inversion (Type 1). According to our OSSE, in this case, the provided initial condition is far from the truth. Figure 7.4c, corresponding to its posterior emissions, indicates a large deviation from the true emissions. In fact, a significant (downward) over-correction in many regions, especially for the large sources, is obtained along with insufficient correction for the small

sources. This behaviour first implies that relying on a model forecast initial concentrations with a perfect assumption exerts a substantial impact on the state of the system, mainly due to accumulating incorrect emissions before inversion (i.e., prior emissions are incorrect). Hence, this eventually degrades the emissions estimation through inversion. Furthermore, the prior error covariance (**B**) is commonly assumed (uniformly) to be proportional to the emissions (as it is in this study). This itself also limits the ability of the inversion to recover fairly large and small (scattered) sources, as also suggested by Yu et al. (2021). In addition, the inversion performance can be worsened in the presence of an incorrect state and missing modelling error correlations.

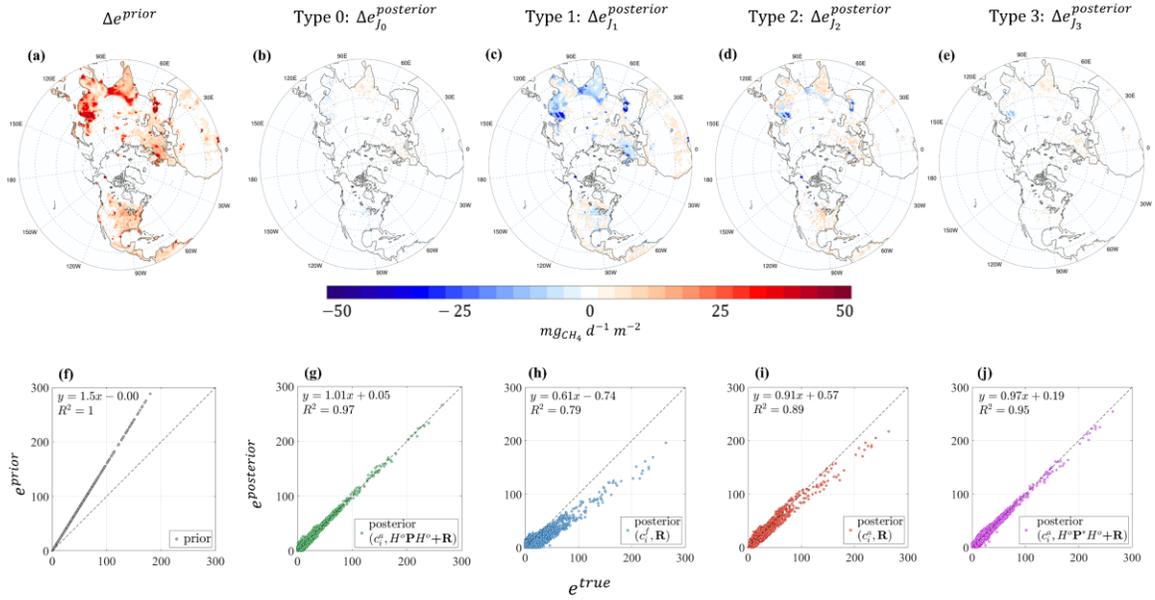
In Type 2 (Figure 7.4d), the inversion is performed with the PvKF analysis ( $c_1^a$ ) that is used as the closest estimate to the true state, but analysis uncertainty is not taken into account to propagate errors during the inversion, meaning that the model state during inversion is perfectly known. Thus, we only rely on the observation error variance (diagonal matrix **R**) in the corresponding cost function (Equation (7.18)). For this case, although Figure 7.4d shows an overall improvement in the posterior emissions estimates compared to Figure 7.4c, it remains inaccurate, particularly in the regions with large amounts of methane emissions, such as East Asia and around the Persian Gulf, and for emission sectors with larger area coverage such as Agriculture in the Midwestern United States and India.

However, comparing the posterior emission of Type 2 and Type 3, we might infer that, besides the analysis field, accounting for the analysis error covariance and propagating it during the inversion can significantly improve the emissions estimates. This is mainly due to the structures of the correlations that exist in the model forecast (see Figure 7.2),

while it is usually neglected due to perfect model assumptions or for computational purposes (Lu et al. 2022). Errors in the forecast during the inversion can be produced by model transport due to various effects (Locatelli et al. 2015; Stanevich et al. 2020) or can be the result of an initial error that is propagated through the model, both of which are important to be considered besides the observation errors for a realistic estimation problem.

In Type 3 inversion, a similar condition as Type 0 is considered, except that the modelling error ( $\mathbf{Q}$ ) is turned to zero. This depicts a scenario where the model transport is assumed perfect, but in our experiments (and in reality), it may not necessarily maintain the optimal state (closest to the true). Drawing a comparison between Figure 7.4e (Type 3) and Figure 7.4b (Type 0) shows minor global differences, although, over some larger emissions regions, such as East Asia, that discrepancy still remains noticeable. This is likely due to the lower density of the GOSAT observations over those regions, for which adding model error ( $\mathbf{Q}$ ) makes those observations more impactful for the inversion. Hence, accounting model error  $\mathbf{Q}$  in 4D-Var inversion provide room for improvement, particularly when observations are insufficient. On the other hand, comparing either Type 0 or Type 3 against Type 2 inversion indicates that a considerable improvement occurs for both large and small sources. This emphasizes the key role of the model-propagated error correlations (initialized by the analysis error covariance,  $H^o \mathbf{P}_t^f(\mathbf{A}_1, \mathbf{Q}) H^{o T}$ , that are overlooked in an inversion with a perfectly known state and diagonal observation error covariance.

total emissions perturbation (uniform)



**Figure 7.4.** (a) prior – true emissions (+50% uniform perturbation); (b) posterior – true emissions in Type 0 inversion using analysis initial ( $c_1^a$ ) and both observation R and model propagated analysis error covariance  $H^o P_i^f(A_1, Q) H^{o T}$ ; (c) posterior – true emissions in Type 1 inversion using forecast initial ( $c_1^f$ ) and observation error covariance R, (d) posterior – true emissions in Type 2 inversion using analysis initial ( $c_1^a$ ) and observation error covariance R; (e) posterior – true emissions in Type 3 inversion using analysis initial ( $c_1^a$ ) and both observation and model propagated analysis error covariance  $H^o P_i^f(A_1) H^{o T}$ , but without model error. Statistical comparison of the (f) prior emissions and (g-j) posterior emissions of Type 0-3 inversion, respectively. x-axis and y-axis represent the true and prior/posterior emissions, respectively. In (f-j),  $P^f(A_1, Q)$  is shown as P, and  $P^f(A_1)$  is shown as  $P^*$ . Synthetic observations are generated using the nature run initialized by the analysis, and a 2-week spin-up is used for the initialization.

OSSE experiments aid in determining the statistics of the posterior emissions without a need to estimate those along with the inversion, which otherwise entails a high computational cost (Bousserez et al. 2016; Bousserez and Henze 2018). Accordingly,

besides the spatial maps, the prior and those four optimized emissions are demonstrated in scatter plots as in Figure 7.4f-j. Figure 7.4h indicates that Type 1 inversion, which integrates the biased model initial forecast (due to biased emissions before the inversion), can result in fairly biased posterior emissions along with significant variance. Accounting for the initial analysis field in Type 2 inversion improves on the bias of the optimized emissions while slightly decreasing the variance (Figure 7.4i). On the other hand, propagating the analysis error (comparison between Figure 7.4j and Figure 7.4h) can largely affect the variance of posterior emissions, although with a small bias improvement.

Overall, adding the model error,  $\mathbf{Q}$ , (comparison between Figure 7.4g and Figure 7.4j) can further improve the bias (and slightly variance) in posterior emissions, particularly for the larger sources. (as in Figure 7.4e and Figure 7.4b). Additionally, the posterior emissions in Figure 7.4g maintain an estimate that is also statistically more reliable than other inversion types, although it tends to be less reliable for estimating small sources. This deficiency is likely due to limited information in determining the prior error covariance ( $\mathbf{B}$ ) for properly weighting the prior emissions (Yu et al. 2021; Lu et al. 2022).

## **7.5.2 Perturbation of Different Sectors**

In the second form of perturbation defined in our OSSE experiments (see Figure 7.3), we examine the ability of our inversion to reproduce true emissions when only one particular emission source sector is perturbed. Following Table 7.1, each source sector is uniformly perturbed in the same way as described in Section 7.5.1 (i.e., scaled up by 50%) and then recovered in an individual OSSE inversion. We repeat a similar assessment as in Section 7.5.1, with different inversion cost functions (Type 0-3 inversion or Equations (7.16) to (7.19)). Figure 7.5 presents the spatial distributions of the differences between

total prior/posterior methane emissions and true emissions, in which every map corresponds to a specific sectoral perturbation that is performed with a particular inversion type.

The overall spatial pattern of the posterior emissions in each emissions sector shows the same order of improvement as obtained for the total emissions in Section 7.5.1. In fact, the contribution of the PvKF analysis field, used as the initial condition, and its error covariance to the inversion leads to a better constraint for every individual sector. However, the responses of each source sector to different types of inversions are not the same. Before discussing the details of the emissions correction in each sector, we need to reacquaint ourselves with some characteristics of each source sector, which can be addressed from the figures with prior perturbations (Figure 7.5a,f,k,p). Besides the geographical locations of the prior emissions in each sector, it is important to identify their spatial distribution along with the level of magnitude of each sector's prior emissions. According to those criteria, we consider agriculture and wetland emissions as area sources (cover large areas with an almost uniform level of emissions), while the emissions of the energy sector are considered as local or point sources (cover small or local areas with large amounts of emissions; hotspots). Although waste emissions appear in both area and point sources depending on their location, overall, it is distributed more like local sources across the domain (Figure 7.5k). All these specific characteristics enable us to evaluate and distinguish the differences between the responses of each sector to the inversion. Note that the magnitude of those emissions sectors is largest for wetlands, followed by agriculture, energy, and waste.

For the Type 1 inversion, the posterior emissions for each sector show an overestimation of large emissions and an underestimation of smaller emissions in the same

sector. More importantly, the biased forecast initial state can cause a miscorrection on other emissions sectors, even though they are not perturbed (i.e., taken as true emissions) in the prior emissions—hereafter referred to as the cross-sectoral effect. For example, the posterior of the agriculture emissions in Figure 7.5c shows a noticeable overestimation of emissions at the locations of the energy sector's emissions (e.g., over Russia and near the Persian Gulf). However, this effect is largely removed in the Type 0 inversion of agriculture emissions (Figure 7.5b). This is consistent across perturbations in all sectors, and implies that a configuration of inversion that integrates the optimal initial analysis and its error covariance propagation not only performs reasonably for the incorrect emissions of the same (perturbed) sector, but can also prevent misrepresentation of other sectors when they are precisely provided in the prior.

An examination of the results of Type 1 and Type 2 inversions displays the effect of initializing the inversion with the analysis field rather than the forecast field from the state assimilation. Figure 7.5c-d (same as 7.5h-i, 7.5m-n, and 7.5r-s) shows meaningful improvements in the emissions from Type 1 to Type 2 for all sectors; however, those improvements are slightly greater for the energy and waste sectors with localized emissions than for agriculture and wetlands. A similar improvement (in the same perturbed sector) has been found when we compare Type 2 with Type 3 inversions, for which the forecast of the analysis error covariances is considered besides the analysis field (Figure 7.5d-e). Although those improvements are more substantial for the energy and the waste sectors, a slight degradation occurs in other unperturbed sectors due to a cross-sectoral effect. For instance, comparing Figure 7.5i and Figure 7.5j of the energy sector indicates that despite the improvement of the emissions in the location of the energy sources (e.g., East Asia and

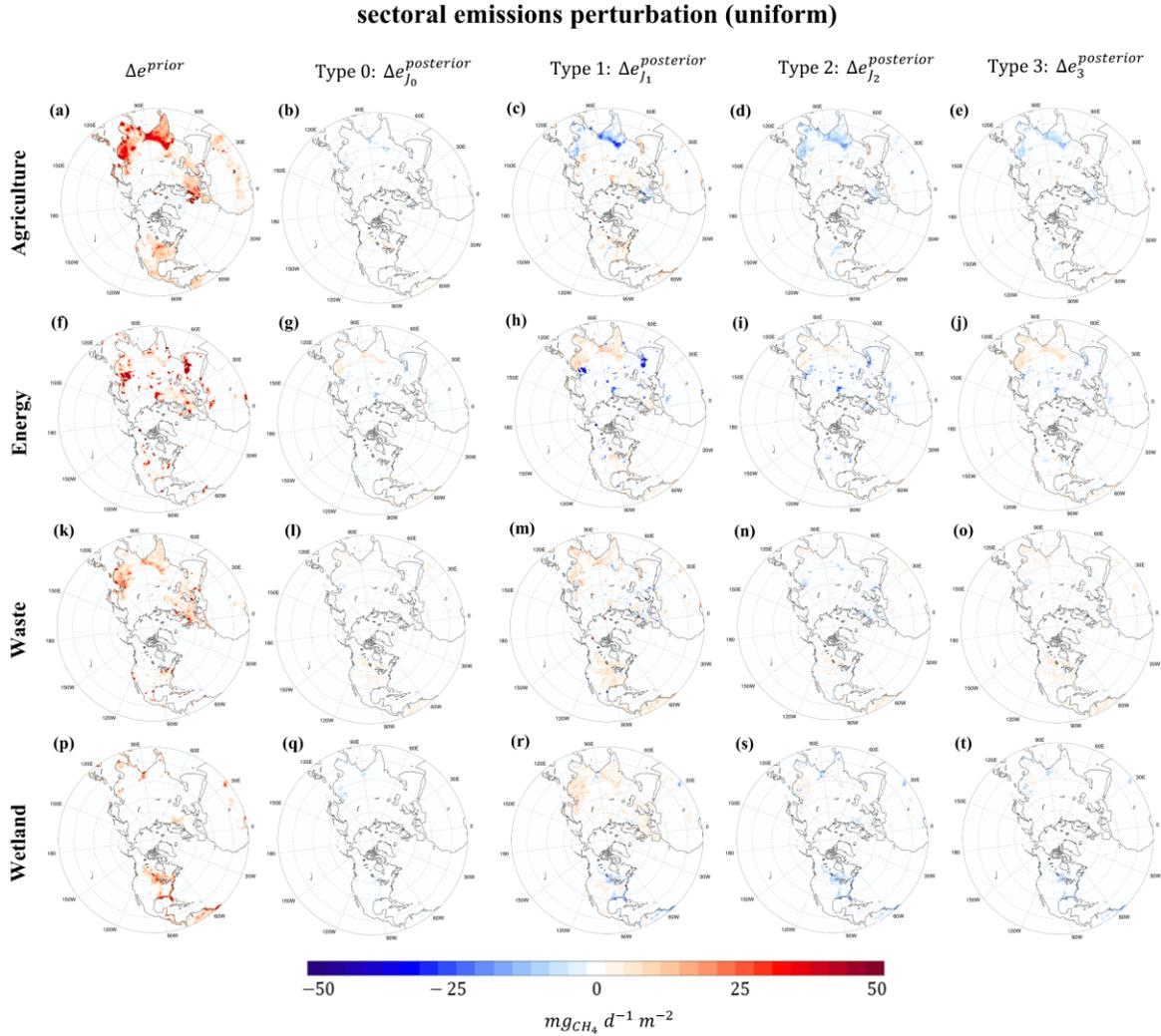
Russia), it causes cross-sectoral effects on the agriculture emissions resulting in degraded estimate over those areas (e.g., India and Southeastern Asia). A similar pattern is also shown in the waste sector (Figure 7.5n-o), where the estimation over agriculture and wetland areas is slightly worsened (e.g., Midwestern U.S. and Boreal regions in North America). Nevertheless, we found that the cross-sectoral effect is not noticeably destructive to the inversion of agriculture and wetland emissions (Figure 7.5d-e and Figure 7.5s-t) when we account for the analysis error covariances in Type 0 inversion. All that considered, the undesirable effect between sectors suggests that the approximated correlation model (Equation (7.7)) with a fixed correlation length across the domain is not sufficiently informative to resolve the structure of the large and localized emissions in the energy and waste sectors. Furthermore, we imply that the weight of the forecast error covariance ( $H^o \mathbf{P}_t^f (\mathbf{A}_1) H^{o T}$ ), compared to the observation error ( $\mathbf{R}$ ), is likely too small for those sectors; thus, the estimation system relies more on the model and prior information (which are more uncertain) than on the observations to correct emissions. One way to partly alleviate this is to increase the forecast error covariance by adding extra (bulk) model transport error ( $\mathbf{Q}$ ). We consider the implications of this model transport error in the remainder of this section.

Comparing Type 3 with Type 0 inversion helps us understand the influence of model transport error  $\mathbf{Q}$  in reproducing true emissions. From an estimation point of view, model transport error  $\mathbf{Q}$  compensates for those missing error variance and correlations that are, in fact, unexplainable by the error propagation scheme (i.e., advection of variance herein; see Section 7.6.1 for details). Our results (Figure 7.5b,r; Figure 7.5h,j; Figure 7.5m,o; Figure 7.5q,t) show that accounting for the model transport error  $\mathbf{Q}$  improves

emissions estimation for all sectors. However, this level of improvement is tangibly larger for the agriculture and wetland than the energy and waste sectors. For example, adding the estimated model error to the Type 3 inversion of agriculture emissions (Figure 7.5b,e) can substantially remove the underestimate in the posterior emissions, particularly over Northern India and Southeast Asia. A similar level of improvement for the wetland emissions over the boreal region and the southeastern United States is obtained (Figure 7.5r,t). The influence of model error, however, is not as significant for the point sources of energy and waste emissions. Nevertheless, adding model error can still be important for the cross-sectoral effects of a sector primarily composed of point sources (e.g. energy) on a sector that is more spatially extensive. For instance, for the inversion of energy (Figure 7.5j) and waste (Figure 7.5o) emissions, the overestimation of the agriculture (e.g., Southeast Asia) and wetland (e.g., boreal area) emissions are decreased by considering model transport error  $\mathbf{Q}$  (by comparing them with Figure 7.5g and Figure 7.5i).

In our estimation, the model transport error ( $\mathbf{Q}$ ) is assumed to be proportional to the variance field, in line with the previous study by Voshtani et al. (2022a), it provides a uniform and bulk impact on the total error across the domain (see Section 7.6.1 and Figure 7.8) due to the rather homogeneous distribution of methane. Therefore, as expected, those emissions sectors covering the broader area, such as agriculture and wetland, are more susceptible to being influenced by the model transport error. On the other hand, the model transport error has less spatial variability than the emissions and thus has little chance of affecting a point source estimation, even though it may have a large magnitude. Perhaps, another form of model transport error, for which the spatial pattern is different, may

improve the influence of model error on the local and large sources, such as energy and waste emissions.



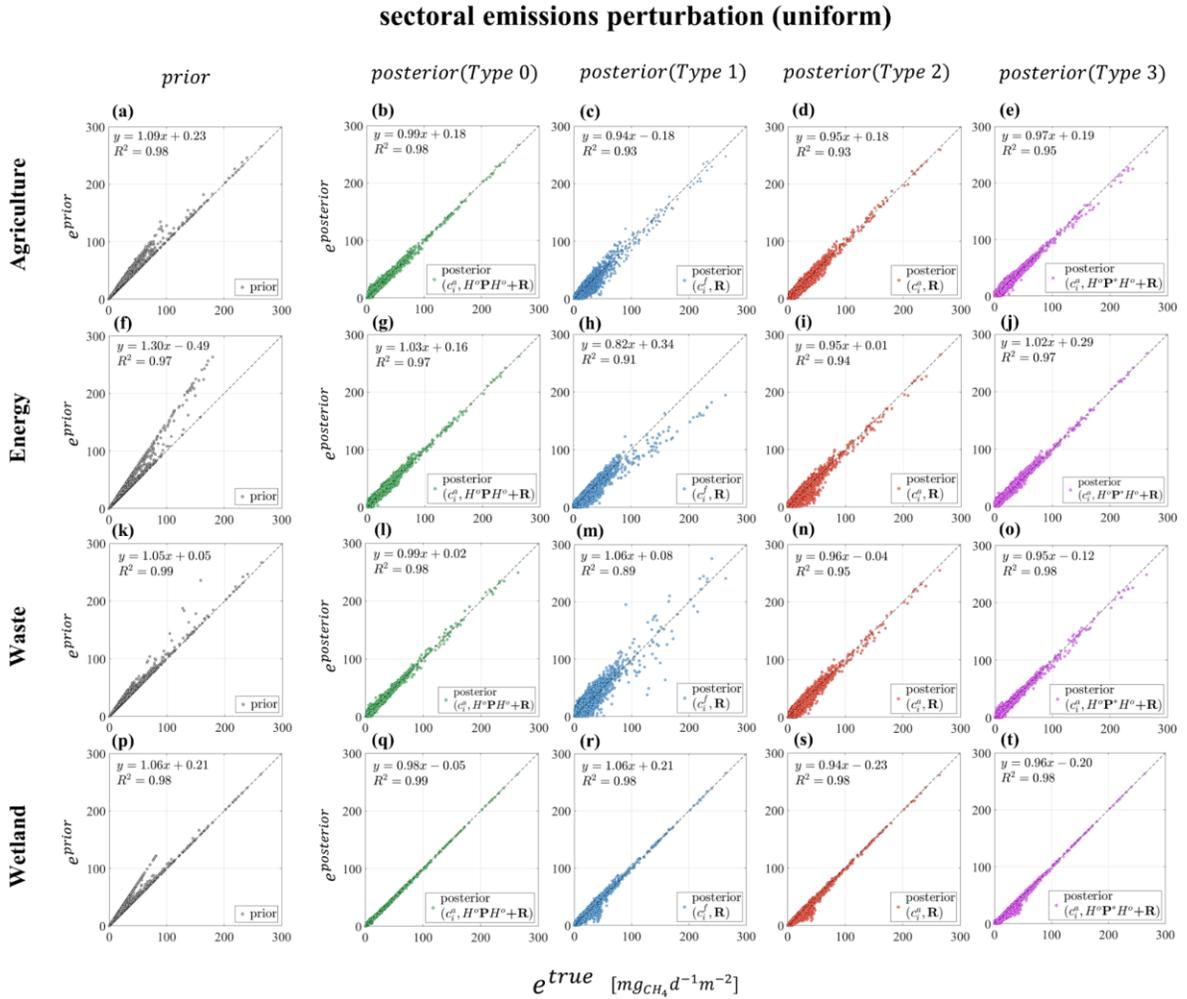
**Figure 7.5.** (a-e) prior – true emissions and comparison of the posterior against true emissions for the only perturbed agriculture sector, (f-j) only perturbed energy sector, (k-o) only perturbed waste sector, and (p-t) only perturbed wetland sector. Prior emissions are generated using 50% uniform perturbation. Type 1 OSSE uses  $(c_1^f)$  and (R). Type 2 performs with analysis initial  $(c_1^a)$  and (R). Type 3 OSSE operates with analysis initial  $(c_1^a)$  and forecast of analysis error covariance  $(H^o P^f(A_1) H^{oT})$ , and Type 0 OSSE works with analysis initial  $(c_1^a)$  and forecast of analysis error covariance with model error  $(H^o P^f(A_1, Q) H^{oT})$ .

The simple fact that the correction for the point sources is not as efficient may indicate that the estimation of the point sources and area sources should be treated with a different modelling framework. In general, due to limited information and large uncertainty about the origin of model error, finely resolving its spatial structure is typically a nontrivial task (Tandeo et al. 2020; Stanevich et al. 2021). Note that we do not separately obtain an estimation of model transport error for each inversion process of this study; instead, we use the estimated parameter following Voshtani et al. (2022b) (see Section 7.3.1 for details). Thus, the error variance associated with the model transport may also not be the optimal one in our analyses here.

Figure 7.6 compares the statistics of the sectoral perturbation cases against four types of inversion schemes, analogous to the maps in Figure 7.5. Comparing Type 1 and Type 2 inversions shows an improvement of the fit to the true emissions for the energy and waste sectors, such that their  $R^2$  increase from 0.91 to 0.94 and from 0.89 to 0.95, respectively (Figure 7.6h-i, Figure 7.6n-m). Those improvements, in fact, occur in the form of mainly bias removal together with variance reduction of the posterior emissions. For the agriculture and wetlands emissions, the posterior emissions do not retain a better fit to true emissions ( $R^2$  remains unchanged), but some reduction is observed in the variance (Figure 7.6c-d, Figure 7.6r-s).

The effect of accounting for the forecast of analysis error covariance (comparison between Type 2 and Type 3) appears as a moderate improvement in the fit of all sectors except wetlands. This is perhaps due to the larger magnitude of other sectors' emissions relative to that of wetlands, which results in an overall small change in the total posterior

emissions. Lastly, we compare the effect of model transport error ( $\mathbf{Q}$ ) on the fit to the true emissions for the four sectors.



**Figure 7.6.** (a-e) statistical comparison of prior and posterior emissions against true emissions in scatter plots for the only perturbed agriculture sector, (f-j) only perturbed energy sector, (k-o) only perturbed waste sector, and (p-t) only perturbed wetland sector. Prior emissions are generated using 50% uniform perturbation of each sector individually.

While updating the initialization field to use the analysis rather than the forecast does not improve the inversion for spatially extensive sources like agriculture and wetlands (compare Type 1 and Type 2), if we additionally account for model transport error (compare Type 3 and Type 0) we observe improvements to the fit to the true emissions for

the agriculture and waste sectors, as  $R^2$  increases from 0.95 to 0.98 and from 0.98 to 0.99, respectively (Figure 7.6b-e, Figure 7.6q-t). This small improvement also occurs in the form of both bias removal and variance reduction of posterior emissions. The further details of the statistics are expressed later in Section 7.5.4.

### 7.5.3 More Realistic Perturbations

OSSE experiments in this section consider a more realistic inversion scenario, aiming to provide a random-like (more objective) perturbation in the prior emissions. One way to achieve this is to perturb each sector individually with different weights and signs ( $\pm$ ) of perturbations while taking them all together for the inversion analysis. Previous methane inversion studies (Bergamaschi et al. 2018; Maasakkers et al. 2019; Janardanan et al. 2020; Saunois et al. 2020; Qu et al. 2021) showed that, overall, agriculture and waste emissions are underestimated, whereas the energy and wetlands are overestimated globally in the prior (mainly based on EDGAR for the anthropogenic inventory and WetCHARTS for the wetlands). Thus, the energy and wetland sectors are perturbed upwards by 50% and 25%, respectively, and the waste and agriculture sectors are perturbed downwards by 50% and 25%, respectively.

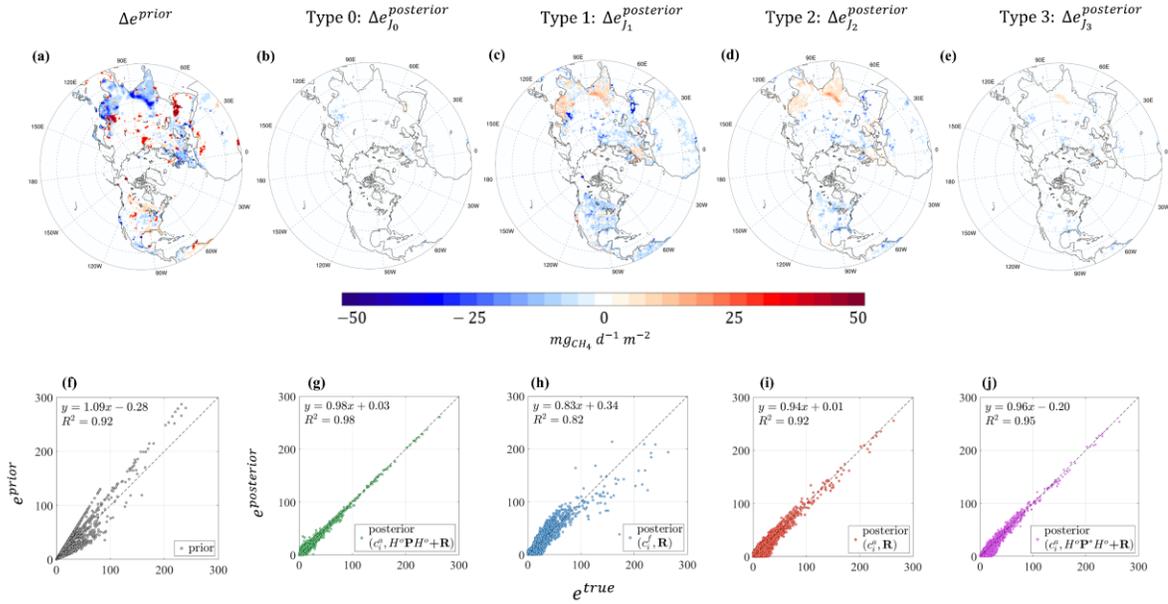
Figure 7.7 summarizes the performance of our four different types of inversion using these prior emissions. In Type 1 inversion, the posterior emissions exhibit significant over- and underestimations (Figure 7.7c). This plot also shows that the weight of the spatial biases in the posterior is almost proportional to the prior emissions, yet agriculture and waste retain a positive bias contrary to the energy and wetlands with a negative bias. Overall, it implies that simply relying on a perfect model initial field results in overcorrections of posterior emissions in many regions. The statistics of the posterior

emissions in Figure 7.7h also indicate that besides a large domain-wide variance, they are negatively biased in large emissions areas over East Asia (from coal emissions) and near the Persian Gulf (likely due to oil and gas emissions). Furthermore, a lower  $R^2$  and weaker regression line of the posterior than the prior suggests that the posterior emissions are likely less reliable than the prior.

When we compare Type 1 inversion with Type 2, where the PvKF analysis provides the initial state, but the inversion relies on observation error covariance  $\mathbf{R}$ , we find a significant improvement in the spatial biases of posterior emissions. This is particularly true for the large sources, which further suggests that the large domain-wide variance and bias of emissions are corrected. The  $R^2$  of the Type 2 inversion also exhibits a large increase compared to Type 1 but is still comparable to the prior statistics. Type 3 inversion (Figure 7.7e,j), which includes the model-propagation of analysis error covariance, maintains a large consistent improvement everywhere compared to the Type 2 inversion. This improvement is reflected in  $R^2$  and the slope of the regression line.

The effect of adding model error ( $\mathbf{Q}$ ) is also examined and seen in the further improvement of posterior emissions in Type 0 (Figure 7.7b) relative to Type 3, which is also confirmed by their statistics (Figure 7.7g). It indicates that adding model error ( $\mathbf{Q}$ ) to the forecast error ( $H^o \mathbf{P}_t^f (\mathbf{A}_t) H^{o T}$ ) can, in fact, help better constrain the emissions through inversion. In addition to this experiment, we conduct another perturbation of those sectors but using different sectoral weights (see Figure E.1 in Appendix E1), and overall, it shows a similar correction behaviour as this experiment in Figure 7.7.

**total emissions perturbation (non-uniform)**



**Figure 7.7.** (a) prior – true emissions ( $\pm 25\text{-}50\%$  variable perturbation); posterior – true emissions in (b) Type 0 inversion, (c) Type 1 inversion, (d) Type 2 inversion, and (e) Type 3 inversion. Statistical comparison of the (f) prior emissions and (g-j) posterior emissions of Type 0-3 inversions, respectively.

### 7.5.4 Overview of the Different Experiments

Three forms of perturbations in our OSSE experiments are described in Sections 7.5.1 to 7.5.3. We examined the ability of four different cost functions for each perturbation type to reproduce true emissions across the domain. Here, we summarize those experiments with further evaluations in terms of three metrics, including normalized mean bias (NMB), normalized mean error (NME), and Pearson's correlation coefficient. Accordingly, we have

$$\text{NMB} = \frac{\sum_{k=1}^N (e_k - e_k^t)}{\sum_{k=1}^N (e_k^t)} \quad (7.22)$$

$$\text{NME} = \frac{\sum_{k=1}^N |e_k - e_k^t|}{\sum_{k=1}^N (e_k^t)} \quad (7.23)$$

$$R = \frac{\sum_{k=1}^N |(e_k^t - \bar{e}_k^t)(e_k - \bar{e}_k)|}{\sqrt{\sum_{k=1}^N (e_k^t - \bar{e}_k^t)^2 \sum_{k=1}^N (e_k - \bar{e}_k)^2}} \quad (7.24)$$

where  $e$  and  $e^t$  denote the posterior and true emissions, respectively, and  $N$  represents the number of grid cells with emissions.

The results in Table 7.3 indicate that NMB significantly decreased between Type 1 and Type 2 inversion in almost all emissions perturbation cases, particularly in Case 1 with a 28% reduction and Case 6 with a 15% reduction, where all sectors are perturbed. It suggests that initialization of the inversion with a biased model forecast ( $c_1^f$ ) reflects on the posterior emissions mainly as a form of residual biases. In addition, NME decreased and  $R$  increased for all cases, indicating that the posterior emissions residuals are smaller for Type 2 inversions. In fact, at the grid level, they are closer to the true emissions. Overall, a similar reduction of NMB and NME and an increase of  $R$  occurs between Type 2 and Type 3 inversion for all cases. This suggests that incorporating the model-propagated analysis error covariance ( $H^o \mathbf{P}_t^f (\mathbf{A}_1) H^{oT}$ ) can also substantially impact our inversion results to recover the true emissions. Finally, the effect of the model error ( $\mathbf{Q}$ ) is shown by comparing Type 3 and Type 0 inversion. Although we implemented a simple form of  $\mathbf{Q}$ , the results of all metrics slightly improved for perturbations with all emissions (Case 1 and Case 6); however, it may influence each sector differently.

**Table 7.3. Normalized mean bias (NMB), the normalized mean error (NME), and Pearson's correlation coefficient ( $R$ ) for each emissions perturbation case and inversion cost function (Equations (7.16) to (7.19)).**

Cost function Perturbation	Type 0: $J_0(c_i^a, \mathbf{P}_i^f(\mathbf{A}_1, \mathbf{Q}), \mathbf{R})$			Type 1: $J_1(c_i^f, \mathbf{R})$			Type 2: $J_2(c_i^a, \mathbf{R})$			Type 3: $J_3(c_i^a, \mathbf{P}_i^f(\mathbf{A}_1), \mathbf{R})$		
	NMB	NME	$R$	NMB	NME	$R$	NMB	NME	$R$	NMB	NME	$R$
Case 1: All sectors/Uniform	+0.02	0.06	0.98	-0.39	0.57	0.88	-0.11	0.29	0.94	-0.03	0.10	0.97
Case 2: Agriculture/Uniform	+0.01	0.04	0.99	-0.07	0.28	0.96	-0.05	0.12	0.93	0.00	0.06	0.95
Case 3: Energy/Uniform	+0.03	0.03	0.98	-0.18	0.31	0.95	-0.09	0.22	0.94	+0.03	0.03	0.97
Case 4: Waste/Uniform	+0.02	0.02	0.99	+0.11	0.45	0.94	-0.03	0.10	0.95	-0.02	0.03	0.98
Case 5: Wetland/Uniform	-0.01	0.01	0.99	-0.06	0.11	0.99	-0.05	0.09	0.99	-0.05	0.04	0.99
Case 6: All sectors/Non-uniform	-0.02	0.05	0.99	+0.22	0.37	0.90	-0.07	0.19	0.92	-0.05	0.11	0.95

In fact, agriculture and wetland are more sensitive to model error as all the metrics are altered to provide a better fit to the true emission. On the other hand, it has little impact on the energy and waste sector. Those behaviours are mainly attributed to the spatial characteristics of model error that are more consistent with those sectors with broader areas and uniform emissions (area sources). We detail the underlying assumptions for model transport error ( $\mathbf{Q}$ ) as well as its effect on the inversion in the next section.

## 7.6 Additional Implications for the Proposed Source Estimation

### 7.6.1 Implications for the Forecast Model Error

Model-propagated forecast error,  $\mathbf{P}^f(\mathbf{A}_1, \mathbf{Q})$ , (Equation (7.14)) not only depends on the error covariance of the initial state, but can also be produced due to imperfections in the model transport. We recall that for practical purposes, our inversions neither rely on a perfectly known initial state nor a perfect CTM. We integrate the PvKF formulation that can cost-effectively provide error information on the initial state and the model transport (Voshtani et al. 2022b) (see Section 7.3.3). Furthermore, as shown earlier in Section 7.5.1- to Section 7.5.4, accounting for those errors during the inversion process can improve emissions estimations; hence, it is quite important also to understand the causes of those improvements.

Using the PvKF assimilation, we obtain the optimal estimation of the analysis error covariance ( $\mathbf{A}$ ) that is used to initialize the propagation of errors during the inversion. Propagation of this error covariance ( $H^o \mathbf{P}_t^f(\mathbf{A}_1) H^{oT}$ ) is the key element in forming correlations that exist by nature in the model space but are often missed in a typical 4D-Var inversion (see Figure 7.2 and Sections 7.3.3 and 7.3.4); thus, not accounting for this error may cause significant degradation of the inversion results (Figure 7.4, Figure 7.5, and Figure 7.7).

Nevertheless, the entire error during inversion does not necessarily originate from the initial state, but may also be compounded by the model transport, given that the model is not perfect. As in any Kalman filter, the PvKF does not depend on a perfect model assumption and thus allows here for the inclusion of the model transport error during the inversion window. Contrary to the model propagation of the initial analysis error

covariance ( $H^o \mathbf{P}^f(\mathbf{A}_1)H^{oT}$ ), which involves the finer structure of the model background correlations, model (transport) error covariance ( $\mathbf{Q}$ ) is assumed to be proportional to the field due to a lack of information about it. In fact, there is little knowledge about the underlying processes driving model error covariance  $\mathbf{Q}$  (Locatelli et al. 2015; Stanevich et al. 2021).

The origin of  $\mathbf{Q}$  is generally unknown, so that quantifying its underlying structure is almost implausible (Tandeo et al. 2020). It has been shown in previous studies as well as here that the effect of model transport error is not negligible (Polavarapu et al. 2016; Stanevich et al. 2021), while there are several explanations for their causes. In this study, the effect of adding model transport error can be explained in two ways based on their forms and effects (although their origins are not quantifiable). The first type is a domain-wide stationery (bulk) error associated with various modelling characteristics. Using criteria based on innovation variance consistency in PvKF assimilation according to Voshtani et al. (2022b), we obtain the estimated value of the corresponding model error covariance  $\mathbf{Q}$ , hereafter referred to as  $\mathbf{Q}_{PvKF}$ , which is used in our OSSE experiments (i.e.,  $\mathbf{Q} = \mathbf{Q}_{PvKF}$ ). Note that numerical discretization error was identified as the first source of error to be associated with the need for this model error covariance  $\mathbf{Q}$  (Pannekoucke et al. 2016, Menard et al. 2021, Gilpin 2022). In particular, Menard et al. (2021) argue that although we rely on the continuous behaviour of the linear advection, a discretized scheme used in the model can lead to a loss of total variance. Therefore, part of this bulk model transport error tends to compensate for that loss. Here in this section, we also discuss the true model error, which may be caused by neglecting the explicit diffusion of the model.

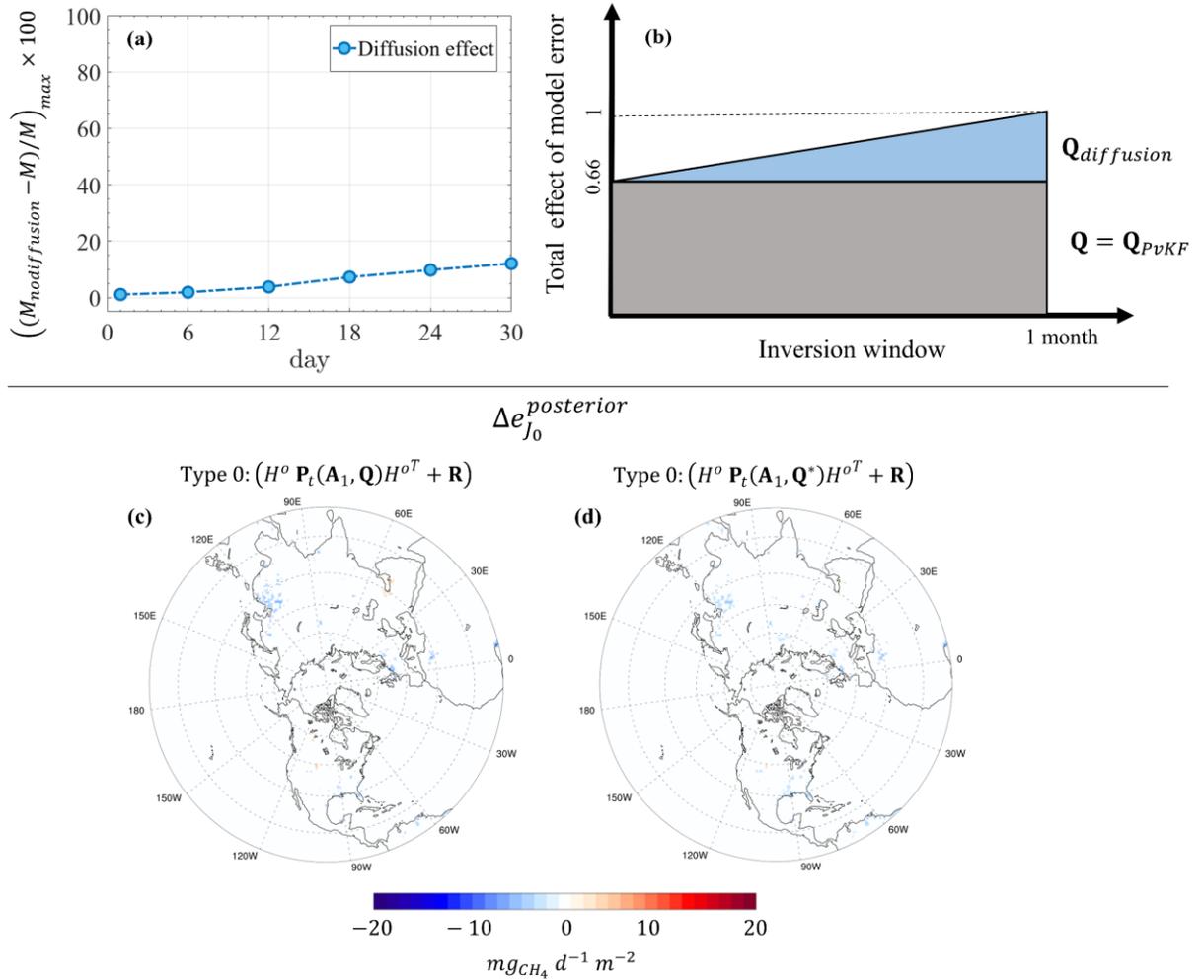
The model error can also be produced due to limited assumptions employed in our inversion/assimilation system. For example, here, the propagation of the error during inversion or assimilation is performed using a continuous formulation based on the advection of variance (Equation (7.6)). It has been shown previously that over the 3-day revisit time of GOSAT, that assumption is fairly reliable since the total variance remains conserved to a great extent (>95%) (Voshtani et al. 2022a). Relying on that assumption for a month of inversion in our experiments may lead to a degradation in realistically simulating the forecast model error. In fact, for the extended period of a month, diffusion spreads a portion of the variance in the form of spatial correlation, which is not considered in error propagation formulation based on the advection-only scheme used in PvKF. Those missing correlations might be addressed using a general form of parametric Kalman filter (Pannekoucke et al. 2016), where besides the variance, the evolution of characteristic parameters of correlations are computed, yet at a sizeable extra cost. Nevertheless, using a simple form of model error, we can compensate for the missing correlations due to diffusion by approximating its impact on the model concentrations.

We perform a series of forecast simulations experiments (similar to a one-observation experiment according to Voshtani et al. (2022a)), while testing how the forecast error variance remains conserved after a month of integration. We approximate the evolution of error variance as proportional to model forecast concentrations (see Section 3.1 for the relative weights of the initial and model errors to the field) while considering the model with an active or deactivated diffusion scheme. Our results indicate that a maximum of 12% violation of the innovation covariance consistency occurs after one month if we do not account for the model error covariance  $\mathbf{Q}$  (Figure 7.8a). Figure E.4

in Appendix E3 also shows the map of methane concentrations of that difference in the first layer of the model after one month. We consider this modelling effect (unaccounted diffusion of error variance) on the forecast error covariance as the diffusion effect ( $\mathbf{Q}_{diffusion}$ ) of model error covariance. Thus, it is recommended to include  $\mathbf{Q}_{diffusion}$  for the extended period of error propagation (e.g.,  $\Delta T \geq$  one month) using the PvKF advection-only scheme. We also found an approximately linear behaviour of this effect on the total model error, as shown in Figure 7.8a. To compare  $\mathbf{Q}_{diffusion}$  with the model error covariance that is already estimated ( $\mathbf{Q}_{PvKF}$ ), we perform the same procedure to retain its impact on domain concentrations over the same month of integration. Our results indicate that the effect of  $\mathbf{Q}_{PvKF}$  is nearly two times larger than the effect of  $\mathbf{Q}_{diffusion}$  at the end of a month-long simulation. Figure 8b shows the relative weight of these two forms of model error.

To test the effect of  $\mathbf{Q}_{diffusion}$  on emissions inversion results, we assume a simple form of that error (Figure 8b) as explained earlier (proportional to the field and linear increment over time). We repeat our OSSE experiment in Section 5.3 with the same inputs and configuration, except that the model transport error covariance  $\mathbf{Q}$  is replaced with  $\mathbf{Q}^*$  in Type 0 inversion (see Table 2). Besides the estimated model transport errors ( $\mathbf{Q}_{PvKF}$ ),  $\mathbf{Q}^*$  includes model error covariance due to the diffusion effect ( $\mathbf{Q}_{diffusion}$ ), thus  $\mathbf{Q}^* = \mathbf{Q}_{PvKF} + \mathbf{Q}_{diffusion}$ . Figure 8c-d compares the spatial distribution of posterior – true emissions for two inversion cases: (i) with  $\mathbf{Q} = \mathbf{Q}_{PvKF}$  and (ii) with  $\mathbf{Q}^*$ . The result shows a similar spatial distribution of posterior emissions with minor changes in magnitude over some large emissions areas. It implies that the additional part of the model error covariance

due to diffusion ( $\mathbf{Q}_{diffusion}$ ) has a rather small impact on recovering true emissions, although it was shown earlier that removing the entire model transport error covariance can exert a substantial impact on the inversion results.



**Figure 7.8.** (a) Diffusion effect of model transport error ( $\mathbf{Q}_{diffusion}$ ) estimated using a series of forecast simulations. (b) schematic form and weight of estimated part of the model error or  $\mathbf{Q}_{PvKF}$  using the advection-only scheme, compared with  $\mathbf{Q}_{diffusion}$ . Comparison between posterior – true emissions of Type 1 of Section 5.3 (based on  $\pm 25$ -50% non-uniform perturbation) where (c) uses model error covariance  $\mathbf{Q} = \mathbf{Q}_{PvKF}$  and (d) uses model error covariance  $\mathbf{Q}^* = \mathbf{Q}_{PvKF} + \mathbf{Q}_{diffusion}$ .

## 7.6.2 Computational Timing of Different Inversions

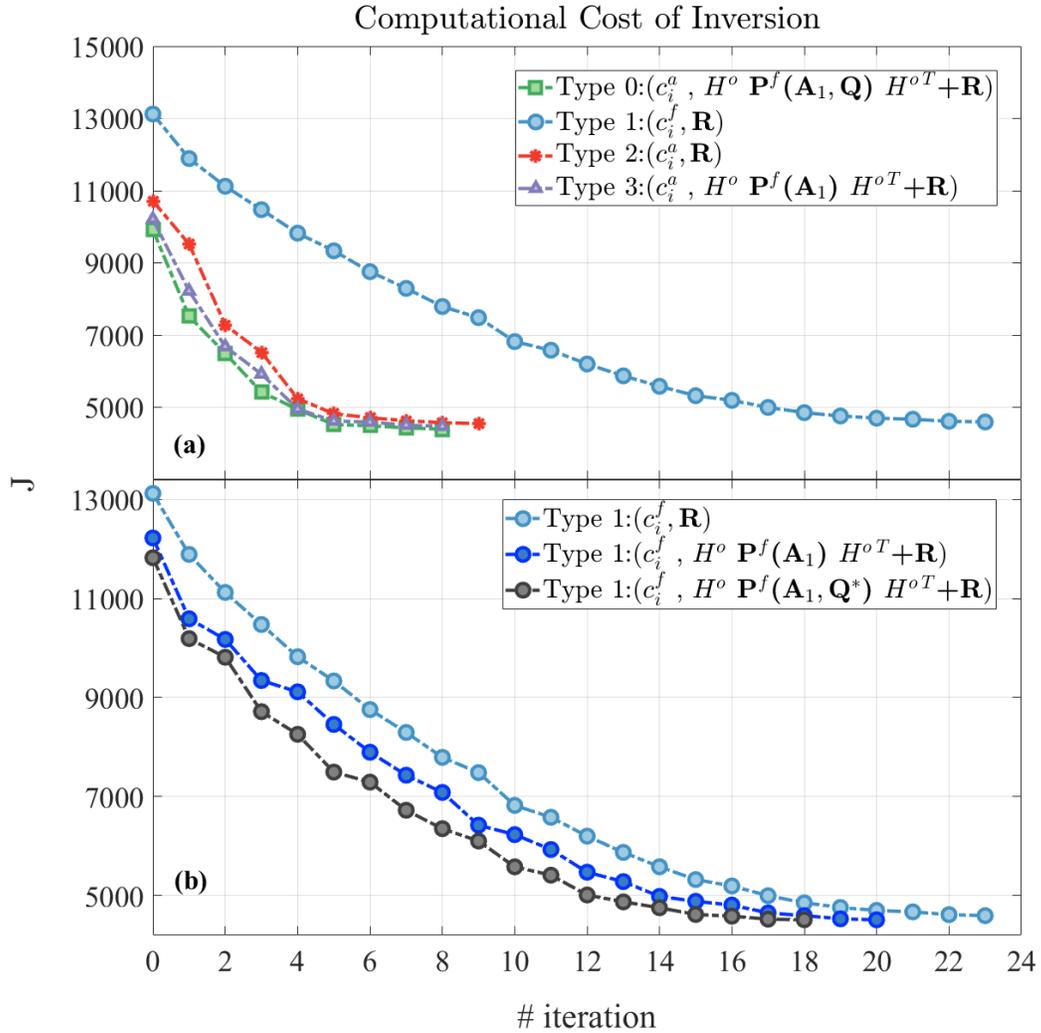
We showed earlier in our OSSE results in Section 7.5 that both initial analysis field and model-propagated analysis error covariance can exert a tangible impact on reproducing the true emissions. Here, we evaluate the computational aspect of our modified inversion schemes that link PvKF assimilation to 4D-Var inversion of methane emissions. We recall that the 4D-Var inversion is designed in a way that the estimation iteratively converges to a local minimum (considered as an optimal solution) by minimizing a quadratic cost function ( $J$ ) of the residuals between model and observations (Equation (7.8)). It is also assumed that when the decrease in the cost function between two successive iterations is less than 1%, the iteration process will be terminated (see Section 7.3.3). Using consistent convergence criteria for all experiments, we compare the computational time of employing different cost functions (Type 0-3 inversions; see Section 7.4.2, Equations (7.16) to (7.19)). Accordingly, Figure 7.9 depicts the value of the cost functions  $J$  at each iteration until convergence for all types of inversions. Note that the results here are associated with the OSSE of variable perturbations as described in Section 7.5.3, Figure 7.7.

Our results in Figure 7.9 indicate that performing Type 1 inversions with a biased model forecast initial and assuming a perfectly known state of the model (light blue line) requires 23 iterations for convergence. However, only accounting for an improved initial field (red line) from our optimal PvKF analysis ( $c_i^a$ ) significantly reduces the computational cost of inversion. In fact, the number of iterations reaches 9 in Type 2 (~ one-third compared to the cost of Type 1). Now, when model-propagated error covariance ( $H^o \mathbf{P}^f (\mathbf{A}_1) H^{oT}$ ) is also considered in our cost function of Type 3 (purple line), the number of iterations for convergence becomes 8, indicating that the cost of inversion does not

change significantly. A similar effect is observed when the model transport error is considered (green line) in our cost function Type 0 ( $H^o \mathbf{P}^f(\mathbf{A}_1, \mathbf{Q}) H^{oT}$ ), and the number of iterations remains the same. These comparisons imply that although accounting for the error covariance, either from imperfect transport or the initial state, does not provide a computational benefit, it exerts a substantial impact on the quality of the posterior emissions (e.g., see Table 7.3). On the other hand, considering the optimal initial analysis field maintains not only a better emissions constraint but also a lower cost of inversion. Note that the PvKF before the inversion window is computationally inexpensive, as it requires a little more than two model runs (Voshtani et al. 2022a). Hence, the overall cost of inversion when the PvKF assimilation cost is also considered remains low enough compared to the cost of Type 1 inversion (without PvKF assimilation).

Consistent with the experiments in Section 7.5, in Figure 7.9a, we only look at the marginal difference in adding improved error statistics to the inversion once the initial state has already been corrected with assimilation analysis. Now, in another experiment, we keep the initial field of the model forecast and replace only the error covariance term in the cost function (Figure 7.9b). In fact, during the assimilation window, besides the model forecast ( $c^f$ ), we account for the model propagation of error covariance without updating them by observations assimilation. Accordingly, at the end of the assimilation window, instead of the analysis error covariance matrix  $\mathbf{A}_1$ , we obtain the forecast error covariance  $\mathbf{P}_1$  to initialize the inversion. In addition, an updated form of model transport error ( $\mathbf{Q}^*$ ), as described in Section 7.6.1, is used in this experiment. Therefore, we can test if the error statistics alone (due to initial or model transport) are sufficient to reduce the computational cost.

Figure 7.9 presents three forms of inversions, all of which carries on the same model forecast ( $c^f$ ). Besides the Type 1 baseline inversion (light blue line), the figure shows an inversion (dark blue line) where the model-propagated forecast error is considered ( $H^o \mathbf{P}^f (\mathbf{P}_1) H^{o T}$ ), and another inversion (grey line) where model transport error is added to the forecast error ( $H^o \mathbf{P}^f (\mathbf{P}_1, \mathbf{Q}^*) H^{o T}$ ). The results of the computational cost of inversions indicate that accounting for those error covariances in the cost function can reduce the number of iterations to 20 and 18, respectively. It suggests that the effect of error statistics alone would reduce the computational cost by up to ~20%. Although this amount is greater than the cost of the marginal difference of adding error statistics once the analysis is already used as the initial field (~10% reduction in cost; Figure 7.9a), it is still significantly small compared to the cost of replacing the initial forecast field by the analysis (~65% reduction). Note that the effect of error statistics alone (given the same  $c^f$ ) on recovering the spatial distribution of true emissions is also more significant than the marginal effect of adding those errors when the analysis field is used (results are not shown here). Finally, we remark that using the updated form of model transport error ( $\mathbf{Q}^*$ ) than the initial form ( $\mathbf{Q}$ ) provides no further reduction of the computational cost of inversion (see Figure E.3 in Appendix E3) and only a minor improvement for the spatial distribution of posterior emissions (Figure 7.8c-d).



**Figure 7.9.** (a) Comparison between the value of cost function against the number of iterations to show the computational cost of four different inversion types (Type 0-3) provided in Equations (18-21); (b) Comparison between the computational cost of inversions when the initial field is provided with the model forecast ( $c_i^f$ ) and same for all the cases (Type 1 inversion), but the impact of adding only error statistics to the cost function is shown.

## 7.7 Summary and Conclusions

We present a new approach for performing methane source estimation, where the PvKF assimilation system of methane concentrations is combined with the 4D-Var source inversion. Previous methane inversion studies typically assume that the initial state

uncertainties are negligible compared to the effect of accumulated emissions uncertainties for a typical duration of methane inversion (e.g., one month) in a limited domain. As a result, the state is considered to be close to the truth, and thus the inversion is nearly insensitive to the initial state uncertainties. However, in this study, we not only produce an assimilation analysis with small uncertainties to initialize the 4D-Var inversion but also account for those state uncertainties for the duration of the inversion. Our PvKF assimilation scheme provides this information. It is a lightweight assimilation system that allows for the propagation of errors using an advection scheme while it is capable of taking the model transport error approximation into account. These state estimation properties qualify us to examine their effect on the source estimation when it is linked to an inversion system.

It is also commonly assumed in methane inversion studies that errors in observation space are not correlated (or correlations are insignificant), and thus independent measurement errors dominate the total error weight, resulting in a diagonal observation error covariance. This assumption is indirectly attributed to the perfect model assumption made in many methane inversion studies. However, in our proposed assimilation-inversion system, the effect of the model forecast errors in the observation space leads to a nondiagonal error covariance matrix. This covariance matrix aims to provide appropriate correlations (corresponding to the optimal solution) when the state of the system is not considered as perfectly known, which is the case in reality. Accordingly, besides the impact of the initial analysis field provided by our PvKF assimilation, we examine the influence of forecast error covariance (model-propagated initial error and transport error included) on the inversion results.

We design observing system simulation experiments (OSSEs) to achieve our goals using the hemispheric CMAQ model and GOSAT methane simulated observations. Our source estimation system considers a monthly mean correction on methane emissions at the grid level across the domain. We construct modified inversion cost functions to account for those state characteristics, including (i) the effect of the optimal initial analysis field, (ii) the forecast of analysis error covariance, and (iii) the approximated transport error, in reproducing true emissions. In addition, different perturbations of prior methane emissions, including (i) uniform perturbations of all sectors together, (ii) individual sectoral perturbations, and (iii) variable sectoral perturbations, are generated to address the limitation of a typical 4D-Var inversion that relies on perfect state assumptions and a diagonal observation error covariance.

Our base case OSSE with uniform perturbation of total methane emissions indicate that not only the initial analysis concentrations but their model-propagated uncertainties have a substantial impact on recovering the true emissions. Comparing the proposed modified inversion cost function, which is fully linked to the assimilation of state and uncertainty (Type 0), with the regular cost function that only relies on the (biased) model forecast with the perfect assumption (Type 1) shows a considerable improvement in posterior emissions statistics. As a result, NMB and NME indicate a 37% and 51% reduction while the correlation  $R$  increase from 0.88 to 0.98. In addition, using a biased initial state (model forecast with perfect model assumption instead of analysis concentrations), results in a significant overestimation of posterior emissions in many regions with large sources. Accounting for the initial analysis field instead of forecast concentrations improves inversion results but still remains inaccurate, particularly over the

large local sources. However, including the model-propagation analysis uncertainty can significantly improve emissions constraints over those areas. This is mainly due to the structures of the error correlations that exist in the model forecast but is usually ignored by making perfect model assumptions in 4D-Var inversion. We also found that by taking the estimated model error  $\mathbf{Q}$  into account, on top of the analysis field and model propagation of uncertainties, slight overall improvements are obtained for the posterior emissions. This impact is more effective in areas where the density of observations is smaller, indicating that added model error  $\mathbf{Q}$  makes those observations more impactful to recover the true emissions.

Our results using individual sectoral perturbation also emphasize the importance of considering both the analysis field and model propagation of errors for each sectoral inversion experiment. Nevertheless, the analysis field reflects a more tangible impact on improving local (or point) sources, such as those in the energy sector, while the influence of the model error propagation is more substantial on area sources, such as those in the agriculture sector. In addition, when the initial state is biased or when the model is assumed to be perfect, inversion with only one sector perturbation negatively impacts the posterior emissions of other sectors, which were initially unperturbed. This effect is also resolved when we use the initial analysis together with model-propagated uncertainties (Type 0). Finally, variable perturbations of different sectors together (a more realistic case) are examined in our OSSE experiments. The results overall come along with our previous finding with uniform perturbations, indicating a 20% and 32% decrease in NMB and NME, respectively and an increase from 0.90 to 0.99 in  $R$ .

The computational cost of using different inversion cost functions is also examined. Using a modified form of 4D-Var inversion, as proposed in this study (Type 0), suggests a significant reduction (more than 60%) in the computational cost of inversion. This reduction occurs mainly when the initial analysis field from PvKF assimilation, instead of the biased model forecast, is used. On the other hand, the propagation of analysis errors (and transport error) in our OSSE experiment (Type 3 and Type 0) shows a negligible improvement in the computational time of inversion, although considerable improvements occur in recovering true methane emissions.

One main limitation of this modified 4D-Var source estimation, despite its practical application as well as high computational efficiency, is in the simplified assumptions for simulating and evolving errors. In fact, using an advection-only scheme for propagating errors over a month-long inversion window causes a loss of variance, which eventually may impact our ability to constrain emissions with our inversion system. Although we can partially compensate for that loss by using an extra modelling error in a simple form (see Section 7.6.1), a more precise solution for an extended period of inversion is to account for that loss by propagating the correlations using diffusion schemes (Voshtani et al. 2022b), besides the advection of variance. This, however, entails extra computational costs.

Another limitation of this approach is associated with the simple form of model transport error,  $\mathbf{Q}$ . We assumed a fairly primitive form of the model transport error, which is spatially proportional to the methane field; however, this leads to inadequate correction for the energy and waste sector contrary to the agriculture and wetlands. Thus, the structure of the model transport error can be designed more sophisticated in the future to resolve

emissions inversions of large and local sources, such as those in the energy and waste sectors.

Finally, the proposed source estimation framework provides a practical application for real observations of different types to address the limitations of similar inverse modelling and to improve the current inventories. However, the current method is incapable of estimating the error statistics of the emissions. To efficiently provide that information using the current source estimation approach, we need to further develop a coupled system of source-state estimation with joint assimilation-inversion capabilities in future. This allows one to propagate the emissions error besides the state error and estimate their uncertainties as part of the solution.

## Chapter 8: Conclusions and Future Work

Atmospheric methane is a critical short-lived greenhouse gas whose global concentrations growth has accelerated over the past decade, mainly due to anthropogenic emissions. Because of its tangible climate and air-quality impact, quite a few recent studies have mostly focused on its source estimate, particularly using satellite observations within an inverse modelling system aiming to improve the current emissions inventories. However, noticeable discrepancies have been reported among different methane emissions inventories, implying that some gaps and limitations are still involved with the current approaches. Those are partly due to errors (e.g., in the transport model and observations) that are either simply neglected, not sufficiently configured, or require extensive computational resources to be quantified using conventional approaches.

An atmospheric chemical data assimilation framework, contrary to inversion, predicts the system's concentration state (as opposed to emissions parameters). Due to the importance of methane emissions estimates, as mentioned earlier, there is a lack of methane assimilation studies in the literature. However, an assimilation system, once it is efficiently built, can be used on its own to ensure a reliable and precise estimate of concentrations, or it can be used to enable a reliable inversion by accurately providing initial and boundary conditions together with their error estimates. Nonetheless, this research aims to explore methane assimilation in-depth, not only to improve the methane model forecast, but to also examine its influence on performing an accurate source inversion. Therefore, as the first novelty, this thesis has designed a lightweight and efficient data assimilation framework, called parametric variance Kalman filter (PvKF), to estimate the methane state analysis

together with its error statistics, using GOSAT satellite observations and the CMAQ air quality model.

The PvKF assimilation scheme is well adapted to methane (quasi-) linear behaviour; it runs sequentially and relies on the continuous propagation of error covariances based on CMAQ's advection scheme. The cost of assimilation, capable of computing errors, has been estimated as approximately twice the CMAQ simulation time, which is regarded as fairly computationally inexpensive compared to other popular assimilation schemes such as 4D-Var and EnKF (as they typically cost as much as a few tens of CMAQ simulation). Furthermore, the assimilation does not rely on perfect model assumptions, contrary to the typical 4D-Var assimilation. It also avoids the loss of error variance, which is considered as a common phenomenon in standard Kalman filtering (e.g., in EnKF), causing a deficiency of the assimilation performance (i.e., perhaps filter divergence).

Before applying the newly developed assimilation to the real observations, we perform several experiments to verify that the assimilation system works as expected. Our results using synthetic observations indicate that the expected behaviour of the analysis error variance and analysis increment is fairly consistent, while the information content (i.e., total variance) is conserved. This may not be the case for EnKF if inflation is not added. We also demonstrate that the effect of a single observation can persist within a period of the GOSAT revisiting time. In addition, it is shown that the vertical error correlation could assist in deducing quantities near the surface.

The PvKF assimilation system was applied to the actual GOSAT observations within the hemispheric CMAQ domain for one month in April 2010. Our primary objective

is to obtain a high-quality analysis of methane concentrations that most accurately represent the real atmosphere. Although the analysis is generally derived from a minimum variance estimation theory in most assimilation schemes, implying an already assumed optimal solution, this optimal value may not represent the true analysis unless the input error covariances are taken as true. Therefore, we aim at the analysis that is not only an optimized state, but its error statistics have to reflect the realistic error covariances.

The cross-validation technique is an alternative to obtain the true analysis error covariances. It does not assume that the analysis is optimal, yet offers a procedure to compute one. We adopt a 3-fold cross-validation method and extend its applicability to satellite observations using an observation thinning (~10 km cut-off distance). This aims to decrease the correlation in observation space. Using the cross-validation technique and maintaining observation variance consistency, we estimate five main parameters involved in our covariance structures, including observation, background length scales, model, and initial error parameters. The analysis derived from the estimated parameters is then considered as our optimal analysis. Our finding highlights the superiority of the optimal analysis against another analysis produced using common but arbitrary values of covariance parameters.

The comparison of the model and the optimal analysis against independent and accurate observations, including TCCON and NOAA/ObsPack, suggests an overestimation of surface methane concentrations while it shows an underestimation of the upper-tropospheric model methane, both of which, regardless of their origin, are sufficiently addressed by PvKF assimilation. We also found that the assimilation of GOSAT methane induces a larger correction near the surface, where presumably there are larger errors than

in the upper troposphere due to mostly incorrect emissions. Our analysis increment of spatial patterns near the surface is coherent with recent emissions inverse modelling studies (Wang et al. 2019; Lu et al. 2021). On top of that, our estimation of the error variances shows how reliable those are. Evaluating the temporal behaviour of our optimal analysis results show a better agreement than the model against the TCCON time series. Furthermore, we found that using a set of non-optimal yet commonly used error covariance parameters can also result in an agreement with TCCON even worse than the model (without assimilation). This implies that the optimal error covariance parameters are quite essential for maintaining a reliable and high-quality analysis that improves the methane representation both spatially and temporally.

Since the PvKF assimilation system provides us with both an optimal state (i.e., analysis) and its uncertainty estimates, it can be a desirable scheme to be coupled with a source inversion system to estimate methane emissions (i.e., or any uncertain model parameters), particularly in a limited regional domain. From an estimation point of view, the contribution of emissions to the state (on a short time scale) is much smaller than the background initial or the inflow mass from the boundaries. This entails the most accurate and realistic estimate of the state. In other words, it is necessary to reduce the state uncertainty (whether used as the boundary, initial, or background conditions) to a level comparable to the emissions signal that we want to extract in inverse modelling. Therefore, in a separate analysis, we investigate the possibility of using the optimal analysis (previously demonstrated) for performing methane inverse modelling.

We evaluate the use of PvKF assimilation in conjunction with a 4D-Var source inversion. Using observing system simulation experiments (OSSEs), we verify the ability

of our modified inversion framework to recover the true emissions. Our results indicate that both the analysis field and its error covariance exert a tangible influence in lowering the bias and variance of optimized emissions. We perform different perturbations of prior emissions, aiming to address the limitation of a formal 4D-Var inversion, especially in properly resolving large sources (e.g., hotspots) and sectoral correlations for short-term analysis.

Using PvKF analysis to initialize the inversion, besides eliminating the need for a fairly extended model spin-up, improves posterior emissions estimate mainly by reducing biases (35% reduction in NMB) across the domain. On the other hand, the PvKF analysis error aims to retain the off-diagonal observational error in a dynamically coherent manner. It compensates for the error correlations in the observation network or the model forecast, which are often missed from a typical 4D-Var inversion—assuming a diagonal observational error covariance. Hence, although this can cause a minor domain-wide bias reduction, the variance of posterior emissions will be tangibly lowered in most cases (e.g., more than 50% reduction in NME). Similar results have been found against individual sectoral emissions inversion. However, the effect of PvKF analysis error in reducing emissions variance is more significant for sectors with larger area sources (i.e., agriculture and wetlands), while the impact of the analysis field initializing the inversion is more considerable for sectors with higher localized (or point) sources (i.e., energy and waste). Overall, the novelty and contribution of this research to the field can be listed as:

- The development of a new chemical assimilation framework (i.e., PvKF) for quantifying the states and uncertainties of long-lived species such as methane that

features a considerably lower cost algorithm than conventional assimilations (e.g., 4D-Var and EnKF). Chapter 5 responds to this contribution.

- An extension to the covariance modelling capable of cross-validation technique for assimilating satellite observations. This provides appropriate covariance parameters corresponding to the optimal solution (i.e., optimal analysis). Chapter 6 provides the details of this original finding.
- A suite of experiments to examine and present the significance of estimating key error parameters such as correlation length scales and observation and modelling error parameters for assimilation of long-lived species such as methane. Chapters 5 and 6 mainly correspond to this novelty.
- Optimal estimation of the state concentrations and propagation of their uncertainties exert a substantial impact on 4D-Var inversion, which is quantifiable. The new framework that links PvKF assimilation to 4D-Var inversion helps improve methane emissions constraints using satellite observations. Chapter 7 details this contribution.

Although the PvKF assimilation system offers a computationally cost-effective and robust estimation, it imposes some limitations. The main limitation of this method is related to the species' lifetime. PvKF is well-adapted to long-lived pollutants with near-linear behaviour, such as methane, and still applies to shorter lifetime chemical species, but with the caveat that a smaller fraction of the total forecast error variance is explained by the advection of error variance. In this case, the residual error variance (i.e., unexplained error variance) is captured by the stationary model error, which usually has a rather primitive

structure, as in this study. (i.e., a more sophisticated design of the model error is required along in this case)

Another limitation is that the framework's feasibility depends on the observation characteristics (e.g., observation number and density). The larger number of observations we assimilate, the more accurate analysis we may obtain. However, the assimilation scheme is limited to a certain number of observations due to the computational capacity of PvKF, as explained in 5.5.2. In addition, increasing the number of observations can result in a higher spatial density, which increases the error correlation in observation space. This contradicts the necessary condition to obtain an optimal PvKF analysis (see Section 6.2 and Section 6.3). Therefore, the number of observations may limit both the efficiency and the quality of the assimilation. In spite of that, using GOSAT observations (with < 300/hr retrievals), we found that the PvKF algorithm is sufficiently adaptable as a lightweight scheme for carrying long-lived tracers (e.g., methane) inside a chemistry-enabled atmospheric model. Some suggestions for future work are expressed in the following:

- The work concluded in this thesis showed the significant role of a (near) optimal analysis derived from a novel PvKF assimilation. This becomes more important when it comes to the source inversion problem. Although it is demonstrated in Chapter 7 that the state analysis and analysis errors have a substantial influence in a 4D-Var source inversion, they may not fully inform the inversion system. In other words, since the state and source are not estimated in a manner to interactively impact each other, the finest estimate may not yet be obtained. One apparent attempt to achieve this in future studies is to design and apply a joint source-state system for estimating methane emissions and concentration simultaneously with

their error statistics. Given the capability of the PvKF assimilation to rapidly and optimally maintain the state estimate along with the error covariances, one can extend its application to an augmented state (i.e., state vector of concentrations and emissions) aiming at a powerful inversion-assimilation system.

- It has been shown that the PvKF assimilation is well adapted to methane state estimation. One can expand the application to other species with different characteristics, perhaps a short-lived pollutant with high chemical effects such as Ozone. In this case, the system does not follow a linear behaviour anymore; hence we may not simply rely on the advection of error variance for the propagation of errors. Instead, a more sophisticated framework to update and propagate the error covariances is required. One can perform an ensemble approach or compute the evolutions of higher moments of errors within the parametric approach, using a more general form of parametric Kalman filtering (see Section 4.4)
- The application of PvKF assimilation can also be investigated against other types and sets of observations with different properties. As mentioned earlier, the current PvKF assimilation is limited to the number and density of the observations that occurred for a particular time. In this case, one needs to expand the ability of the assimilation system to mitigate either computational or systematical limitations. For instance, a more sophisticated matrix inversion approach or a novel technique to treat a large number of observations (see Section 7.3.4) is required.
- Since PvKF assimilation is a newly developed system, there is significant room for improvement, either on the methodology or for the applications. For example, given the accurate estimate of the assimilation analysis, one can develop a bias correction

process over remote areas such as ocean observations. In addition, a more complicated form of the (model) error covariance, or with a larger number of parameters that better represent the error correlations, can also be examined with a modified assimilation system in future.

## Bibliography

- Alexe M, Bergamaschi P, Segers A, Detmers R, Butz A, Hasekamp O, Guerlet S, Parker R, Boesch H, Frankenberg C et al. 2015. Inverse modelling of ch<sub>4</sub> emissions for 2010-2011 using different satellite retrieval products from gosat and sciamachy. *Atmospheric Chemistry and Physics*. 15(1):113-133.
- Arellano AF, Kasibhatla PS, Giglio L, van der Werf GR, Randerson JT, Collatz GJ. 2006. Time-dependent inversion estimates of global biomass-burning co emissions using measurement of pollution in the troposphere (mopitt) measurements. *Journal of Geophysical Research-Atmospheres*. 111(D9):17.
- Arora VK, Melton JR, Plummer D. 2018. An assessment of natural methane fluxes simulated by the class-ctem model. *Biogeosciences*. 15(15):4683-4709.
- Asch M, Bocquet M, Nodet M. 2016. *Data assimilation: Methods, algorithms, and applications*. Philadelphia, USA: SIAM.
- Aydin M, Verhulst KR, Saltzman ES, Battle MO, Montzka SA, Blake DR, Tang Q, Prather MJ. 2011. Recent decreases in fossil-fuel emissions of ethane and methane derived from firm air. *Nature*. 476(7359):198-201.
- Babenhauserheide A, Basu S, Houweling S, Peters W, Butz A. 2015. Comparing the carbontracker and tm5-4dvar data assimilation systems for co<sub>2</sub> surface flux inversions. *Atmospheric Chemistry and Physics*. 15(17):9747-9763.
- Bader W, Bovy B, Conway S, Strong K, Smale D, Turner AJ, Blumenstock T, Boone C, Coen MC, Coulon A et al. 2017. The recent increase of atmospheric methane from 10 years of ground-based ndacc fir observations since 2005. *Atmospheric Chemistry and Physics*. 17(3):2255-2277.
- Banda N, Krol M, van Noije T, van Weele M, Williams JE, Le Sager P, Niemeier U, Thomason L, Rockmann T. 2015. The effect of stratospheric sulfur from mount pinatubo on tropospheric oxidizing capacity and methane. *Journal of Geophysical Research-Atmospheres*. 120(3):1202-1220.
- Baray S, Darlington A, Gordon M, Hayden KL, Leithead A, Li SM, Liu PSK, Mittermeier RL, Moussa SG, O'Brien J et al. 2018. Quantification of methane sources in the athabasca oil sands region of alberta by aircraft mass balance. *Atmospheric Chemistry and Physics*. 18(10):7361-7378.
- Basu S, Guerlet S, Butz A, Houweling S, Hasekamp O, Aben I, Krummel P, Steele P, Langenfelds R, Torn M et al. 2013. Global co<sub>2</sub> fluxes estimated from gosat retrievals of total column co<sub>2</sub>. *Atmospheric Chemistry and Physics*. 13(17):8695-8717.

- Basu S, Lan X, Dlugokencky E, Michel S, Schwietzke S, Miller JB, Bruhwiler L, Oh Y, Tans PP, Apadula F et al. 2022. Estimating emissions of methane consistent with atmospheric measurements of methane and  $\delta^{13}\text{C}$  of methane. *Atmos Chem Phys Discuss.* 2022:1-38.
- Berchet A, Pison I, Chevallier F, Bousquet P, Conil S, Geever M, Laurila T, Lavric J, Lopez M, Moncrieff J et al. 2013. Towards better error statistics for atmospheric inversions of methane surface fluxes. *Atmospheric Chemistry and Physics.* 13(14):7115-7132.
- Berchet A, Sollum E, Thompson RL, Pison I, Thanwerdas J, Broquet G, Chevallier F, Aalto T, Bergamaschi P, Brunner D et al. 2021. The community inversion framework v1.0: A unified system for atmospheric inversion studies. *Geoscientific Model Development.* 14(8):5331-5354.
- Bergamaschi P, Frankenberg C, Meirink JF, Krol M, Dentener F, Wagner T, Platt U, Kaplan JO, Korner S, Heimann M et al. 2007. Satellite cartography of atmospheric methane from sciamachyon board envisat: 2. Evaluation based on inverse model simulations. *Journal of Geophysical Research-Atmospheres.* 112(D2).
- Bergamaschi P, Houweling S, Segers A, Krol M, Frankenberg C, Scheepmaker R, Dlugokencky E, Wofsy S, Kort E, Sweeney C. 2013. Atmospheric  $\text{CH}_4$  in the first decade of the 21st century: Inverse modeling analysis using sciamachy satellite retrievals and noaa surface measurements. *Journal of Geophysical Research: Atmospheres.* 118(13):7350-7369.
- Bergamaschi P, Corazza M, Karstens U, Athanassiadou M, Thompson RL, Pison I, Manning AJ, Bousquet P, Segers A, Vermeulen A. 2015. Top-down estimates of european  $\text{CH}_4$  and  $\text{N}_2\text{O}$  emissions based on four different inverse models. *Atmospheric Chemistry and Physics.* 16:715-736.
- Bergamaschi P, Karstens U, Manning AJ, Saunio M, Tsuruta A, Berchet A, Vermeulen AT, Arnold T, Janssens-Maenhout G, Hammer S et al. 2018. Inverse modelling of european  $\text{CH}_4$  emissions during 2006-2012 using different inverse models and reassessed atmospheric observations. *Atmospheric Chemistry and Physics.* 18(2):901-920.
- Berry T, Sauer T. 2013. Adaptive ensemble kalman filtering of non-linear systems. *Tellus Series a-Dynamic Meteorology and Oceanography.* 65:16.
- Blake DR, Mayer EW, Tyler SC, Makide Y, Montague DC, Rowland FS. 1982. Global increase in atmospheric methane concentrations between 1978 and 1980. *Geophysical Research Letters.* 9(4):477-480.
- Bloom AA, Bowman KW, Lee M, Turner AJ, Schroeder R, Worden JR, Weidner R, McDonald KC, Jacob DJ. 2017. A global wetland methane emissions and

- uncertainty dataset for atmospheric chemical transport models (wetcharts version 1.0). *Geoscientific Model Development*. 10(6):2141-2156.
- Blumenstock T, Hase F, Schneider M, García O, Sepúlveda E. 2017. Tcon data from izana (es), release ggg2014. R1. TCCON data archive, hosted by CaltechDATA.
- Bocquet M, Elbern H, Eskes H, Hirtl M, Zabkar R, Carmichael GR, Flemming J, Inness A, Pagowski M, Camano JLP et al. 2015. Data assimilation in atmospheric chemistry models: Current status and future prospects for coupled chemistry meteorology models. *Atmospheric Chemistry and Physics*. 15(10):5325-5358.
- Bocquet M. 2016. Localization and the iterative ensemble kalman smoother. *Quarterly Journal of the Royal Meteorological Society*. 142(695):1075-1089.
- Bohn TJ, Melton JR, Ito A, Kleinen T, Spahni R, Stocker BD, Zhang B, Zhu X, Schroeder R, Glagolev MV et al. 2015. Wetchimp-wsl: Intercomparison of wetland methane emissions models over west siberia. *Biogeosciences*. 12(11):3321-3349.
- Bormann N, Bauer P. 2010. Estimates of spatial and interchannel observation-error characteristics for current sounder radiances for numerical weather prediction. I: Methods and application to atovs data. *Quarterly Journal of the Royal Meteorological Society*. 136(649):1036-1050.
- Bormann N, Collard A, Bauer P. 2010. Estimates of spatial and interchannel observation-error characteristics for current sounder radiances for numerical weather prediction. II: Application to airs and iasi data. *Quarterly Journal of the Royal Meteorological Society*. 136(649):1051-1063.
- Bousserez N, Henze DK, Perkins A, Bowman KW, Lee M, Liu J, Deng F, Jones DBA. 2015. Improved analysis-error covariance matrix for high-dimensional variational inversions: Application to source estimation using a 3d atmospheric transport model. *Quarterly Journal of the Royal Meteorological Society*. 141(690):1906-1921.
- Bousserez N, Henze DK, Rooney B, Perkins A, Wecht KJ, Turner AJ, Natraj V, Worden JR. 2016. Constraints on methane emissions in north america from future geostationary remote-sensing measurements. *Atmospheric Chemistry and Physics (Online)*. 16(10).
- Bousserez N, Henze DK. 2018. Optimal and scalable methods to approximate the solutions of large-scale bayesian problems: Theory and application to atmospheric inversion and data assimilation. *Quarterly Journal of the Royal Meteorological Society*. 144(711):365-390.
- Bovensmann H, Burrows J, Buchwitz M, Frerick J, Noël S, Rozanov V, Chance K, Goede A. 1999. Sciamachy: Mission objectives and measurement modes. *Journal of the atmospheric sciences*. 56(2):127-150.

- Bradley ES, Leifer I, Roberts DA, Dennison PE, Washburn L. 2011. Detection of marine methane emissions with aviris band ratios. *Geophysical Research Letters*. 38:4.
- Brasseur GP, Solomon S. 2005. *Dynamics and transport*. Springer.
- Brasseur GP, Jacob DJ. 2017. *Modeling of atmospheric chemistry*. New York, USA: Cambridge University Press.
- Bruhwiller L, Michalak A, Peters W, Baker D, Tans P. 2005. An improved kalman smoother for atmospheric inversions.
- Bruhwiller L, Dlugokencky E, Masarie K, Ishizawa M, Andrews A, Miller J, Sweeney C, Tans P, Worthy D. 2014. Carbontracker-ch4: An assimilation system for estimating emissions of atmospheric methane. *Atmospheric Chemistry and Physics*. 14(16):8269.
- Brunner D, Henne S, Keller C, Reimann S, Vollmer M, O'Doherty S, Maione M. 2012a. An extended kalman-filter for regional scale inverse emission estimation. *Atmospheric Chemistry and Physics*. 12(7):3455-3478.
- Brunner D, Henne S, Keller CA, Reimann S, Vollmer MK, O'Doherty S, Maione M. 2012b. An extended kalman-filter for regional scale inverse emission estimation. *Atmospheric Chemistry and Physics*. 12(7):3455-3478.
- Buchwitz M, Reuter M, Bovensmann H, Pillai D, Heymann J, Schneising O, Rozanov V, Krings T, Burrows JP, Boesch H et al. 2013. Carbon monitoring satellite (carbonsat): Assessment of atmospheric co2 and ch4 retrieval errors by error parameterization. *Atmospheric Measurement Techniques*. 6(12):3477-3500.
- Buchwitz M, Reuter M, Schneising O, Boesch H, Guerlet S, Dils B, Aben I, Armante R, Bergamaschi P, Blumenstock T et al. 2015. The greenhouse gas climate change initiative (ghg-cci): Comparison and quality assessment of near-surface-sensitive satellite-derived co2 and ch4 global data sets. *Remote Sensing of Environment*. 162:344-362.
- Buchwitz M, Reuter M, Schneising O, Hewson W, Detmers RG, Boesch H, Hasekamp OP, Aben I, Bovensmann H, Burrows JP et al. 2017. Global satellite observations of column-averaged carbon dioxide and methane: The ghg-cci xco2 and xch4 crdp3 data set. *Remote Sensing of Environment*. 203:276-295.
- Buehner M. 2005. Ensemble-derived stationary and flow-dependent background-error covariances: Evaluation in a quasi-operational nwp setting. *Quarterly Journal of the Royal Meteorological Society*. 131(607):1013-1043.
- Buehner M, Houtekamer PL, Charette C, Mitchell HL, He B. 2010. Intercomparison of variational data assimilation and the ensemble kalman filter for global deterministic nwp. Part i: Description and single-observation experiments. *Monthly Weather Review*. 138(5):1550-1566.

- The noaa annual greenhouse gas index (aggi). 2020. [accessed 5 October 2021]. <https://gml.noaa.gov/aggi/aggi.html>.
- Butler TM, Simmonds I, Rayner PJ. 2004. Mass balance inverse modelling of methane in the 1990s using a chemistry transport model. *Atmospheric Chemistry and Physics*. 4:2561-2580.
- Butz A, Guerlet S, Hasekamp O, Schepers D, Galli A, Aben I, Frankenberg C, Hartmann JM, Tran H, Kuze A et al. 2011. Toward accurate co2 and ch4 observations from gosat. *Geophysical Research Letters*. 38.
- Byrd RH, Lu PH, Nocedal J, Zhu CY. 1995. A limited memory algorithm for bound constrained optimization. *Siam Journal on Scientific Computing*. 16(5):1190-1208.
- Byun D, Schere KL. 2006. Review of the governing equations, computational algorithms, and other components of the models-3 community multiscale air quality (cmaq) modeling system. *Applied Mechanics Reviews*. 59(1-6):51-77.
- Chatterjee A, Michalak AM, Anderson JL, Mueller KL, Yadav V. 2012. Toward reliable ensemble kalman filter estimates of co2 fluxes. *Journal of Geophysical Research-Atmospheres*. 117:17.
- Chatterjee A, Michalak AM. 2013. Technical note: Comparison of ensemble kalman filter and variational approaches for co2 data assimilation. *Atmospheric Chemistry and Physics*. 13(23):11643-11660.
- Chen YH, Prinn RG. 2006. Estimation of atmospheric methane emissions between 1996 and 2001 using a three-dimensional global chemical transport model. *Journal of Geophysical Research-Atmospheres*. 111(D10):25.
- Chen YL, Shen HZ, Kaiser J, Hu YT, Capps SL, Zhao SL, Hakami A, Shih JS, Pavur GK, Turner MD et al. 2021. High-resolution hybrid inversion of iasi ammonia columns to constrain us ammonia emissions using the cmaq adjoint model. *Atmospheric Chemistry and Physics*. 21(3):2067-2082.
- Ciais P, Sabine C, Bala G, Bopp L, Brovkin V, Canadell J, Chhabra A, DeFries R, Galloway J, Heimann M. 2014. Carbon and other biogeochemical cycles. *Climate change 2013: The physical science basis contribution of working group i to the fifth assessment report of the intergovernmental panel on climate change*. Cambridge University Press. p. 465-570.
- CMAQ tutorials. 2021. [accessed 14 October 2021]. <https://www.epa.gov/cmaq/cmaq-documentation>.
- CMAQv5.3 user's guide. 2019. [accessed 14 October 2021]. [https://github.com/USEPA/CMAQ/blob/5.3/DOCS/Users\\_Guide/README.md](https://github.com/USEPA/CMAQ/blob/5.3/DOCS/Users_Guide/README.md).

- Cohn SE. 1993. Dynamics of short-term univariate forecast error covariances. *Monthly Weather Review*. 121(11):3123-3149.
- Cohn SE, Sivakumaran N, Todling R. 1994. A fixed-lag kalman smoother for retrospective data assimilation. *Monthly Weather Review*. 122(12):2838-2867.
- Cohn SE. 1997. An introduction to estimation theory. *Journal of the Meteorological Society of Japan*. 75(1B):257-288.
- Cohn SE, da Silva A, Guo J, Sienkiewicz M, Lamich D. 1998. Assessing the effects of data selection with the dao physical-space statistical analysis system. *Monthly Weather Review*. 126(11):2913-2926.
- Cosme E, Brankart J-M, Verron J, Brasseur P, Krysta M. 2010. Implementation of a reduced rank square-root smoother for high resolution ocean data assimilation. *Ocean Modelling*. 33(1-2):87-100.
- Cosme E, Verron J, Brasseur P, Blum J, Auroux D. 2012. Smoothing problems in a bayesian framework and their linear gaussian solutions. *Monthly Weather Review*. 140(2):683-695.
- Courtier P, Thepaut JN, Hollingsworth A. 1994. A strategy for operational implementation of 4d-var, using an incremental approach. *Quarterly Journal of the Royal Meteorological Society*. 120(519):1367-1387.
- Courtier P. 1997. Dual formulation of four-dimensional variational assimilation. *Quarterly Journal of the Royal Meteorological Society*. 123(544):2449-2461.
- Cressot C, Chevallier F, Bousquet P, Crevoisier C, Dlugokencky EJ, Fortems-Cheiney A, Frankenberg C, Parker R, Pison I, Scheepmaker RA et al. 2014. On the consistency between global and regional methane emissions inferred from sciamachy, tanso-fts, iasi and surface measurements. *Atmospheric Chemistry and Physics*. 14(2):577-592.
- Crippa M, Solazzo E, Huang GL, Guizzardi D, Koffi E, Muntean M, Schieberle C, Friedrich R, Janssens-Maenhout G. 2020. High resolution temporal profiles in the emissions database for global atmospheric research. *Scientific Data*. 7(1).
- Crippa M, Guizzardi D, Muntean M, Schaaf E, Lo Vullo E, Solazzo E, Monforti-Ferrario F, Olivier J, Vignati E. 2021. EDGAR v6.0 Greenhouse Gas Emissions. European Commission, Joint Research Centre (JRC).  
PID: <http://data.europa.eu/89h/97a67d67-c62e-4826-b873-9d972c4f670b>
- EDGAR v6.0 greenhouse gas emissions. 2021. European Commission, Joint Research Centre (JRC); [accessed 5 November 2021]. <http://data.europa.eu/89h/97a67d67-c62e-4826-b873-9d972c4f670b>.

- Dai T, Cheng Y, Goto D, Schutgens NA, Kikuchi M, Yoshida M, Shi G, Nakajima T. 2019. Inverting the east asian dust emission fluxes using the ensemble kalman smoother and himawari-8 aods: A case study with wrf-chem v3. 5.1. *Atmosphere*. 10(9):543.
- Daley R. 1992a. Estimating model-error covariances for application to atmospheric data assimilation. *Monthly Weather Review*. 120(8):1735-1746.
- Daley R. 1992b. Forecast-error statistics for homogeneous and inhomogeneous observation networks. *Monthly Weather Review*. 120(4):627-643.
- Daley R. 1992c. The effect of serially correlated observation and model error on atmospheric data assimilation. *Monthly Weather Review*. 120(1):164-177.
- Daley R. 1992d. The lagged innovation covariance - a performance diagnostic for atmospheric data assimilation. *Monthly Weather Review*. 120(1):178-196.
- Daley R, Menard R. 1993. Spectral characteristics of kalman filter systems for atmospheric data assimilation. *Monthly Weather Review*. 121(5):1554-1565.
- Dalsoren SB, Myhre CL, Myhre G, Gomez-Pelaez AJ, Sovde OA, Isaksen ISA, Weiss RF, Harth CM. 2016. Atmospheric methane evolution the last 40 years. *Atmospheric Chemistry and Physics*. 16(5):3099-3126.
- Dean JF, Middelburg JJ, Rockmann T, Aerts R, Blauw LG, Egger M, Jetten MSM, de Jong AEE, Meisel OH, Rasigraf O et al. 2018. Methane feedbacks to the global climate system in a warmer world. *Reviews of Geophysics*. 56(1):207-250.
- Dee DP. 1995. Online estimation of error covariance parameters for atmospheric data assimilation. *Monthly Weather Review*. 123(4):1128-1145.
- Dee DP, da Silva AM. 1999. Maximum-likelihood estimation of forecast and observation error covariance parameters. Part i: Methodology. *Monthly Weather Review*. 127(8):1822-1834.
- Dee DP, Gaspari G, Redder C, Rukhovets L, da Silva AM. 1999. Maximum-likelihood estimation of forecast and observation error covariance parameters. Part ii: Applications. *Monthly Weather Review*. 127(8):1835-1849.
- Deng F, Jones DBA, Henze DK, Bousserez N, Bowman KW, Fisher JB, Nassar R, O'Dell C, Wunch D, Wennberg PO et al. 2014. Inferring regional sources and sinks of atmospheric co2 from gosat xco2 data. *Atmospheric Chemistry and Physics*. 14(7):3703-3727.
- Desroziers G, Berre L, Chapnik B, Poli P. 2005. Diagnosis of observation, background and analysis-error statistics in observation space. *Quarterly Journal of the Royal Meteorological Society*. 131(613):3385-3396.

- Deutscher N, Notholt J, Messerschmidt J, Weinzierl C, Warneke T, Petri C, Grupe P, Katrynski K. 2019. Tccon data from bialystok (pl), release ggg2014. R1, tccon data archive, hosted by caltechdata.
- Dlugokencky EJ, Steele LP, Lang PM, Masarie KA. 1994. The growth-rate and distribution of atmospheric methane. *Journal of Geophysical Research-Atmospheres*. 99(D8):17021-17043.
- Dlugokencky EJ, Nisbet EG, Fisher R, Lowry D. 2011. Global atmospheric methane: Budget, changes and dangers. *Philosophical Transactions of the Royal Society a-Mathematical Physical and Engineering Sciences*. 369(1943):2058-2072.
- Dlugokencky. 2022. *Noaa/gml (www.Esrl.Noaa.Gov/gmd/ccgg/trends\_ch4/)*.
- Dutaur L, Verchot LV. 2007. A global inventory of the soil ch(4) sink. *Global Biogeochemical Cycles*. 21(4):9.
- Dyer ELE, Jones DBA, Nusbaumer J, Li H, Collins O, Vettoretti G, Noone D. 2017. Congo basin precipitation: Assessing seasonality, regional interactions, and sources of moisture. *Journal of Geophysical Research-Atmospheres*. 122(13):6882-6898.
- Elbern H, Strunk A, Schmidt H, Talagrand O. 2007. Emission rate and chemical state estimation by 4-dimensional variational inversion. *Atmospheric Chemistry and Physics*. 7(14):3749-3769.
- Errera Q, Daerden F, Chabrillat S, Lambert JC, Lahoz WA, Viscardy S, Bonjean S, Fonteyn D. 2008. 4d-var assimilation of mipas chemical observations: Ozone and nitrogen dioxide analyses. *Atmospheric Chemistry and Physics*. 8(20):6169-6187.
- Errera Q, Menard R. 2012. Technical note: Spectral representation of spatial correlations in variational assimilation with grid point models and application to the belgian assimilation system for chemical observations (bascoe). *Atmospheric Chemistry and Physics*. 12(21):10015-10031.
- Errera Q, Ceccherini S, Christophe Y, Chabrillat S, Hegglin MI, Lambert A, Menard R, Raspollini P, Skachko S, van Weele M et al. 2016. Harmonisation and diagnostics of mipas esa ch4 and n2o profiles using data assimilation. *Atmospheric Measurement Techniques*. 9(12):5895-5909.
- Eskes HJ, Van Velthoven PFJ, Valks PJM, Kelder HM. 2003. Assimilation of gome total-ozone satellite observations in a three-dimensional tracer-transport model. *Quarterly Journal of the Royal Meteorological Society*. 129(590):1663-1681.
- Etheridge DM, Steele LP, Francey RJ, Langenfelds RL. 1998. Atmospheric methane between 1000 ad and present: Evidence of anthropogenic emissions and climatic variability. *Journal of Geophysical Research-Atmospheres*. 103(D13):15979-15993.

- Etiopie G, Milkov AV. 2004. A new estimate of global methane flux from onshore and shallow submarine mud volcanoes to the atmosphere. *Environmental Geology*. 46(8):997-1002.
- Etiopie G, Feyzullayev A, Baciuc CL. 2009. Terrestrial methane seeps and mud volcanoes: A global perspective of gas origin. *Marine and Petroleum Geology*. 26(3):333-344.
- Etminan M, Myhre G, Highwood EJ, Shine KP. 2016. Radiative forcing of carbon dioxide, methane, and nitrous oxide: A significant revision of the methane radiative forcing. *Geophysical Research Letters*. 43(24):12614-12623.
- Evensen G. 1994. Sequential data assimilation with a nonlinear quasi-geostrophic model using monte-carlo methods to forecast error statistics. *Journal of Geophysical Research-Oceans*. 99(C5):10143-10162.
- Evensen G. 2009a. *Data assimilation: The ensemble kalman filter*. Springer Science & Business Media.
- Evensen G. 2009b. The ensemble kalman filter for combined state and parameter estimation monte carlo techniques for data assimilation in large systems. *Ieee Control Systems Magazine*. 29(3):83-104.
- Feng L, Palmer PI, Parker RJ, Lunt MF, Boesch H. 2022. Methane emissions responsible for record-breaking atmospheric methane growth rates in 2020 and 2021. *Atmos Chem Phys Discuss*. 2022:1-23.
- Ferretti DF, Miller JB, White JWC, Etheridge DM, Lassey KR, Lowe DC, Meure CMM, Dreier MF, Trudinger CM, van Ommen TD et al. 2005. Unexpected changes to the global methane budget over the past 2000 years. *Science*. 309(5741):1714-1717.
- Fillion A, Bocquet M, Gratton S, Gürol S, Sakov P. 2020. An iterative ensemble kalman smoother in presence of additive model error. *SIAM/ASA Journal on Uncertainty Quantification*. 8(1):198-228.
- Fiore AM, Jacob DJ, Field BD, Streets DG, Fernandes SD, Jang C. 2002. Linking ozone pollution and climate change: The case for controlling methane. *Geophysical Research Letters*. 29(19).
- Fiore AM,. Background error covariance modelling. Seminar on Recent Development in Data Assimilation for Atmosphere and Ocean; 2003: Shinfield Park, Reading.
- Fletcher SEM, Schaefer H. 2019. Rising methane: A new climate challenge. *Science*. 364(6444):932-933.

- Forster A, Schouten S, Baas M, Damste JSS. 2007a. Mid-cretaceous (albian-santonian) sea surface temperature record of the tropical atlantic ocean. *Geology*. 35(10):919-922.
- Forster P, Ramaswamy V, Artaxo P, Bernsten T, Betts R, Fahey DW, Haywood J, Lean J, Lowe DC, Myhre G. 2007b. Changes in atmospheric constituents and in radiative forcing. Chapter 2. *Climate change 2007 the physical science basis*.
- Fox TA, Barchyn TE, Risk D, Ravikumar AP, Hugenholtz CH. 2019. A review of close-range and screening technologies for mitigating fugitive methane emissions in upstream oil and gas. *Environmental Research Letters*. 14(5):18.
- France JL, Fisher RE, Lowry D, Allen G, Andrade MF, Bauguitte SJB, Bower K, Broderick TJ, Daly MC, Forster G et al. 2022. Delta c-13 methane source signatures from tropical wetland and rice field emissions. *Philosophical Transactions of the Royal Society a-Mathematical Physical and Engineering Sciences*. 380(2215):18.
- Frankenberg C, Thorpe AK, Thompson DR, Hulley G, Kort EA, Vance N, Borchardt J, Krings T, Gerilowski K, Sweeney C et al. 2016. Airborne methane remote measurements reveal heavy-tail flux distribution in four corners region. *Proceedings of the National Academy of Sciences of the United States of America*. 113(35):9734-9739.
- Fung I, John J, Lerner J, Matthews E, Prather M, Steele LP, Fraser PJ. 1991. 3-dimensional model synthesis of the global methane cycle. *Journal of Geophysical Research-Atmospheres*. 96(D7):13033-13065.
- Ganesan AL, Schwietzke S, Poulter B, Arnold T, Lan X, Rigby M, Vogel FR, van der Werf GR, Janssens-Maenhout G, Boesch H et al. 2019. Advancing scientific understanding of the global methane budget in support of the paris agreement. *Global Biogeochemical Cycles*. 33(12):1475-1512.
- Gaspari G, Cohn SE. 1999. Construction of correlation functions in two and three dimensions. *Quarterly Journal of the Royal Meteorological Society*. 125(554):723-757.
- Gauthier P, Charette C, Fillion L, Koclas P, Laroche S. 1999. Implementation of a 3d variational data assimilation system at the canadian meteorological centre. Part i: The global analysis. *Atmosphere-Ocean*. 37(2):103-156.
- Gauthier P, Tanguay M, Laroche S, Pellerin S, Morneau J. 2007. Extension of 3dvar to 4dvar: Implementation of 4dvar at the meteorological service of canada. *Monthly Weather Review*. 135(6):2339-2354.
- Gelb A. 1974. *Applied optimal estimation*. MIT press.

- Giglio L, Randerson JT, van der Werf GR. 2013. Analysis of daily, monthly, and annual burned area using the fourth-generation global fire emissions database (gfed4). *Journal of Geophysical Research-Biogeosciences*. 118(1):317-328.
- Gilliland AB, Dennis RL, Roselle SJ, Pierce TE. 2003. Seasonal nh3 emission estimates for the eastern united states based on ammonium wet concentrations and an inverse modeling method. *Journal of Geophysical Research-Atmospheres*. 108(D15):12.
- Gilliland AB, Appel KW, Pinder RW, Dennis RL. 2006. Seasonal nh3 emissions for the continental united states: Inverse model estimation and evaluation. *Atmospheric Environment*. 40(26):4986-4998.
- Gilpin S, Matsuo T, Cohn SE. 2022. Continuum covariance propagation for understanding variance loss in advective systems. *SIAM/ASA Journal on Uncertainty Quantification*. 886-914.
- Gómez-Sanabria A, Höglund-Isaksson L, Rafaj P, Schöpp W. 2018. Carbon in global waste and wastewater flows – its potential as energy source under alternative future waste management regimes. *Adv Geosci*. 45:105-113.
- Greybush SJ, Kalnay E, Miyoshi T, Ide K, Hunt BR. 2011. Balance and ensemble kalman filter localization techniques. *Monthly Weather Review*. 139(2):511-522.
- Hakami A, Henze DK, Seinfeld JH, Chai T, Tang Y, Carmichael GR, Sandu A. 2005. Adjoint inverse modeling of black carbon during the asian pacific regional aerosol characterization experiment. *Journal of Geophysical Research-Atmospheres*. 110(D14):17.
- Hakami A, Seinfeld JH, Chai TF, Tang YH, Carmichael GR, Sandu A. 2006. Adjoint sensitivity analysis of ozone nonattainment over the continental united states. *Environmental Science & Technology*. 40(12):3855-3864.
- Hakami A, Henze DK, Seinfeld JH, Singh K, Sandu A, Kim ST, Byun DW, Li QB. 2007. The adjoint of cmaq. *Environmental Science & Technology*. 41(22):7807-7817.
- Hamill TM, Whitaker JS, Snyder C. 2001. Distance-dependent filtering of background error covariance estimates in an ensemble kalman filter. *Monthly Weather Review*. 129(11):2776-2790.
- Hansen PC. 1999. The l-curve and its use in the numerical treatment of inverse problems.
- Heald CL, Jacob DJ, Jones DBA, Palmer PI, Logan JA, Streets DG, Sachse GW, Gille JC, Hoffman RN, Nehr Korn T. 2004. Comparative inverse analysis of satellite (mopitt) and aircraft (trace-p) observations to estimate asian sources of carbon monoxide. *Journal of Geophysical Research-Atmospheres*. 109(D23).

- Heimann M. 2011. Atmospheric science enigma of the recent methane budget. *Nature*. 476(7359):157-158.
- Henze DK, Hakami A, Seinfeld JH. 2007. Development of the adjoint of geos-chem. *Atmospheric Chemistry and Physics*. 7(9):2413-2433.
- Hintsala EJ, Moore FL, Hurst DF, Dutton GS, Hall BD, Nance JD, Miller BR, Montzka SA, Wolton LP, McClure-Begley A. 2021. Uas chromatograph for atmospheric trace species (ucats)—a versatile instrument for trace gas measurements on airborne platforms. *Atmospheric Measurement Techniques Discussions*.1-30.
- Hoesly RM, Smith SJ, Feng LY, Klimont Z, Janssens-Maenhout G, Pitkanen T, Seibert JJ, Vu L, Andres RJ, Bolt RM et al. 2018. Historical (1750-2014) anthropogenic emissions of reactive gases and aerosols from the community emissions data system (ceds). *Geoscientific Model Development*. 11(1):369-408.
- Hoglund-Isaksson L. 2012. Global anthropogenic methane emissions 2005-2030: Technical mitigation potentials and costs. *Atmospheric Chemistry and Physics*. 12(19):9079-9096.
- Hoglund-Isaksson L. 2017. Bottom-up simulations of methane and ethane emissions from global oil and gas systems 1980 to 2012. *Environmental Research Letters*. 12(2):10.
- Hollingsworth A, Lonnberg P. 1986. The statistical structure of short-range forecast errors as determined from radiosonde data .1. The wind-field. *Tellus Series a-Dynamic Meteorology and Oceanography*. 38(2):111-136.
- Houtekamer PL, Mitchell HL. 1998. Data assimilation using an ensemble kalman filter technique. *Monthly Weather Review*. 126(3):796-811.
- Houtekamer PL, Mitchell HL. 2001. A sequential ensemble kalman filter for atmospheric data assimilation. *Monthly Weather Review*. 129(1):123-137.
- Houtekamer PL, Mitchell HL. 2005. Ensemble kalman filtering. *Quarterly Journal of the Royal Meteorological Society*. 131(613):3269-3289.
- Houtekamer PL, Zhang FQ. 2016. Review of the ensemble kalman filter for atmospheric data assimilation. *Monthly Weather Review*. 144(12):4489-4532.
- Houweling S, Bergamaschi P, Chevallier F, Heimann M, Kaminski T, Krol M, Michalak AM, Patra PK. 2017. Global inverse modeling of ch4 sources and sinks: An overview of methods. *Atmospheric chemistry and physics*. 17(1):235-256.
- Hulley GC, Duren RM, Hopkins FM, Hook SJ, Vance N, Guillevic P, Johnson WR, Eng BT, Mihaly JM, Jovanovic VM et al. 2016. High spatial resolution imaging of methane and other trace gases with the airborne hyperspectral thermal emission spectrometer (hytes). *Atmospheric Measurement Techniques*. 9(5):2393-2408.

- Inoue M, Morino I, Uchino O, Nakatsuru T, Yoshida Y, Yokota T, Wunch D, Wennberg PO, Roehl CM, Griffith DWT et al. 2016. Bias corrections of gosat swir xco2 and xch4 with tcon data and their evaluation using aircraft measurement data. *Atmospheric Measurement Techniques*. 9(8):3491-3512.
- IPCC. 2006. IPCC guidelines for national greenhouse gas inventories The national greenhouse gas Inventories programme, edited by: Eggleston, H. S., Buendia, L., Miwa, K., Ngara, T., and Tanabe, K., The Intergovernmental Panel on Climate Change, IPCC TSU NGGIP, IGES, Institute for Global Environmental Strategies Hayama, Japan.
- IPCC. 2013. The physical science basis. *IPCC*, Cambridge Univ Press, New York.
- Jacob DJ. 1999. Introduction to atmospheric chemistry. Princeton University Press.
- Jacob DJ, Turner AJ, Maasackers JD, Sheng JX, Sun K, Liu X, Chance K, Aben I, McKeever J, Frankenberg C. 2016. Satellite observations of atmospheric methane and their value for quantifying methane emissions. *Atmospheric Chemistry and Physics*. 16(22):14371-14396.
- Jacob DJ, Varon DJ, Cusworth DH, Dennison PE, Frankenberg C, Gautam R, Guanter L, Kelley J, McKeever J, Ott LE et al. 2022. Quantifying methane emissions from the global scale down to point sources using satellite observations of atmospheric methane. *Atmos Chem Phys Discuss*. 2022:1-44.
- Jacobson MZCUP. 2020. Fundamentals of atmospheric modeling. Cambridge: Cambridge University Press.
- Janardanan R, Maksyutov S, Tsuruta A, Wang FJ, Tiwari YK, Valsala V, Ito A, Yoshida Y, Kaiser JW, Janssens-Maenhout G et al. 2020. Country-scale analysis of methane emissions with a high-resolution inverse model using gosat and surface observations. *Remote Sensing*. 12(3):24.
- Janjic T, Bormann N, Bocquet M, Carton JA, Cohn SE, Dance SL, Losa SN, Nichols NK, Potthast R, Waller JA et al. 2018. On the representation error in data assimilation. *Quarterly Journal of the Royal Meteorological Society*. 144(713):1257-1278.
- Janssens-Maenhout G, Crippa M, Guizzardi D, Muntean M, Schaaf E, Dentener F, Bergamaschi P, Pagliari V, Olivier JG, Peters JA. 2019. Edgar v4. 3.2 global atlas of the three major greenhouse gas emissions for the period 1970–2012. *Earth System Science Data*. 11(3):959-1002.
- Jazwinski AH. 1970. Stochastic processes and filtering theory. New York, NY: Acad. Press.

- Jervis D, McKeever J, Durak BOA, Sloan JJ, Gains D, Varon DJ, Ramier A, Strupler M, Tarrant E. 2021. The ghgsat-d imaging spectrometer. *Atmospheric Measurement Techniques*. 14(3):2127-2140.
- Jiang Z, Jones DBA, Worden J, Worden HM, Henze DK, Wang YX. 2015. Regional data assimilation of multi-spectral mopitt observations of co over north america. *Atmospheric Chemistry and Physics*. 15(12):6801-6814.
- Johnson MR, Tyner DR, Conley S, Schwietzke S, Zavala-Araiza D. 2017. Comparisons of airborne measurements and inventory estimates of methane emissions in the alberta upstream oil and gas sector. *Environmental Science & Technology*. 51(21):13008-13017.
- Kai FM, Tyler SC, Randerson JT, Blake DR. 2011. Reduced methane growth rate explained by decreased northern hemisphere microbial sources. *Nature*. 476(7359):194-197.
- Kaiser K, Benner R. 2012. Characterization of lignin by gas chromatography and mass spectrometry using a simplified cuo oxidation method. *Analytical Chemistry*. 84(1):459-464.
- Kalman RE. 1960. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*. 82(1):35-45.
- Kalman RE, Bucy RS. 1961. New results in linear filtering and prediction theory.
- Kaplan JO. 2002. Wetlands at the last glacial maximum: Distribution and methane emissions. *Geophysical Research Letters*. 29(6):4.
- Kasibhatla PS, Heimann M, Rayner PJ, Mahowald NM, Prinn RG, Hartley DE. 2000. Inverse methods in global biogeochemical cycles. Washington, D.C.: American Geophysical Union.
- Khattatov BV, Lamarque JF, Lyjak LV, Menard R, Levelt P, Tie XX, Brasseur GP, Gille JC. 2000. Assimilation of satellite observations of long-lived chemical species in global chemistry transport models. *Journal of Geophysical Research-Atmospheres*. 105(D23):29135-29144.
- Kirschke S, Bousquet P, Ciais P, Saunois M, Canadell JG, Dlugokencky EJ, Bergamaschi P, Bergmann D, Blake DR, Bruhwiler L et al. 2013. Three decades of global methane sources and sinks. *Nature Geoscience*. 6(10):813-823.
- Kivi R, Heikkinen P, Kyrö E. 2014. Tccon data from sodankylä (fi), release ggg2014. R0. TCCON data archive, hosted by CaltechDATA. 10.
- Kong L, Tang X, Zhu J, Wang ZF, Pan YP, Wu HJ, Wu L, Wu QZ, He YX, Tian SL et al. 2019. Improved inversion of monthly ammonia emissions in china based on

- the chinese ammonia monitoring network and ensemble kalman filter. *Environmental Science & Technology*. 53(21):12529-12538.
- Kopacz M, Jacob DJ, Fisher JA, Logan JA, Zhang L, Megretskaya IA, Yantosca RM, Singh K, Henze DK, Burrows JP et al. 2010. Global estimates of CO<sub>2</sub> sources with high resolution by adjoint inversion of multiple satellite datasets (MOPITT, AIRS, SCIAMACHY, TES). *Atmospheric Chemistry and Physics*. 10(3):855-876.
- Kort EA, Wofsy SC, Daube BC, Diao M, Elkins JW, Gao RS, Hintsa EJ, Hurst DF, Jimenez R, Moore FL et al. 2012. Atmospheric observations of arctic ocean methane emissions up to 82 degrees north. *Nature Geoscience*. 5(5):318-321.
- Krishnamoorthy A, Menon D. 2013. Matrix inversion using Cholesky decomposition. *2013 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*.70-72.
- Kurosawa K, Poterjoy J. 2021. Data assimilation challenges posed by nonlinear operators: A comparative study of ensemble and variational filters and smoothers. *Monthly Weather Review*. 149(7):2369-2389.
- Kuze A, Suto H, Nakajima M, Hamazaki T. 2009. Thermal and near infrared sensor for carbon observation Fourier-transform spectrometer on the greenhouse gases observing satellite for greenhouse gases monitoring. *Appl Opt*. 48(35):6716-6733.
- Kuze A, Kikuchi N, Kataoka F, Suto H, Shiomi K, Kondo Y. 2020. Detection of methane emission from a local source using GOSAT target observations. *Remote Sensing*. 12(2):15.
- Lahoz WA, Schneider P. 2014. Data assimilation: Making sense of earth observation. *Frontiers in Environmental Science*. 2(16).
- Lahoz WA,. *Advanced interdisciplinary data assimilation: Filtering and smoothing via error subspace statistical estimation. OCEANS'02 MTS/IEEE; 2002: IEEE.*
- Lambert G, Schmidt S. 1993. Reevaluation of the oceanic flux of methane - uncertainties and long-term variations. *Chemosphere*. 26(1-4):579-589.
- Lehner B, Doll P. 2004. Development and validation of a global database of lakes, reservoirs and wetlands. *Journal of Hydrology*. 296(1-4):1-22.
- Liang AL, Gong W, Han G, Xiang CZ. 2017. Comparison of satellite-observed XCO<sub>2</sub> from GOSAT, OCO-2, and ground-based TCCON. *Remote Sensing*. 9(10):27.
- Liang X, Zheng XG, Zhang SP, Wu GC, Dai YJ, Li Y. 2012. Maximum likelihood estimation of inflation factors on error covariance matrices for ensemble Kalman filter assimilation. *Quarterly Journal of the Royal Meteorological Society*. 138(662):263-273.

- Locatelli R, Bousquet P, Chevallier F, Fortems-Cheney A, Szopa S, Saunois M, Agustí-Panareda A, Bergmann D, Bian H, Cameron-Smith P et al. 2013. Impact of transport model errors on the global and regional methane emissions estimated by inverse modelling. *Atmospheric Chemistry and Physics*. 13(19):9917-9937.
- Locatelli R, Bousquet P, Saunois M, Chevallier F, Cressot C. 2015. Sensitivity of the recent methane budget to lmdz sub-grid-scale physical parameterizations. *Atmospheric Chemistry and Physics*. 15(17):9765-9780.
- Lonnberg P, Hollingsworth A. 1986. The statistical structure of short-range forecast errors as determined from radiosonde data .2. The covariance of height and wind errors. *Tellus Series a-Dynamic Meteorology and Oceanography*. 38(2):137-161.
- Lorenç AC. 2003. The potential of the ensemble kalman filter for nwp - a comparison with 4d-var. *Quarterly Journal of the Royal Meteorological Society*. 129(595):3183-3203.
- Lu X, Jacob DJ, Zhang YZ, Maasakkers JD, Sulprizio MP, Shen L, Qu Z, Scarpelli TR, Nesser H, Yantosca RM et al. 2021. Global methane budget and trend, 2010-2017: Complementarity of inverse analyses using in situ (globalviewplus ch4 obspack) and satellite (gosat) observations. *Atmospheric Chemistry and Physics*. 21(6):4637-4657.
- Lu X, Jacob DJ, Wang HL, Maasakkers JD, Zhang YZ, Scarpelli TR, Shen L, Qu Z, Sulprizio MP, Nesser H et al. 2022. Methane emissions in the united states, canada, and mexico: Evaluation of national methane emission inventories and 2010-2017 sectoral trends by inverse analysis of in situ (globalviewplus ch4 obspack) and satellite (gosat) atmospheric observations. *Atmospheric Chemistry and Physics*. 22(1):395-418.
- Ma S, Worden JR, Bloom AA, Zhang YZ, Poulter B, Cusworth DH, Yin Y, Pandey S, Maasakkers JD, Lu X et al. 2021. Satellite constraints on the latitudinal distribution and temperature sensitivity of wetland methane emissions. *Agu Advances*. 2(3):12.
- Maasakkers JD, Jacob DJ, Sulprizio MP, Turner AJ, Weitz M, Wirth T, Hight C, DeFigueiredo M, Desai M, Schmeltz R et al. 2016. Gridded national inventory of us methane emissions. *Environmental Science & Technology*. 50(23):13123-13133.
- Maasakkers JD, Jacob DJ, Sulprizio MP, Scarpelli TR, Nesser H, Sheng JX, Zhang YZ, Hersher M, Bloom AA, Bowman KW et al. 2019. Global distribution of methane emissions, emission trends, and oh concentrations and trends inferred from an inversion of gosat satellite data for 2010-2015. *Atmospheric Chemistry and Physics*. 19(11):7859-7881.
- Maasakkers JD, Jacob DJ, Sulprizio MP, Scarpelli TR, Nesser H, Sheng JX, Zhang YZ, Lu X, Bloom AA, Bowman KW et al. 2021. 2010-2015 north american methane

- emissions, sectoral contributions, and trends: A high-resolution inversion of gosat observations of atmospheric methane. *Atmospheric Chemistry and Physics*. 21(6):4339-4356.
- Marseille GJ, Barkmeijer J, de Haan S, Verkley W. 2016. Assessment and tuning of data assimilation systems using passive observations. *Quarterly Journal of the Royal Meteorological Society*. 142(701):3001-3014.
- Massart S, Agusti-Panareda A, Aben I, Butz A, Chevallier F, Crevoisier C, Engelen R, Frankenberg C, Hasekamp O. 2014. Assimilation of atmospheric methane products into the macc-ii system: From sciamachy to tanso and iasi. *Atmospheric Chemistry and Physics*. 14(12):6139-6158.
- Mathur R, Xing J, Gilliam R, Sarwar G, Hogrefe C, Pleim J, Pouliot G, Roselle S, Spero TL, Wong DC et al. 2017. Extending the community multiscale air quality (cmaq) modeling system to hemispheric scales: Overview of process considerations and initial applications. *Atmospheric Chemistry and Physics*. 17(20):12449-12474.
- McNorton J, Bousserez N, Agustí-Panareda A, Balsamo G, Engelen R, Huijnen V, Inness A, Kipling Z, Parrington M, Ribas R. 2022. Quantification of methane emissions from hotspots and during covid-19 using a global atmospheric inversion. *Atmos Chem Phys Discuss*. 2022:1-33.
- Menard R, Daley R. 1996. The application of kalman smoother theory to the estimation of 4dvar error statistics. *Tellus A*. 48(2):221-237.
- Menard R. 2000. Tracer assimilation. *GEOPHYSICAL MONOGRAPH-AMERICAN GEOPHYSICAL UNION*. 114:67-106.
- Menard R, Chang LP. 2000. Assimilation of stratospheric chemical tracer observations using a kalman filter. Part ii: Chi(2)-validated results and analysis of variance and correlation dynamics. *Monthly Weather Review*. 128(8):2672-2686.
- Menard R, Cohn SE, Chang LP, Lyster PM. 2000. Assimilation of stratospheric chemical tracer observations using a kalman filter. Part i: Formulation. *Monthly Weather Review*. 128(8):2654-2671.
- Menard R. Convergence and stability of estimated error variances derived from assimilation residuals in observation space. *Proceedings of ECMWF Workshop on diagnostics of data assimilation system performance*; 2009.
- Menard R, Deshaies-Jacques M, Gasset N. 2016. A comparison of correlation-length estimation methods for the objective analysis of surface pollutants at environment and climate change canada. *Journal of the Air & Waste Management Association*. 66(9):874-895.
- Menard R. 2016. Error covariance estimation methods based on analysis residuals: Theoretical foundation and convergence properties derived from simplified

- observation networks. *Quarterly Journal of the Royal Meteorological Society*. 142(694):257-273.
- Menard R, Deshaies-Jacques M. 2018a. Evaluation of analysis by cross-validation, part ii: Diagnostic and optimization of analysis error covariance. *Atmosphere*. 9(2):21.
- Menard R, Deshaies-Jacques M. 2018b. Evaluation of analysis by cross-validation. Part i: Using verification metrics. *Atmosphere*. 9(3):16.
- Menard R, Gauthier P, Rochon Y, Robichaud A, de Grandpre J, Yang Y, Charrette C, Chabrilat S. 2019. Coupled stratospheric chemistry-meteorology data assimilation. Part ii: Weak and strong coupling. *Atmosphere*. 10(12):45.
- Menard R, Skachko S, Pannekoucke O. 2021. Numerical discretization causing error variance loss and the need for inflation. *Quarterly Journal of the Royal Meteorological Society*. 23.
- Metia S, Oduro SD, Duc HN, Ha Q. 2016. Inverse air-pollutant emission and prediction using extended fractional kalman filtering. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*. 9(5):2051-2063.
- Meure CM, Etheridge D, Trudinger C, Steele P, Langenfelds R, van Ommen T, Smith A, Elkins J. 2006. Law dome co<sub>2</sub>, ch<sub>4</sub> and n<sub>2</sub>o ice core records extended to 2000 years bp. *Geophysical Research Letters*. 33(14):4.
- Migliorini S. 2013. Information-based data selection for ensemble data assimilation. *Quarterly Journal of the Royal Meteorological Society*. 139(677):2033-2054.
- Miller SM, Wofsy SC, Michalak AM, Kort EA, Andrews AE, Biraud SC, Dlugokencky EJ, Eluszkiewicz J, Fischer ML, Janssens-Maenhout G. 2013. Anthropogenic emissions of methane in the united states. *Proceedings of the National Academy of Sciences*. 110(50):20018-20022.
- Miller SM, Michalak AM. 2017. Constraining sector-specific co<sub>2</sub> and ch<sub>4</sub> emissions in the us. *Atmospheric Chemistry and Physics*. 17(6):3963-3985.
- Miller SM, Michalak AM, Detmers RG, Hasekamp OP, Bruhwiler LMP, Schwietzke S. 2019. China's coal mine methane regulations have not curbed growing emissions. *Nature Communications*. 10.
- Minx JC, Lamb WF, Andrew RM, Canadell JG, Crippa M, Dobbeling N, Forster PM, Guizzardi D, Olivier J, Peters GP et al. 2021. A comprehensive and synthetic dataset for global, regional, and national greenhouse gas emissions by sector 1970-2018 with an extension to 2019. *Earth System Science Data*. 13(11):5213-5252.

- Miyazaki K, Eskes HJ, Sudo K. 2012. Global nox emission estimates derived from an assimilation of omi tropospheric no2 columns. *Atmospheric Chemistry and Physics*. 12(5):2263-2288.
- Miyoshi T. 2011. The gaussian approach to adaptive covariance inflation and its implementation with the local ensemble transform kalman filter. *Monthly Weather Review*. 139(5):1519-1535.
- Monge-Sanz BM, Chipperfield MP, Untch A, Morcrette JJ, Rap A, Simmons AJ. 2013. On the uses of a new linear scheme for stratospheric methane in global models: Water source, transport tracer and radiative forcing. *Atmospheric Chemistry and Physics*. 13(18):9641-9660.
- Myhre G, Shindell D, Bréon F, Collins W, Fuglestedt J, Huang J, Koch D, Lamarque J, Lee D, Mendoza B et al. 2013. Anthropogenic and natural radiative forcing. *Climate change 2013: The physical science basis contribution of working group i to the fifth assessment report of the intergovernmental panel on climate change*. Cambridge, UK: Cambridge University Press. p. 659-740.
- Naftel JC. 2009. Nasa global hawk: A new tool for earth science research. National Aeronautics and Space Administration, Dryden Flight Research Center.
- Naik V, Voulgarakis A, Fiore AM, Horowitz LW, Lamarque JF, Lin M, Prather MJ, Young PJ, Bergmann D, Cameron-Smith PJ et al. 2013. Preindustrial to present-day changes in tropospheric hydroxyl radical and methane lifetime from the atmospheric chemistry and climate model intercomparison project (accmip). *Atmospheric Chemistry and Physics*. 13(10):5277-5298.
- Napelenok SL, Pinder RW, Gilliland AB, Martin RV. 2008. A method for evaluating spatially-resolved nox emissions using kalman filter inversion, direct sensitivities, and space-based no2 observations. *Atmospheric Chemistry and Physics*. 8(18):5603-5614.
- NASA's earth observing system data and information system (eosdis), data processing levels. 2021. [accessed 1 November 2021]. <https://earthdata.nasa.gov/collaborate/open-data-services-and-software/data-information-policy/data-levels>.
- National Academies of Sciences E, and Medicine. 2018. Improving characterization of anthropogenic methane emissions in the united states. Washington, DC: The National Academies Press.
- Nisbet EG, Fisher RE, Lowry D, France JL, Allen G, Bakkaloglu S, Broderick TJ, Cain M, Coleman M, Fernandez J et al. 2020. Methane mitigation: Methods to reduce emissions, on the path to the paris agreement. *Reviews of Geophysics*. 58(1):51.

- Nisbet EG, Jones AE, Pyle JA, Skiba U. 2022. Rising methane: Is there a methane emergency? Preface. *Philosophical Transactions of the Royal Society a-Mathematical Physical and Engineering Sciences*. 380(2215):4.
- Notholt J, Petri C, Warneke T, Deutscher N, Buschmann M, Weinzierl C, Macatangay R, Grupe P. 2019. Tccon data from bremen (de), release ggg2014r0, tccon data archive, hosted by caltechdata.
- Olsen E, Fetzer E, Hulley G, Manning E, Blaisdell J, Iredell L, Susskind J, Warner J, Wei Z, Blackwell W. 2013. *Airs/amsu/hsb version 6 level 2 product user guide*. USA: NASA-JPL Tech Rep.
- Orbe C, Waugh DW, Yang H, Lamarque JF, Tilmes S, Kinnison DE. 2017. Tropospheric transport differences between models using the same large-scale meteorological fields. *Geophysical Research Letters*. 44(2):1068-1078.
- Oshio H, Yoshida Y, Matsunaga T, Deutscher NM, Dubey M, Griffith DWT, Hase F, Iraci LT, Kivi R, Liu C et al. 2020. Bias correction of the ratio of total column ch4 to co2 retrieved from gosat spectra. *Remote Sensing*. 12(19).
- Osinski R, Bouttier F. 2018. Short-range probabilistic forecasting of convective risks for aviation based on a lagged-average-forecast ensemble approach. *Meteorological Applications*. 25(1):105-118.
- Otte TL, Pleim JE. 2010. The meteorology-chemistry interface processor (mcip) for the cmaq modeling system: Updates through mcipv3.4.1. *Geoscientific Model Development*. 3(1):243-256.
- Palmer PI, Feng L, Lunt MF, Parker RJ, Bosch H, Lan X, Lorente A, Borsdorff T. 2021. The added value of satellite observations of methane for understanding the contemporary methane budget. *Philosophical Transactions of the Royal Society a-Mathematical Physical and Engineering Sciences*. 379(2210):21.
- Pannekoucke O, Ricci S, Barthelemy S, Menard R, Thual O. 2016. Parametric kalman filter for chemical transport models. *Tellus Series a-Dynamic Meteorology and Oceanography*. 68:14.
- Pannekoucke O, Bocquet M, Menard R. 2018a. Parametric covariance dynamics for the nonlinear diffusive burgers equation. *Nonlinear Processes in Geophysics*. 25(3):481-495.
- Pannekoucke O, Ricci S, Barthelemy S, Menard R, Thual O. 2018b. Parametric kalman filter for chemical transport models (vol 68, 31547, 2016). *Tellus Series a-Dynamic Meteorology and Oceanography*. 70.
- Pannekoucke O, Menard R, El Aabaribaoune M, Plu M. 2021. A methodology to obtain model-error covariances due to the discretization scheme from the parametric kalman filter perspective. *Nonlinear Processes in Geophysics*. 28(1):1-22.

- Parker RJ, Boesch H, Byckling K, Webb AJ, Palmer PI, Feng L, Bergamaschi P, Chevallier F, Notholt J, Deutscher N et al. 2015. Assessing 5 years of gosat proxy xch4 data and associated uncertainties. *Atmospheric Measurement Techniques*. 8(11):4785-4801.
- Parker RJ, Webb A, Boesch H, Somkuti P, Guillo RB, Di Noia A, Kalaitzi N, Anand JS, Bergamaschi P, Chevallier F et al. 2020. A decade of gosat proxy satellite ch4 observations. *Earth System Science Data*. 12(4):3383-3412.
- Parrington M, Jones DBA, Bowman KW, Horowitz LW, Thompson AM, Tarasick DW, Witte JC. 2008. Estimating the summertime tropospheric ozone distribution over north america through assimilation of observations from the tropospheric emission spectrometer. *Journal of Geophysical Research-Atmospheres*. 113(D18):18.
- Parrish DF, Derber JC. 1992. The national-meteorological-centers spectral statistical-interpolation analysis system. *Monthly Weather Review*. 120(8):1747-1763.
- Patra PK, Houweling S, Krol M, Bousquet P, Belikov D, Bergmann D, Bian H, Cameron-Smith P, Chipperfield MP, Corbin K et al. 2011. Transcom model simulations of ch4 and related species: Linking transport, surface flux and chemical loss with ch4 variability in the troposphere and lower stratosphere. *Atmospheric Chemistry and Physics*. 11(24):12813-12837.
- Peng Z, Zhang M, Kou X, Tian X, Ma X. 2015. A regional carbon data assimilation system and its preliminary evaluation in east asia. *Atmospheric Chemistry and Physics*. 15(2):1087-1104.
- Peters W, Miller JB, Whitaker J, Denning AS, Hirsch A, Krol MC, Zupanski D, Bruhwiler L, Tans PP. 2005. An ensemble data assimilation system to estimate co2 surface fluxes from atmospheric trace gas observations. *Journal of Geophysical Research-Atmospheres*. 110(D24):21.
- Petit JR, Jouzel J, Raynaud D, Barkov NI, Barnola JM, Basile I, Bender M, Chappellaz J, Davis M, Delaygue G et al. 1999. Climate and atmospheric history of the past 420,000 years from the vostok ice core, antarctica. *Nature*. 399(6735):429-436.
- Polavarapu SM, Neish M, Tanguay M, Girard C, de Grandpre J, Semeniuk K, Gravel S, Ren SZ, Roche S, Chan D et al. 2016. Greenhouse gas simulations with a coupled meteorological and transport model: The predictability of co2. *Atmospheric Chemistry and Physics*. 16(18):12005-12038.
- Pickett-Heaps CA, Jacob DJ, Wecht KJ, Kort EA, Wofsy SC, Diskin GS, Worthy DEJ, Kaplan JO, Bey I, Drevet J. 2011. Magnitude and seasonality of wetland methane emissions from the hudson bay lowlands (canada). *Atmospheric Chemistry and Physics*. 11(8):3773-3779.

- Prather MJ, Holmes CD, Hsu J. 2012. Reactive greenhouse gas scenarios: Systematic exploration of uncertainties and the role of atmospheric chemistry. *Geophysical Research Letters*. 39:5.
- Prinn R, Weiss R. 2022. Advanced global atmospheric gases experiment (<https://agage.Mit.Edu/>).
- Prinn RG, Huang J, Weiss RF, Cunnold DM, Fraser PJ, Simmonds PG, McCulloch A, Harth C, Reimann S, Salameh P et al. 2005. Evidence for variability of atmospheric hydroxyl radicals over the past quarter century. *Geophysical Research Letters*. 32(7):4.
- Pulido M, Tandeo P, Bocquet M, Carrassi A, Lucini M. 2018. Stochastic parameterization identification using ensemble kalman filtering combined with maximum likelihood methods. *Tellus Series a-Dynamic Meteorology and Oceanography*. 70:17.
- Qu Z, Jacob DJ, Shen L, Lu X, Zhang YZ, Scarpelli TR, Nesser H, Sulprizio MP, Maasakkers JD, Bloom AA et al. 2021. Global distribution of methane emissions: A comparative inverse analysis of observations from the tropomi and gosat satellite instruments. *Atmospheric Chemistry and Physics*. 21(18):14159-14175.
- Rauch HE, Tung F, Striebel CT. 1965. Maximum likelihood estimates of linear dynamic systems. *AIAA journal*. 3(8):1445-1450.
- Raynaud L, Berre L, Desroziers G. 2009. Objective filtering of ensemble-based background-error variances. *Quarterly Journal of the Royal Meteorological Society*. 135(642):1177-1199.
- Reichstein M, Bahn M, Ciais P, Frank D, Mahecha MD, Seneviratne SI, Zscheischler J, Beer C, Buchmann N, Frank DC et al. 2013. Climate extremes and the carbon cycle. *Nature*. 500(7462):287-295.
- Rice AL, Butenhoff CL, Teama DG, Roger FH, Khalil MAK, Rasmussen RA. 2016. Atmospheric methane isotopic record favors fossil sources flat in 1980s and 1990s with recent increase. *Proceedings of the National Academy of Sciences of the United States of America*. 113(39):10791-10796.
- Richter JH, Solomon A, Bacmeister JT. 2014. Effects of vertical resolution and nonorographic gravity wave drag on the simulated climate in the community atmosphere model, version 5. *Journal of Advances in Modeling Earth Systems*. 6(2):357-383.
- Rigby M, Prinn RG, Fraser PJ, Simmonds PG, Langenfelds RL, Huang J, Cunnold DM, Steele LP, Krummel PB, Weiss RF et al. 2008. Renewed growth of atmospheric methane. *Geophysical Research Letters*. 35(22):6.

- Rigby M, Montzka SA, Prinn RG, White JWC, Young D, O'Doherty S, Lunt MF, Ganesan AL, Manning AJ, Simmonds PG et al. 2017. Role of atmospheric oxidation in recent methane growth. *Proceedings of the National Academy of Sciences of the United States of America*. 114(21):5373-5377.
- Rodgers CD. 2000. *Inverse methods for atmospheric sounding : Theory and practice*. Singapore ; [River Edge, N.J.] : World Scientific, [2000] [©2000].
- Rosevall JD, Murtagh DP, Urban J, Jones AK. 2007. A study of polar ozone depletion based on sequential assimilation of satellite data from the envisat/mipas and odin/smr instruments. *Atmospheric Chemistry and Physics*. 7:899-911.
- Rutherford ID. 1972. Data assimilation by statistical interpolation of forecast error fields. *Journal of the Atmospheric Sciences*. 29(5):809-+.
- Saad KM, Wunch D, Deutscher NM, Griffith DWT, Hase F, De Maziere M, Notholt J, Pollard DF, Roehl CM, Schneider M et al. 2016. Seasonal variability of stratospheric methane: Implications for constraining tropospheric methane budgets using total column observations. *Atmospheric Chemistry and Physics*. 16(21):14003-14024.
- Sandu A, Daescu DN, Carmichael GR, Chai TF. 2005. Adjoint sensitivity analysis of regional air quality models. *Journal of Computational Physics*. 204(1):222-252.
- Sass RL, Andrews JA, Ding AJ, Fisher FM. 2002. Spatial and temporal variability in methane emissions from rice paddies: Implications for assessing regional methane budgets. *Nutrient Cycling in Agroecosystems*. 64(1-2):3-7.
- Satterfield EA, Hodyss D, Kuhl DD, Bishop CH. 2018. Observation-informed generalized hybrid error covariance models. *Monthly Weather Review*. 146(11):3605-3622.
- Saunio M, Bousquet P, Poulter B, Peregon A, Ciais P, Canadell JG, Dlugokencky EJ, Etiope G, Bastviken D, Houweling S et al. 2016a. The global methane budget 2000-2012. *Earth System Science Data*. 8(2):697-751.
- Saunio M, Jackson RB, Bousquet P, Poulter B, Canadell JG. 2016b. The growing role of methane in anthropogenic climate change. *Environmental Research Letters*. 11(12).
- Saunio M, Bousquet P, Poulter B, Peregon A, Ciais P, Canadell JG, Dlugokencky EJ, Etiope G, Bastviken D, Houweling S. 2017. Variability and quasi-decadal changes in the methane budget over the period 2000–2012. *Atmospheric Chemistry and Physics*. 17(18):11135-11161.
- Saunio M, Stavert AR, Poulter B, Bousquet P, Canadell JG, Jackson RB, Raymond PA, Dlugokencky EJ, Houweling S, Patra PK et al. 2020. The global methane budget 2000-2017. *Earth System Science Data*. 12(3):1561-1623.

- Scarpelli TR, Jacob DJ, Maasackers JD, Sulprizio MP, Sheng JX, Rose K, Romeo L, Worden JR, Janssens-Maenhout G. 2020. A global gridded (0.1 degrees x 0.1 degrees) inventory of methane emissions from oil, gas, and coal exploitation based on national reports to the united nations framework convention on climate change. *Earth System Science Data*. 12(1):563-575.
- Scarpelli TR, Jacob DJ, Moran M, Reuland F, Gordon D. 2022. A gridded inventory of canada's anthropogenic methane emissions. *Environmental Research Letters*. 17(1):14.
- Scheepmaker RA, Frankenberg C, Deutscher NM, Schneider M, Barthlott S, Blumenstock T, Garcia OE, Hase F, Jones N, Mahieu E et al. 2015. Validation of sciamachy hdo/h2o measurements using the tcon and ndacc-musica networks. *Atmospheric Measurement Techniques*. 8(4):1799-1818.
- Schepers D, Guerlet S, Butz A, Landgraf J, Frankenberg C, Hasekamp O, Blavier JF, Deutscher NM, Griffith DWT, Hase F et al. 2012. Methane retrievals from greenhouse gases observing satellite (gosat) shortwave infrared measurements: Performance comparison of proxy and physics retrieval algorithms. *Journal of Geophysical Research-Atmospheres*. 117.
- Schuldt KN, Aalto T, Andrews A, Aoki S, Arduini J, Baier B, Bergamaschi P, Biermann T, Biraud SC, Boenisch H et al. 2021. Multi-laboratory compilation of atmospheric methane data for the period 1983-2020; obspack\_ch4\_1\_globalviewplus\_v3.0\_2021-05-07. NOAA Earth System Research Laboratory, Global Monitoring Laboratory.
- Seber GA, Lee AJ. 2012. *Linear regression analysis*. John Wiley & Sons.
- Segers AJ, Eskes HJ, Van der A RJ, Van Oss F, Van Velthoven PFJ. 2005. Assimilation of gome ozone profiles and a global chemistry-transport model using a kalman filter with anisotropic covariance. *Quarterly Journal of the Royal Meteorological Society*. 131(606):477-502.
- Seinfeld JH, Pandis SN. 2016. *Atmospheric chemistry and physics : From air pollution to climate change*. Hoboken, New Jersey: John Wiley & Sons.
- Sheng JX, Jacob DJ, Turner AJ, Maasackers JD, Benmergui J, Bloom AA, Arndt C, Gautam R, Zavala-Araiza D, Boesch H et al. 2018. 2010-2016 methane trends over canada, the united states, and mexico observed by the gosat satellite: Contributions from different source sectors. *Atmospheric Chemistry and Physics*. 18(16):12257-12267.
- Sherwen T, Schmidt JA, Evans MJ, Carpenter LJ, Grossmann K, Eastham SD, Jacob DJ, Dix B, Koenig TK, Sinreich R et al. 2016. Global impacts of tropospheric halogens (cl, br, i) on oxidants and composition in geos-chem. *Atmospheric Chemistry and Physics*. 16(18):12239-12271.

- Shindell D, Kuylenstierna JCI, Vignati E, van Dingenen R, Amann M, Klimont Z, Anenberg SC, Muller N, Janssens-Maenhout G, Raes F et al. 2012. Simultaneously mitigating near-term climate change and improving human health and food security. *Science*. 335(6065):183-189.
- Simon D. 2006. Optimal state estimation: Kalman, h infinity, and nonlinear approaches. John Wiley & Sons.
- Simpson IJ, Andersen MPS, Meinardi S, Bruhwiler L, Blake NJ, Helmig D, Rowland FS, Blake DR. 2012. Long-term decline of global atmospheric ethane concentrations and implications for methane. *Nature*. 488(7412):490-494.
- Skachko S, Errera Q, Menard R, Christophe Y, Chabrillat S. 2014. Comparison of the ensemble kalman filter and 4d-var assimilation methods using a stratospheric tracer transport model. *Geoscientific Model Development*. 7(4):1451-1465.
- Skachko S, Menard R, Errera Q, Christophe Y, Chabrillat S. 2016. Enkf and 4d-var data assimilation with chemical transport model bascoe (version 05.06). *Geoscientific Model Development*. 9(8):2893-2908.
- Skamarock WC, Klemp JB, Dudhia J, Gill DO, Barker DM, Wang W, Powers JG. 2008. A description of the advanced research wrf version 3. Near technical note-475+ str.
- Stanevich I, Jones D, Strong K, Parker RJ, Boesch H, Wunch D, Notholt J, Petri C, Warneke T, Sussmann R. 2020. Characterizing model errors in chemical transport modeling of methane: Impact of model resolution in versions v9-02 of geos-chem and v35j of its adjoint model. *Geoscientific Model Development*. 13(9):3839-3862.
- Stanevich I, Jones DBA, Strong K, Keller M, Henze DK, Parker RJ, Boesch H, Wunch D, Notholt J, Petri C et al. 2021. Characterizing model errors in chemical transport modeling of methane: Using gosat xch4 data with weak-constraint four-dimensional variational data assimilation. *Atmospheric Chemistry and Physics*. 21(12):9545-9572.
- Staniaszek Z, Griffiths PT, Folberth GA, O'Connor FM, Abraham NL, Archibald AT. 2022. The role of future anthropogenic methane emissions in air quality and climate. *Npj Climate and Atmospheric Science*. 5(1):8.
- Stoffelen A. 1998. Toward the true near-surface wind speed: Error modeling and calibration using triple collocation. *Journal of Geophysical Research-Oceans*. 103(C4):7755-7766.
- Strang G. 1968. On construction and comparison of difference schemes. *Siam Journal on Numerical Analysis*. 5(3):506-+.
- Strang G, Borre K. 1997. Linear algebra, geodesy, and gps. Siam.

- Szenasi B, Berchet A, Broquet G, Segers A, van der Gon HD, Krol M, Hullegie JJS, Kiesow A, Gunther D, Petrescu AMR et al. 2021. A pragmatic protocol for characterising errors in atmospheric inversions of methane emissions over europe. *Tellus Series B-Chemical and Physical Meteorology*. 73(1):1-23.
- Tandeo P, Ailliot P, Bocquet M, Carrassi A, Miyoshi T, Pulido M, Zhen YC. 2020. A review of innovation-based methods to jointly estimate model and observation error covariance matrices in ensemble data assimilation. *Monthly Weather Review*. 148(10):3973-3994.
- Tang W, Cohan D, Lamsal L, Xiao X, Zhou W. 2013. Inverse modeling of texas no x emissions using space-based and ground-based no 2 observations. *Atmospheric Chemistry and Physics*. 13(21):11005-11018.
- Tangborn A, Menard R, Ortland D. 2002. Bias correction and random error characterization for the assimilation of high-resolution doppler imager line-of-sight velocity measurements. *Journal of Geophysical Research-Atmospheres*. 107(D12):15.
- Tenkanen M, Tsuruta A, Rautiainen K, Kangasaho V, Ellul R, Aalto T. 2021. Utilizing earth observations of soil freeze/thaw data and atmospheric concentrations to estimate cold season methane emissions in the northern high latitudes. *Remote Sensing*. 13(24).
- Tian H, Li TF, Zhang TW, Xiao XM. 2016. Characterization of methane adsorption on overmature lower silurian-upper ordovician shales in sichuan basin, southwest china: Experimental results and geological implications. *International Journal of Coal Geology*. 156:36-49.
- Tremolet Y. 2006. Accounting for an imperfect model in 4d-var (vol 132, pg 2483, 2006). *Quarterly Journal of the Royal Meteorological Society*. 132(621):3127-3127.
- Tubiello FN, Karl K, Flammini A, Gutschow J, Obli-Laryea G, Conchedda G, Pan XY, Qi SY, Heidarsdottir HH, Wanner N et al. 2022. Pre- and post-production processes increasingly dominate greenhouse gas emissions from agri-food systems. *Earth System Science Data*. 14(4):1795-1809.
- Turner AJ, Jacob DJ. 2015. Balancing aggregation and smoothing errors in inverse models. *Atmospheric Chemistry and Physics*. 15(12):7039-7048.
- Turner AJ, Jacob DJ, Wecht KJ, Maasackers JD, Lundgren E, Andrews AE, Biraud SC, Boesch H, Bowman KW, Deutscher NM et al. 2015. Estimating global and north american methane emissions with high spatial resolution using gosat satellite data. *Atmospheric Chemistry and Physics*. 15(12):7049-7069.

- Turner AJ, Jacob D, Benmergui J, Wofsy S, Maasackers J, Butz A, Hasekamp O, Biraud S. 2016. A large increase in us methane emissions over the past decade inferred from satellite data and surface observations. *Geophysical research letters*. 43(5):2218-2224.
- Turner AJ, Frankenberg C, Wennberg PO, Jacob DJ. 2017. Ambiguity in the causes for decadal trends in atmospheric methane and hydroxyl. *Proceedings of the National Academy of Sciences of the United States of America*. 114(21):5367-5372.
- Turner AJ, Jacob DJ, Benmergui J, Brandman J, White L, Randles CA. 2018. Assessing the capability of different satellite observing configurations to resolve the distribution of methane emissions at kilometer scales. *Atmospheric Chemistry and Physics*. 18(11):8265-8278.
- Turner AJ, Frankenberg C, Kort EA. 2019. Interpreting contemporary trends in atmospheric methane. *Proceedings of the National Academy of Sciences of the United States of America*. 116(8):2805-2813.
- Turner MD, Henze DK, Hakami A, Zhao SL, Resler J, Carmichael GR, Stanier CO, Baek J, Sandu A, Russell AG et al. 2015b. Differences between magnitudes and health impacts of bc emissions across the united states using 12 km scale seasonal source apportionment. *Environmental Science & Technology*. 49(7):4362-4371.
- Ueno G, Higuchi T, Kagimoto T, Hirose N. 2010. Maximum likelihood estimation of error covariances in ensemble-based filters and its application to a coupled atmosphere-ocean model. *Quarterly Journal of the Royal Meteorological Society*. 136(650):1316-1343.
- Community modeling and analysis system cmas [www document]. Smoke v3.6 user's man. 2017. Chapel Hill: The University of North Carolina; [accessed 2 November 2021]. <https://www.cmascenter.org/smoke/>.
- UNFCCC. 2020. Ghg data – time series, annex i, available at: [https://di.unfccc.int/time\\_series](https://di.unfccc.int/time_series), last access: December 2020.
- US-EPA. 2016. Inventory of us greenhouse gas emissions and sinks: 1990-2014. EPA 430-R-11-005, Tech Rep.
- van der A RJ, Allaart MAF, Eskes HJ. 2010. Multi sensor reanalysis of total ozone. *Atmospheric Chemistry and Physics*. 10(22):11277-11294.
- van der A RJ, Allaart MAF, Eskes HJ. 2015. Extended and refined multi sensor reanalysis of total ozone for the period 1970-2012. *Atmospheric Measurement Techniques*. 8(7):3021-3035.
- van der Laan-Luijkx IT, van der Velde IR, van der Veen E, Tsuruta A, Stanislawski K, Babenhauserheide A, Zhang HF, Liu Y, He W, Chen HL et al. 2017. The

- carbontracker data assimilation shell (ctdas) v1.0: Implementation and global carbon balance 2001-2015. *Geoscientific Model Development*. 10(7):2785-2800.
- van der Werf GR, Randerson JT, Giglio L, van Leeuwen TT, Chen Y, Rogers BM, Mu MQ, van Marle MJE, Morton DC, Collatz GJ et al. 2017. Global fire emissions estimates during 1997-2016. *Earth System Science Data*. 9(2):697-720.
- Varon DJ, Jervis D, McKeever J, Spence I, Gains D, Jacob DJ. 2021. High-frequency monitoring of anomalous methane point sources with multispectral sentinel-2 satellite observations. *Atmospheric Measurement Techniques*. 14(4):2771-2785.
- Vaughn TL, Bell CS, Pickering CK, Schwietzke S, Heath GA, Petron G, Zimmerle DJ, Schnell RC, Nummedal D. 2018. Temporal variability largely explains top-down/bottom-up difference in methane emission estimates from a natural gas production region. *Proceedings of the National Academy of Sciences of the United States of America*. 115(46):11712-11717.
- Voshtani S, Menard R, Walker TW, Hakami A. 2022a. Assimilation of gosat methane in the hemispheric cmaq; part i: Design of the assimilation system. *Remote Sensing*. 14(2):32.
- Voshtani S, Menard R, Walker TW, Hakami A. 2022b. Assimilation of gosat methane in the hemispheric cmaq; part ii: Results using optimal error statistics. *Remote Sensing*. 14(2):27.
- Voshtani S, Menard R, Walker TW, Hakami A. 2022c (in review). Use of Assimilation Analysis in 4D-Var Source Inversion: Observing System Simulation Experiments (OSSEs) with GOSAT Methane and Hemispheric CMAQ. *Remote Sensing*.
- Waller JA, Ballard SP, Dance SL, Kelly G, Nichols NK, Simonin D. 2016a. Diagnosing horizontal and inter-channel observation error correlations for sevir observations using observation-minus-background and observation-minus-analysis statistics. *Remote Sensing*. 8(7):14.
- Waller JA, Dance SL, Nichols NK. 2016b. Theoretical insight into diagnosing observation error correlations using observation-minus-background and observation-minus-analysis statistics. *Quarterly Journal of the Royal Meteorological Society*. 142(694):418-431.
- Walter KM, Zimov SA, Chanton JP, Verbyla D, Chapin FS. 2006. Methane bubbling from siberian thaw lakes as a positive feedback to climate warming. *Nature*. 443(7107):71-75.
- Wang FJ, Maksyutov S, Tsuruta A, Janardanan R, Ito A, Sasakawa M, Machida T, Morino I, Yoshida Y, Kaiser JW et al. 2019. Methane emission estimates by the global high-resolution inverse model using national inventories. *Remote Sensing*. 11(21).

- Warneke T, Messerschmidt J, Notholt J, Weinzierl C, Deutscher N, Petri C, Grupe P, Vuillemin C, Truong F, Schmidt M. 2017. Tccon data from orléans (fr), release ggg2014. R0. TCCON data archive, hosted by CaltechDATA.
- Wecht KJ, Jacob DJ, Frankenberg C, Jiang Z, Blake DR. 2014a. Mapping of north american methane emissions with high spatial resolution by inversion of sciamachy satellite data. *Journal of Geophysical Research-Atmospheres*. 119(12):7741-7756.
- Wen ZW, Yin WT. 2013. A feasible method for optimization with orthogonality constraints. *Mathematical Programming*. 142(1-2):397-434.
- Wennberg P, Wunch D, Roehl C, Blavier J, Toon G, Allen N, Dowell P, Teske K, Martin C, Martin J. 2016. Tccon data from lamont (us), release ggg2014. R1. TCCON data archive, hosted by CaltechDATA. 10.
- Wennberg P, Roehl C, Wunch D, Toon G, Blavier J, Washenfelder R, Keppel-Aleks G, Allen N, Ayers J. 2017. Tccon data from park falls (us), release ggg2014. R1. TCCON data archive, hosted by CaltechDATA. 10.
- Whitaker JS, Hamill TM. 2002. Ensemble data assimilation without perturbed observations. *Monthly Weather Review*. 130(7):1913-1924.
- Wmo greenhouse gas bulletin (ghg bulletin), no.17: the state of greenhouse gases in the atmosphere based on global observations through 2020. 2021. No. 17. WMO; [accessed 2 November 2021]. [https://library.wmo.int/doc\\_num.php?explnum\\_id=10838](https://library.wmo.int/doc_num.php?explnum_id=10838).
- Wofsy SC, Team HS, Cooperating Modellers T, Satellite T. 2011. Hiaper pole-to-pole observations (hippo): Fine-grained, global-scale measurements of climatically important atmospheric gases and aerosols. *Philosophical Transactions of the Royal Society a-Mathematical Physical and Engineering Sciences*. 369(1943):2073-2086.
- Worden JR, Bloom AA, Pandey S, Jiang Z, Worden HM, Walker TW, Houweling S, Rockmann T. 2017. Reduced biomass burning emissions reconcile conflicting estimates of the post-2006 atmospheric methane budget. *Nature Communications*. 8:11.
- Worden JR, Cusworth DH, Qu Z, Yin Y, Zhang Y, Bloom AA, Ma S, Byrne BK, Scarpelli T, Maasackers JD et al. 2022. The 2019 methane budget and uncertainties at 1° resolution and each country through bayesian integration of gosat total column methane data and a priori inventory estimates. *Atmos Chem Phys*. 22(10):6811-6841.
- Wu XR, Elbern H, Jacob B. 2022. The assessment of potential observability for joint chemical states and emissions in atmospheric modelings. *Stochastic Environmental Research and Risk Assessment*. 36(6):1743-1760.

- Wunch D, Toon GC, Wennberg PO, Wofsy SC, Stephens BB, Fischer ML, Uchino O, Abshire JB, Bernath P, Biraud SC et al. 2010. Calibration of the total carbon column observing network using aircraft profile data. *Atmospheric Measurement Techniques*. 3(5):1351-1362.
- Wunch D, Toon GC, Blavier JFL, Washenfelder RA, Notholt J, Connor BJ, Griffith DWT, Sherlock V, Wennberg PO. 2011. The total carbon column observing network. *Philosophical Transactions of the Royal Society a-Mathematical Physical and Engineering Sciences*. 369(1943):2087-2112.
- Wunch D, Toon G, Sherlock V, Deutscher N, Liu C, Feist D, Wennberg P. 2017. Documentation for the 2014 tcon data release, caltechdata.
- Wunch D, Jones DBA, Toon GC, Deutscher NM, Hase F, Notholt J, Sussmann R, Warneke T, Kuenen J, van der Gon HD et al. 2019. Emissions of methane in Europe inferred by total column measurements. *Atmospheric Chemistry and Physics*. 19(6):3963-3980.
- Xiong XZ, Barnett C, Maddy E, Sweeney C, Liu XP, Zhou LH, Goldberg M. 2008. Characterization and validation of methane products from the atmospheric infrared sounder (airs). *Journal of Geophysical Research-Biogeosciences*. 113.
- Yoshida Y, Kikuchi N, Morino I, Uchino O, Oshchepkov S, Bril A, Saeki T, Schutgens N, Toon GC, Wunch D et al. 2013. Improvement of the retrieval algorithm for gosat swir xco2 and xch4 and their validation using tcon data. *Atmospheric Measurement Techniques*. 6(6):1533-1547.
- Yu K, Keller CA, Jacob DJ, Molod AM, Eastham SD, Long MS. 2018. Errors and improvements in the use of archived meteorological data for chemical transport modeling: An analysis using geos-chem v11-01 driven by geos-5 meteorology. *Geoscientific Model Development*. 11(1):305-319.
- Yu XY, Millet DB, Henze DK. 2021. How well can inverse analyses of high-resolution satellite data resolve heterogeneous methane fluxes? Observing system simulation experiments with the geos-chem adjoint model (v35). *Geoscientific Model Development*. 14(12):7775-7793.
- Zavala-Araiza D, Lyon DR, Alvarez RA, Davis KJ, Harriss R, Herndon SC, Karion A, Kort EA, Lamb BK, Lan X et al. 2015. Reconciling divergent estimates of oil and gas methane emissions. *Proceedings of the National Academy of Sciences of the United States of America*. 112(51):15597-15602.
- Zavala-Araiza D, Alvarez RA, Lyon DR, Allen DT, Marchese AJ, Zimmerle DJ, Hamburg SP. 2017. Super-emitters in natural gas infrastructure are caused by abnormal process conditions. *Nature Communications*. 8:10.

- Zhang X, Heemink A, Janssen L, Janssen P, Sauter F. 1999. A computationally efficient kalman smoother for the evaluation of the ch<sub>4</sub> budget in europe. *Applied Mathematical Modelling*. 23(2):109-129.
- Zhang Y, Jacob DJ, Maasakkers JD, Sulprizio MP, Sheng J-X, Gautam R, Worden J. 2018. Monitoring global tropospheric oh concentrations using satellite observations of atmospheric methane. *Atmospheric Chemistry & Physics*. 18(21).
- Zhang YZ, Jacob DJ, Lu X, Maasakkers JD, Scarpelli TR, Sheng JX, Shen L, Qu Z, Sulprizio MP, Chang JF et al. 2021. Attribution of the accelerating increase in atmospheric methane during 2010-2018 by inverse analysis of gosat observations. *Atmospheric Chemistry and Physics*. 21(5):3643-3666.
- Zhang Z, Poulter B, Knox S, Stavert A, McNicol G, Fluet-Chouinard E, Feinberg A, Zhao YH, Bousquet P, Canadell JG et al. 2022. Anthropogenic emission is the main contributor to the rise of atmospheric methane during 1993-2017. *National Science Review*. 9(5):13.
- Zhao SL, Russell MG, Hakami A, Capps SL, Turner MD, Henze DK, Percell PB, Resler J, Shen H, Russell AG et al. 2020a. A multiphase cmaq version 5.0 adjoint. *Geoscientific Model Development*. 13(7):2925-2944.
- Zhao Y, Saunois M, Bousquet P, Lin X, Berchet A, Hegglin MI, Canadell JG, Jackson RB, Hauglustaine DA, Szopa S. 2019. Nter-model comparison of global hydroxyl radical (oh) distributions and their impact on atmospheric methane over the 2000–2016 period. *Atmospheric Chemistry and Physics*. 19(21):13701-13723.
- Zhao YH, Saunois M, Bousquet P, Lin X, Berchet A, Hegglin MI, Canadell JG, Jackson RB, Deushi M, Jockel P et al. 2020a. On the role of trend and variability in the hydroxyl radical (oh) in the global methane budget. *Atmospheric Chemistry and Physics*. 20(21):13011-13022.
- Zhao YH, Saunois M, Bousquet P, Lin X, Berchet A, Hegglin MI, Canadell JG, Jackson RB, Dlugokencky EJ, Langenfelds RL et al. 2020b. Influences of hydroxyl radicals (oh) on top-down estimates of the global and regional methane budgets. *Atmospheric Chemistry and Physics*. 20(15):9525-9546.
- Zhou MQ, Dils B, Wang PC, Detmers R, Yoshida Y, O'Dell CW, Feist DG, Velazco VA, Schneider M, De Maziere M. 2016. Validation of tanso-fts/gosat xco<sub>2</sub> and xch<sub>4</sub> glint mode retrievals using tcon data from near-ocean sites. *Atmospheric Measurement Techniques*. 9(3):1415-1430.
- Zhu H, Zhang XW, Chu DL, Liao LZ. 2017. Nonconvex and nonsmooth optimization with generalized orthogonality constraints: An approximate augmented lagrangian method. *Journal of Scientific Computing*. 72(1):331-372.

- Zoogman P, Jacob DJ, Chance K, Liu X, Lin M, Fiore A, Travis K. 2014. Monitoring high-ozone events in the us intermountain west using tempo geostationary satellite observations. *Atmospheric Chemistry and Physics*. 14(12):6261-6271.
- Zubrow A, Chen L, Kotamarthi VR. 2008. Eakf-cmaq: Introduction and evaluation of a data assimilation for cmaq based on the ensemble adjustment kalman filter. *Journal of Geophysical Research-Atmospheres*. 113(D9):18.
- Zupanski D, Hou AY, Zhang SQ, Zupanski M, Kummerow CD, Cheung SH. 2007. Applications of information theory in ensemble data assimilation. *Quarterly Journal of the Royal Meteorological Society*. 133(627):1533-1545.

## Appendices

### Appendix A. Parameter Estimation Methods of Covariance (Chapter 3)

#### $\chi^2$ Diagnostic

In linear estimation theory (i.e., Kalman filtering) with a linear observation operator  $\mathbf{H}$ , the perceived (computed) innovation covariance matrix is obtained as  $\mathbf{\Gamma} = \mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R}$  (Kalman and Bucy 1961). On the other hand, a sample covariance of the innovations is denoted as  $\tilde{\mathbf{\Gamma}} = E[\mathbf{d}\mathbf{d}^T]$ . Accordingly, to estimate the true observation and background error covariances, it is necessary that the perceived innovation covariance matrix represents the sample covariance of the innovation ( $\mathbf{\Gamma} = \tilde{\mathbf{\Gamma}}$ ). One way to validate this is to perform the  $\chi^2$  diagnostic (Talagrand 1998; Menard and Chang 2000; Menard et al. 2000) expressed as

$$\langle \chi^2 \rangle = \text{tr}(\mathbf{\Gamma}^{-1} \tilde{\mathbf{\Gamma}}) = m, \quad (\text{A.1})$$

where  $\langle \rangle$  and  $\text{tr}(\ )$  represent the sample mean and trace of a matrix, respectively, and  $m$  denotes the number of observations. Note that the size of  $\mathbf{\Gamma}$  and  $\tilde{\mathbf{\Gamma}}$  are both  $m \times m$ .

This diagnostic is used to estimate one parameter (correlation length or error variance parameter) at a time, taking the others as known. Essentially, for a particular covariance model, the perceived innovation covariance matrix can be expressed as  $\mathbf{\Gamma}(\boldsymbol{\alpha}) = \mathbf{H}\mathbf{B}(\boldsymbol{\alpha})\mathbf{H}^T + \mathbf{R}(\boldsymbol{\alpha})$ , where its value depends on the tunable parameters  $\boldsymbol{\alpha}$ . Hence, an estimate of  $\boldsymbol{\alpha}$  is obtained once the equality of Equation (A.1) is maintained.

## Hollingsworth–Lönnberg

Another well-known method that was originally developed for optimal interpolation is known as Hollingsworth–Lönnberg (Hollingsworth and Lonnberg 1986; Rutherford 1972) method. This method is typically used to estimate the variance parameter ( $f^b$ ) and correlation length scales ( $L_h, L_v$ ) of the background error covariance. Considering a pair of observations at the model grid points  $(i, j)$ , this method performs minimization of a cost function based on a nonlinear fit of the distance between the perceived correlation model  $C(D_{ij}, L_c)$  and the normalized sample covariance

$$\bar{\Gamma}(i, j) = \tilde{\Gamma}(i, j) / \Sigma(i, j)$$

$$J(\boldsymbol{\alpha}) = J(f^b, L_c) = \sum_{i \neq j} \left( f^b C(D_{ij}, L_c) - \bar{\Gamma}(i, j) \right)^2, \quad (\text{A.2})$$

where  $\tilde{\Gamma}(i, j)$  denotes an element  $(i, j)$  of the innovation covariance matrix,  $\Sigma(i, j)$  is a covariance normalization, and  $J$  is Hollingsworth–Lönnberg cost function. Note that the optimization can be devised either using a local fit and local parameter estimation or a global fit. For instance, in the case of a global estimation,  $\Sigma(i, j) = 1$ , and the estimated  $f^b$  represents the global mean background variance parameter. The details and application of this method are provided in Hollingsworth and Lonnberg (1986); Lonnberg and Hollingsworth (1986); Bormann et al. (2010); Menard et al. (2016), and Menard and Deshaies-Jacques (2018a).

## Maximum Likelihood

The maximum likelihood method is another common approach that is based on probability estimation. The covariance parameters in this method are determined to

maximize the probability density function of innovation realizations. In general, it assumes a stochastic process with a normal distribution of innovations  $\mathbf{d} = \{d_1, d_2, \dots, d_m\}$  and covariance matrix  $\mathbf{\Gamma}(\boldsymbol{\alpha})$ . Thus, a conditional probability density function of innovations  $\mathbf{d}$  given parameters  $\boldsymbol{\alpha}$  is of the form

$$p(\mathbf{d} | \boldsymbol{\alpha}) = \frac{1}{\sqrt{2\pi \det(\mathbf{\Gamma}(\boldsymbol{\alpha}))}} \exp\left(-\frac{1}{2} \mathbf{d}^T \mathbf{\Gamma}^{-1}(\boldsymbol{\alpha}) \mathbf{d}\right), \quad (\text{A.3})$$

where  $\det(\bullet)$  denotes the determinant of a matrix. Equivalently, one can construct an analytical form of the log-likelihood function,  $\Psi_m(\boldsymbol{\alpha})$ , as  $p(\mathbf{d} | \boldsymbol{\alpha}) = \exp[-\Psi_m(\boldsymbol{\alpha})]$  which needs to be minimized. Hence, the maximum of the probability density function in Equation (A.3) is the same as the minimum of the maximum likelihood cost function in the form

$$\begin{aligned} \Psi_m(\boldsymbol{\alpha}) &= \left( \log(2\pi) + \frac{1}{2} \log(\det(\mathbf{\Gamma}(\boldsymbol{\alpha})) + \frac{1}{2} \mathbf{d}^T \mathbf{\Gamma}^{-1}(\boldsymbol{\alpha}) \mathbf{d}) \right) \\ &\propto \log(\det(\mathbf{\Gamma}(\boldsymbol{\alpha})) + \text{tr}(\mathbf{\Gamma}^{-1}(\boldsymbol{\alpha}) \tilde{\mathbf{\Gamma}})) \end{aligned} \quad (\text{A.4})$$

The details and derivation of this method are illustrated in Dee and da Silva (1999); and Dee et al. (1999).

### Desroziers Diagnostics

Another popular estimation method in the assimilation/inversion community was developed by Desroziers et al. (2005). This method examines different innovation statistics in the observation space. Besides the diagnostics on the innovation (i.e.,  $\mathbf{\Gamma} = \tilde{\mathbf{\Gamma}}$ ), it provides a consistency check on observation errors, background errors, and analysis errors as follows

$$\mathbf{\Gamma}(\boldsymbol{\alpha}) = \mathbf{H}\mathbf{B}(\boldsymbol{\alpha})\mathbf{H}^T + \mathbf{R}(\boldsymbol{\alpha}) = E\left[\mathbf{d}_b^o (\mathbf{d}_b^o)^T\right] = \tilde{\mathbf{\Gamma}} \quad (\text{A.5})$$

$$\mathbf{HB}(\boldsymbol{\alpha})\mathbf{H}^T = E\left[\mathbf{d}_b^a(\mathbf{d}_b^o)^T\right] \quad (\text{A.6})$$

$$\mathbf{R}(\boldsymbol{\alpha}) = E\left[\mathbf{d}_a^o(\mathbf{d}_b^o)^T\right] \quad (\text{A.7})$$

$$\mathbf{HA}(\boldsymbol{\alpha})\mathbf{H}^T = E\left[\mathbf{d}_b^a(\mathbf{d}_a^o)^T\right] \quad (\text{A.8})$$

where  $\mathbf{d}_b^o$  is observation – background,  $\mathbf{d}_a^o$  denotes observation – analysis, and  $\mathbf{d}_b^a$  represents analysis – background;  $\mathbf{A}$  also denotes the (perceived) analysis error covariance obtained from Kalman filtering equations. Due to mainly its simple implementations, the Desroziers diagnostic method has been widely used in atmospheric assimilation/inversion to determine whether the prescribed observation and background error are appropriate estimates or not. The accuracy, applicability, and convergence properties of this method were also discussed in-depth by Menard et al. (2016) and Waller et al. (2016a; 2016b).

### Cross-Validation

The diagnostics proposed by Desroziers et al. (2005) (also other diagnostics mentioned earlier) can provide the true error covariance only when the data assimilation system is optimal. However, the optimal solution (obtained from minimum estimation theory or Kalman filtering) might not be true unless the input error covariances therein are also determined realistic. Menard (2016) and Menard and Deshaies-Jacques (2018a) show that the necessary and sufficient conditions to estimate the true error covariances (i.e.,  $\mathbf{R}$  and  $\mathbf{HBH}^T$ ) are (i) innovation covariance consistency (i.e.,  $\boldsymbol{\Gamma} = \tilde{\boldsymbol{\Gamma}}$ ) and (ii) optimality of the Kalman gain matrix (i.e.,  $\mathbf{K} = \tilde{\mathbf{K}}$ ). Although the first condition can be demonstrated from previous estimation approaches, the second condition is challenging to perform, mainly due to the underlying correlation that exists between analysis and observations (Menard and Deshaies-Jacques 2018a; 2018b). In other words, the analysis may not be

optimal when it is evaluated through diagnostics with the same set of observations that produced it.

Using the cross-validation method, it was shown that the true Kalman gain does not necessarily rely on the minimum estimation theory, but occurs when the analysis error variance is computed on the passive observations (i.e., observations that are not used to generate the analysis) space. Accordingly, the passive observations are uncorrelated with analysis, and they can provide a realistic measure of the analysis errors. We recall that the analysis error covariance for any (arbitrary) Kalman gain is computed as

$$\mathbf{A} = (\mathbf{I} - \mathbf{KH})\mathbf{B}^{true}(\mathbf{I} - \mathbf{KH})^T + \mathbf{KR}^{true}\mathbf{K}^T, \quad (\text{A.9})$$

where  $\mathbf{B}^{true}$ ,  $\mathbf{R}^{true}$  are the true background error covariance and true observation error covariances, respectively. The optimal Kalman gain can be realistic in the same direction when the total analysis error variance ( $tr(\mathbf{A})$ ) is minimized with respect to the gain matrix under the cross-validation assumptions; thus, we have

$$\arg \min_{\alpha} \{tr(\mathbf{A}(\alpha))\} = \hat{\mathbf{K}} \approx \mathbf{B}^{true}\mathbf{H}^T(\mathbf{HB}^{true}\mathbf{H}^T + \mathbf{R})^{-1} = \mathbf{K}^{true}, \quad (\text{A.10})$$

where  $\mathbf{K}^{true}$  is true Kalman gain. Subsequently, by comparing the passive observations with the analysis interpolated at the passive observation locations, the cross-validation cost function is constructed in a form

$$J_c = \langle (O - A)_c^2 \rangle. \quad (\text{A.11})$$

Thus, the cost function,  $J_c$ , is indeed a measure of the analysis error variance (Marseille et al. 2016; Menard and Deshaies-Jacques 2018a). Assuming that the observation errors are spatially uncorrelated and uncorrelated with the forecast (or background) errors, it is established that

$$\mathbf{R}_c + \mathbf{H}_c \mathbf{A} \mathbf{H}_c^T = \mathbb{E}[(O - A)_c (O - A)_c^T], \quad (\text{A.12})$$

whether the analysis is optimal or not. In Equation (A.8), subscripts  $c$  denotes values estimated in the passive observation space that are independent observations.  $\mathbf{H}_c$  is the observation operator that interpolates the 3D field at the passive observation sites. By searching the values in the parameter space, we can estimate the value of the cost function for each parameter value until we find the minimum of the cost function. We thus have

$$\arg \min_{\boldsymbol{\alpha}} J_c(\boldsymbol{\alpha}) \Rightarrow \text{optimal } \mathbf{A}(\boldsymbol{\alpha}), \quad (\text{A.13})$$

in the subspace spanned by the covariance parameters,  $\boldsymbol{\alpha}$  (Menard and Deshaies-Jacques 2018a). The essential part of this search also consists in maintaining the innovation covariance consistency so that an estimate of the optimal parameters values that satisfies both conditions is obtained. The modelled covariances with these optimal parameter values are then an estimate of the true error covariances Corresponding to the optimal solution that represents a realistic atmosphere. Note that further details of the derivations of the cross-validation method for GOSAT observations are provided in Sections 6.2 and 6.3.

## Appendix B. Kalman Filtering Assimilation and its Variants (Chapter 4)

### Kalman Filter (KF)

Another category of data assimilation approach exists that neither needs an adjoint model nor the optimization based on variational analysis. It is derived from linear estimation theory and is widely known as Kalman filtering. Kalman filtering (KF) and its variants (e.g., extended Kalman filter, ensemble Kalman filter, Kalman smoother, and parametric Kalman filter) allow for an explicit estimation of the forecast and analysis error covariance along with the state estimate. This method generally does not have any assumption on the model error, so that the error can be automatically configured in the assimilation system. Kalman filtering can be formulated in a way that the state variable can evolve in time while observations are modelled sequentially, given the knowledge of the error statistic due to the model state and observations. Another distinctive feature of this approach is that the error covariance can evolve along with the state of the system. These characteristics make KF a favourable estimation method for model state estimation in data assimilation, rather than emissions inverse modelling, yet we will consider the application of these approaches for both assimilation and inversion problems.

A standard Kalman filter data assimilation assumes a linear model  $\mathbf{M}$ , which evolves the state of the system forward in time from  $t$  to  $t+1$ , thus

$$\mathbf{x}_{t+1} = \mathbf{M}_{t+1} \mathbf{x}_t + \boldsymbol{\varepsilon} \quad (\text{B.1})$$

where  $\varepsilon$  represent a stationary model error with zero mean and unbiased Gaussian distribution,  $\varepsilon = N(0, \mathbf{Q})$ . It accounts for the observations that are related to the model through the observation equation,

$$y = \mathbf{H}_t \mathbf{x}_t + \varepsilon^\circ \quad (\text{B.2})$$

where  $\mathbf{H}$ , and  $\varepsilon^\circ = N(0, \mathbf{R})$  are the observation operator and observation error with zero mean and Gaussian distribution. Kalman filter algorithm has two steps at each integration  $[t, t+1]$ . In the first step (i.e., the analysis step), the model state and observations are combined optimally at  $t$ , given their uncertainties estimate (Kalman 1960; Kalman and Bucy 1961). The state and error covariance analyses (posterior) are obtained as  $\mathbf{x}_t^a$  and  $\mathbf{P}_t^a$ , respectively. In the second step (i.e., the forecast step), the model applies to the current analysis to perform a forecast on the state and the error covariance. In summary, one timestep of KF includes

- Analysis step

$$\mathbf{x}_t^a = \mathbf{x}_t^f + \mathbf{K}_t (\mathbf{y}_t^\circ - \mathbf{K}_t \mathbf{x}_t^f) \quad (\text{B.3})$$

$$\mathbf{P}_t^a = \mathbf{P}_t^f - \mathbf{K}_t \mathbf{H}_t \mathbf{P}_t^f \quad (\text{B.4})$$

$$\mathbf{K} = \mathbf{P}_t^f \mathbf{H}_t^T (\mathbf{H}_t \mathbf{P}_t^f \mathbf{H}_t^T + \mathbf{R})^{-1} \quad (\text{B.5})$$

- Forecast step

$$\mathbf{x}_{t+1}^f = \mathbf{M}_{t+1} \mathbf{x}_t^a \quad (\text{B.6})$$

$$\mathbf{P}_{t+1}^f = \mathbf{P}_t^f - \mathbf{M}_{t+1} \mathbf{P}_t^a \mathbf{M}_{t+1}^T + \mathbf{Q}, \quad (\text{B.7})$$

where  $\mathbf{K}_t$  is called the Kalman gain matrix. The next integration  $[t+1, t+2]$  repeats these two steps, starting from the forecast of the state and error covariance to compute the analysis with a new set of observations at  $t+1$ , then performing another forecast. Note that

the integration time of assimilation/inversion should not be confused with the forecast model timestep, which is usually assumed smaller. Furthermore, in a KF inverse modelling problem, the observation operator  $\mathbf{H}_t$  includes the dynamic of the atmosphere while the forecast model  $\mathbf{M}_t$  represents the temporal behaviour of the parameter (e.g., emissions), which is often assumed stationary (i.e.,  $\mathbf{M}_t = \mathbf{I}$ ).

KF systems are based on some assumptions that limit their applicability to a large state-space atmospheric problem (data assimilation and inversion). It relies on a system that is linear by its nature and based on Gaussian error statistics. The real atmosphere, however, is confronted with the nonlinearity of some sort, such as those in the transport processes and chemistry. Another challenge with KF is that the error covariances need to be explicitly computed, stored, and propagated, which is almost intractable due to the large size of the state. Nevertheless, there are some forms of KF that attempt to address the limitations above. We briefly describe those main methods based on Kalman filtering in the following.

### **Extended Kalman Filter (EKF)**

The extended Kalman filter (EKF) is a variant of the Kalman filter method that allows for a small deviation from linearity and Gaussianity in a large state-space problem. It may also be applicable to a moderate nonlinear dynamic system under a certain condition, such as for the reduced state dimension of the system. In this approach, the dynamic system takes a more general form, such that

$$\mathbf{x}_{t+1} = \mathbf{M}_{t+1} \mathbf{x}_t + \boldsymbol{\varepsilon}, \quad (\text{B.8})$$

$$y = H_t \mathbf{x}_t + \varepsilon^\circ, \quad (\text{B.9})$$

where both the model  $M$  and the observation operator  $H$  represent a nonlinear system that will be linearized around the current mean state estimate and covariance. Notes that the linearized model  $\mathbf{M}'_t = \left. \frac{\partial M_t}{\partial \mathbf{x}} \right|_{\mathbf{x}_A}$  and observation operator  $\mathbf{H}'_t = \left. \frac{\partial H_t}{\partial \mathbf{x}} \right|_{\mathbf{x}_A}$  need to be performed for an appropriate timestep, for which the linearity (and Gaussianity) assumptions are held. Given a short enough timestep  $[t, t+1]$ , the steps of deriving the extended Kalman filter assimilation/inversion algorithm are then given as follows:

- Analysis step

$$\mathbf{x}_t^a = \mathbf{x}_t^f + \mathbf{K}_t (\mathbf{y}_t^\circ - \mathbf{H}'_t \mathbf{x}_t^f) \quad (\text{B.10})$$

$$\mathbf{P}_t^a = \mathbf{P}_t^f - \mathbf{K}_t (\mathbf{H}'_t) \mathbf{P}_t^f \quad (\text{B.11})$$

$$\mathbf{K} = \mathbf{P}_t^f (\mathbf{H}'_t)^T \left( (\mathbf{H}'_t) \mathbf{P}_t^f (\mathbf{H}'_t)^T + \mathbf{R} \right)^{-1} \quad (\text{B.12})$$

- Forecast step

$$\mathbf{x}_{t+1}^f = \mathbf{M}'_{t+1} \mathbf{x}_t^a \quad (\text{B.13})$$

$$\mathbf{P}_{t+1}^f = \mathbf{P}_t^f - (\mathbf{M}'_{t+1}) \mathbf{P}_t^a (\mathbf{M}'_{t+1})^T + \mathbf{Q} \quad (\text{B.14})$$

EKF has been used in both numerical weather prediction and atmospheric chemistry assimilation/inversion in the past. It was used in reduced dimensions problems where the size of the state and covariance are tractable to propagate error covariance or to construct a TLM (in reduced dimension). In this case,  $\mathbf{M}'$  and  $\mathbf{H}'$  are constructed explicitly. For example, Gilliland et al. (2006; 2003) used this approach for a low dimension problem with no and very small spatial variability to estimate  $\text{NH}_3$  emissions, with a focus on temporal variability (e.g., seasonality). Chen and Prinn (2006) used EKF for methane emissions in nine inversion cases to examine their temporal and sectoral

behaviour with uncertainties. Napelenok et al. (2008) used the capability of CMAQ-DDM (i.e., a form of TLM) for the estimation of NO<sub>x</sub> emissions using satellite observations. They account for only ten source regions in the U.S. while obtaining the error covariance evolutions. Tang et al. (2013) used the same capabilities for NO<sub>x</sub> emissions inversion using both satellite and ground-based observation. Metia et al. (2016) adopted a version of EKF that entails spectral representation for the evolution of error covariance, with the objective of emissions inversion of NO<sub>x</sub> and VOC in a high-resolution system. Brunner et al. (2012a) also performed an EKF in a reduced dimension of 224 source regions for the estimation of Halocarbon emissions, including HFC-125, HFC-152, and HCFC-141b and examined their result against bottom-up estimates.

Despite its simple implementation and efficient applicability for near-linear estimation systems, EKF is limited to be performed in a large state-space problem (e.g., native resolutions of atmospheric model state). It was shown that for a large assimilation/inversion system (e.g.,  $O(10^5)$ ), even if the TLM can be obtained, the solution may be sensitive to small errors (Evensen 2009b).

### **RTS Extended Kalman Smoother (RTS-EKS)**

The smoothing process consists of estimating the state, given the past, current, and future observations. In hindsight, smoother estimations must improve on filters that only rely on past and present observations. A derivation of optimal smoothing algorithms can be found in (Gelb 1974). Different types of smoothers, such as fixed-lagged, fixed-interval, and fixed-point smoother with or without an ensemble/iterative formulation, have been developed and applied in atmospheric data assimilation and inversion (Cohn et al. 1994; Ménard and Daley 1996; Zhang et al. 1999; Lermusiaux et al. 2002; Bruhwiler et al. 2005;

Cosme et al. 2010; Cosme et al. 2012; Bruhwiler et al. 2014; Peng et al. 2015; Dai et al. 2019; Fillion et al. 2020). A detailed comparison of characteristics and the computational cost between five types of smoothing algorithms in a geophysical Bayesian framework was demonstrated by Cosme et al. (2012), giving insights into adopting an appropriate smoothing algorithm for the problem of interest. Extended Kalman smoothing with Rauch-Tung-Striebel capabilities (RTS-EKS) (Rauch et al. 1965) is a type of forward-backward fixed-interval smoother, maintaining higher computational efficiency than other smoother, mainly because it resolves the backward filter without a need to explicitly integrate the backward estimate.

The RTS-EKS method, in general, has similar limitations as the EKF method, particularly in its applicability to a large state-space atmospheric problem. It requires the computation of TLM and only applies to a weak-nonlinear dynamical system, so that usually, a low-rank estimation or a reduced dimension form of this method is computationally affordable. Despite those similarities with an almost equivalent computational cost, RTS-EKS is expected to provide more consistent assimilation/inversion results than EKF throughout the entire assimilation window. A derivation of the RTS-EKS from a forward-backward smoother is presented in Simon (2006, p286-293). Below, the main steps of the smoothing algorithm are provided for one timestep  $[t, t+1]$ . Note that the computation of the forward filter in RTS-EKS is exactly equivalent to EKF, except that the analysis and its error covariance need to be stored during the forward pass and provided in backward calculations. In the algorithm below, subscript F and RTS correspond to the forward filter and backward computation of RTS smoother,

respectively. To Initialize the backward integration,  $\mathbf{x}_{RTS,t+1} = \mathbf{x}_{F,t+1}^a$  and  $\mathbf{P}_{RTS,t+1} = \mathbf{P}_{F,t+1}^a$  conditions are required, thus we have

#### Forward filter

- Analysis step

$$\mathbf{x}_{F,t}^a = \mathbf{x}_{F,t}^f + \mathbf{K}_{F,t}(\mathbf{y}_t^\circ - \mathbf{H}'_t \mathbf{x}_{F,t}^f) \quad (\text{B.15})$$

$$\mathbf{P}_{F,t}^a = \mathbf{P}_{F,t}^f - \mathbf{K}_{F,t}(\mathbf{H}'_t) \mathbf{P}_{F,t}^f \quad (\text{B.16})$$

$$\mathbf{K}_{F,t} = \mathbf{P}_{F,t}^f (\mathbf{H}'_t)^T \left( (\mathbf{H}'_t) \mathbf{P}_{F,t}^f (\mathbf{H}'_t)^T + \mathbf{R} \right)^{-1} \quad (\text{B.17})$$

- Forecast step

$$\mathbf{x}_{F,t+1}^f = \mathbf{M}'_{t+1} \mathbf{x}_{F,t}^a \quad (\text{B.18})$$

$$\mathbf{P}_{F,t+1}^f = \mathbf{P}_{F,t}^f - (\mathbf{M}'_{t+1}) \mathbf{P}_{F,t}^a (\mathbf{M}'_{t+1})^T + \mathbf{Q} \quad (\text{B.19})$$

#### (backward) RTS smoothing

$$\mathbf{K}_{RTS,t} = \mathbf{P}_{F,t}^a \mathbf{K}_{F,t} \left( \mathbf{P}_{F,t+1}^f \right)^{-1} \quad (\text{B.20})$$

$$\mathbf{P}_{RTS,t} = \mathbf{P}_{F,t}^a - \mathbf{K}_{RTS,t} \left( \mathbf{P}_{F,t+1}^f - \mathbf{P}_{RTS,t+1}^f \right) \mathbf{K}_{RTS,t}^T \quad (\text{B.21})$$

$$\mathbf{x}_{RTS,t} = \mathbf{x}_{F,t}^a + \mathbf{K}_{RTS,t} (\mathbf{x}_{RTS,t+1} - \mathbf{x}_{t+1}^f), \quad (\text{B.22})$$

where  $\mathbf{K}_{RTS,t}$  is a gain matrix of smoother, and  $\mathbf{x}_{RTS,t}$  and  $\mathbf{P}_{RTS,t}$  are smoothing solution of the state and error covariances. Intuitively, the smoothing solution at time  $t$  is superior to the filter as it accounts for the observations in future time (e.g., at  $t+1$ ). Note that the EKF and RTS-EKS smoother are computationally only applicable to the reduced dimension

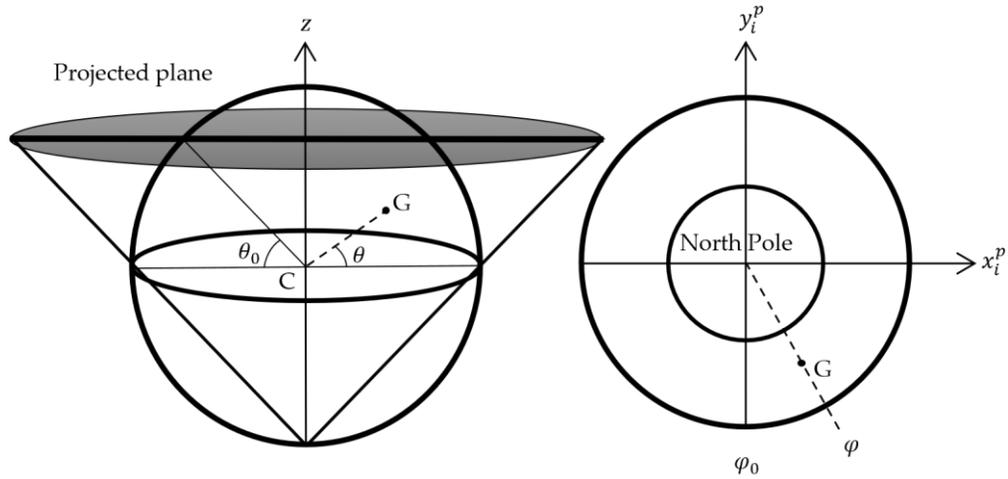
problems; thus, we do not rely on them for our high dimensional estimation problem. Still, we further discuss other data assimilation/inversion methods in the next sections to overcome the limitation of these two approaches.

## Appendix C1. Polar Stereographic Projection (Chapter 5)

H-CMAQ adapts the horizontal coordinates and the projections, including the polar stereographic projection, specified in MCIP (Otte and Pleim 2010). For the Northern Hemisphere, the projection starts from the South Pole and runs through each point on the Earth's surface to the projection plane. Considering the reference projection parameters specified in H-CMAQ (see Figure C.1, i.e., the true latitude  $\theta_0$  and longitude  $\varphi_0$ ), the transformation equations from the Earth's surface in S to the projection plane in  $\mathbb{R}^2(x_i^p, y_i^p)$ , where  $x_i^p$  and  $y_i^p$  denote the coordinates of the projected plane, are

$$\begin{aligned}x^p &= r \sin(\varphi - \varphi_0) \\y^p &= -r \cos(\varphi - \varphi_0) \\r &= \gamma r_e \cos \theta \quad , \\ \gamma &= \frac{1 + \sin(\theta_0)}{1 + \sin(\theta)}\end{aligned} \tag{C.1}$$

where  $\varphi$  and  $\theta$  are the corresponding longitude and the latitude of an arbitrary point (e.g., observation point G in Figure C.1) and  $r_e$  is the radius of the Earth. Note that the  $y$ -axis in the projection plane, as shown in Figure C.1, describes the meridian and  $\gamma$  denotes the polar stereographic map scale factor.



**Figure C.1. Polar stereographic projection geometry**

For the transformation into the horizontal gridded model, we use the latitude and longitude of the center of each gridcell. According to CMAQ documentation (CMAQv5.3 user's guide 2019), we account for the Earth's radius of 6,370,000 m, the true latitude of  $45^\circ$ , and the meridian taken as the true longitude. Note that the polar stereographic projection is integrated into the observation operator, so that each observation point is transformed using the same set of parameters and equations described above.

## Appendix C2. PvKF Data Assimilation Algorithm (Chapter 5)

Table C.1. Algorithm of Parametric variance Kalman filter (PvKF) assimilation

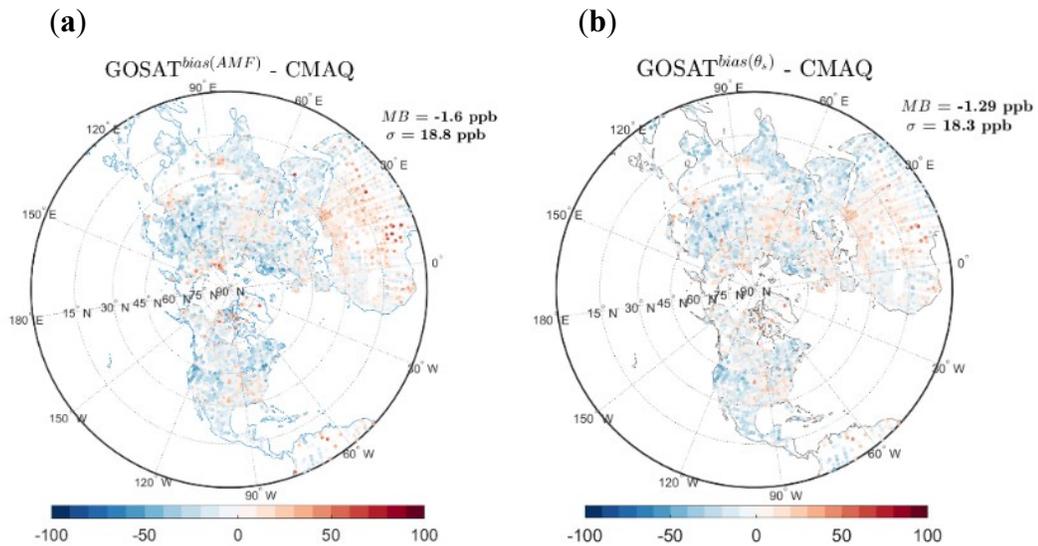
1:	Pre-processing of observations (quality control, bias correction, etc.)
2:	Initialization of state vector, $\mathbf{X}_0^f$ , and covariance matrix, $\mathbf{P}_0^f$
3:	<b>for</b> $k = 0, \dots, K$ <b>do</b>
4:	----- <b>Analysis step</b> -----
5:	<b>if</b> $\mathbf{Y}_k^o \neq \emptyset$ <b>then</b>
6:	$\mathbf{K}_k = (\mathbf{H}_k \mathbf{P}_k^f)^T (\mathbf{H}_k \mathbf{P}_k^f \mathbf{H}_k^T + \mathbf{R}_k)^{-1}$
7:	$\mathbf{X}_k^a = \mathbf{X}_k^f + \mathbf{K}_k (\mathbf{Y}_k^o - \mathbf{H}_k \mathbf{X}_k^f)$
8:	$\mathbf{P}_k^a = (\mathbf{I} - \mathbf{K}_k \mathbf{H}_k) \mathbf{P}_k^f = \mathbf{P}_k^f - (\mathbf{H}_k \mathbf{P}_k^f)^T (\mathbf{H}_k \mathbf{P}_k^f \mathbf{H}_k^T + \mathbf{R}_k)^{-1} (\mathbf{H}_k \mathbf{P}_k^f)$
9:	<b>else</b>
10:	$\mathbf{X}_k^a \leftarrow \mathbf{X}_k^f$
11:	$\mathbf{P}_k^a \leftarrow \mathbf{P}_k^f$
12:	<b>end if</b>
13:	----- <b>Forecast step</b> -----
14:	$\mathbf{X}_{k+1}^f = \mathbf{M}_{k:k+1} \mathbf{X}_k^a$
15:	$\mathbf{P}_{k+1}^f(i,i) = \mathbf{M}_{k:k+1}^* \mathbf{P}_k^a(i,i)$
16:	Note: $\mathbf{M}_{k:k+1}$ : H-CMAQ $\leftarrow$ CH <sub>4</sub>
17:	$\mathbf{M}_{k:k+1}^*$ : H-CMAQ (Advection-only) $\leftarrow$ Inert tracer
18:	-----
19:	<b>end for</b>

### Appendix C3. Regression Tests for GOSAT Bias (Chapter 5)

The air mass factor (AMF) is the ratio of the slant column measurements to the vertical column. It can be approximated with a simple function of the solar zenith angle,  $\theta_s$ , and the satellite viewing angle,  $\theta_v$ , as:

$$\text{AMF} = \frac{1}{\cos \theta_s} + \frac{1}{\cos \theta_v} \quad (\text{C.2})$$

Figure C.2 displays the difference between GOSAT and CMAQ with bias corrections on the GOSAT observations to account for AMF and for solar zenith angle,  $\theta_s$ , for the month of April 2010.



**Figure C.2. (a)  $\text{GOSAT}^{\text{bias(AMF)}} - \text{CMAQ}$  with air mass factor bias correction, (b)  $\text{GOSAT}^{\text{bias}(\theta_s)} - \text{CMAQ}$  with solar zenith angle bias correction. MB and  $\sigma$  denote domain-wide mean bias and standard deviation of the residual.**

The global mean bias is  $-1.6$  ppb ( $\text{GOSAT}^{\text{bias(AMF)}} - \text{CMAQ}$ ) and  $-1.29$  ppb ( $\text{GOSAT}^{\text{bias}(\theta_s)} - \text{CMAQ}$ ) for bias-corrected fields with respect to AMF and  $\theta_s$ , respectively. The residual standard deviation remains in the same range ( $\sim 18$  ppb) for both cases. Both cases provide a larger mean bias and larger residual standard deviation than the bias-corrected data with respect to the latitude.

***Multivariate (iterative) regression:***

The regression lines are computed based on Equation (C.4). A multivariate iterative regression is used to fit the residuals to the latitude, solar zenith angle, and AMF. The algorithm starts by regressing the residual against the first parameter, latitude (*itr 1*); next, the residual from the first iteration ( $C_{Y_i}^{\text{itr}1} - C_{X_i}^{\text{itr}1}$ ) is regressed against the second parameter,  $\theta_s$ , (*itr 2*). Similarly, in the third iteration, the residual of the second iteration ( $C_{Y_i}^{\text{itr}2} - C_{X_i}^{\text{itr}2}$ ) is regressed on the third variable, AMF.

**Table C.2. Evaluating the multiple (iterative) regression algorithm based on two measures: Mean Square Error (*MSE*) and the correlation coefficient *r*. Latitude,  $\theta_s$ , and AMF represent the order of parameters in the multiple regression algorithm, respectively.**

Parameter Measure	Latitude ( <i>itr 1</i> )	$\theta_s$ ( <i>itr 2</i> )	AMF ( <i>itr 3</i> )
<i>MSE (ppb)</i>	334.6	326.5	325.7
<i>r</i>	0.43	0.07	0.005

The significant decline of the correlation coefficient, *r*, in Table C.2 denotes that the three parameters are dependent and highly correlated. In addition, assuming that the

small reduction of *MSE* either from *itr 1* to *itr 2* and from *itr 2* to *itr 3* does not have a substantial impact on every single data point; therefore, we could account for the single parameter regression on the highest correlated parameter to be as effective as multiple regression analysis.

$$MSE = \frac{1}{n} \sum_{i=1}^n (C_{Y_i} - C_{X_i})^2 \quad (C.3)$$

$$r = \frac{\sum_{i=1}^n (C_{X_i} - \bar{C}_{X_i})(C_{Y_i} - \bar{C}_{Y_i})}{\sqrt{\sum_{i=1}^n (C_{X_i} - \bar{C}_{X_i})^2 \sum_{i=1}^n (C_{Y_i} - \bar{C}_{Y_i})^2}} \quad (C.4)$$

$$C_{X_i} = \alpha X_i + \beta \quad (C.5)$$

$C_{Y_i}$  denotes the data point and  $C_{X_i}$  represents the equivalent value from regression model prediction. A different order of parameters in multivariate regression has been tested in Table C.3 and Table C.4

**Table C.3. Evaluating the multiple (iterative) regression algorithm based on two measures: Mean Square Error (*MSE*) and the correlation coefficient *r*.  $\theta_s$ , latitude, and AMF represent the order of parameters in the multiple regression algorithm, respectively.**

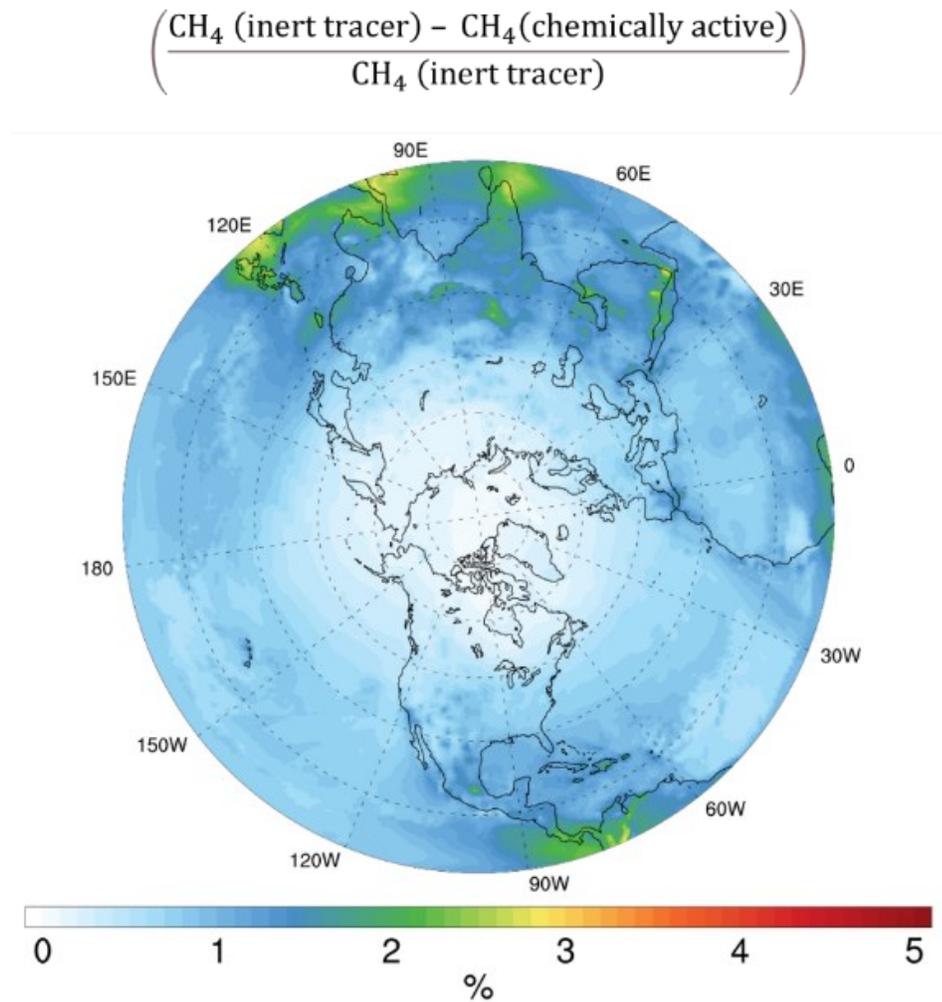
Measure \ Parameter	$\theta_s$ ( <i>itr 1</i> )	Latitude ( <i>itr 2</i> )	AMF ( <i>itr 3</i> )
<i>MSE (ppb)</i>	387.2	329.3	326.4
<i>r</i>	0.40	0.09	0.01

**Table C.4. Evaluating the multiple (iterative) regression algorithm based on two measures: Mean Square Error (*MSE*) and the correlation coefficient *r*. AMF,  $\theta_s$ , and latitude represent the order of parameters in the multiple regression algorithm, respectively.**

Measure \ Parameter	AMF ( <i>itr 1</i> )	$\theta_s$ ( <i>itr 2</i> )	Latitude ( <i>itr 3</i> )
<i>MSE (ppb)</i>	452.8	341.1	331.6
<i>r</i>	0.33	0.12	0.02

Comparing the second and third iterations between Table C.2-Table C.4 can imply that the highest correlated variable (i.e., latitude) is the best choice for the linear regression (with a single parameter) to replace with a multiple (iterative) regression since it leaves a smaller *MSE* and less residual correlation. However, we could also infer that if we start from the less correlated variable, we probably require more iterations of a multivariate regression (than starting with the highest correlated variable) to obtain an efficient regression. In this case, replacing the multivariate regression with the single-variable regression becomes less applicable. Nonetheless, multivariate regression using all parameters ends up with almost the same results regardless of the order of variables in the iterations.

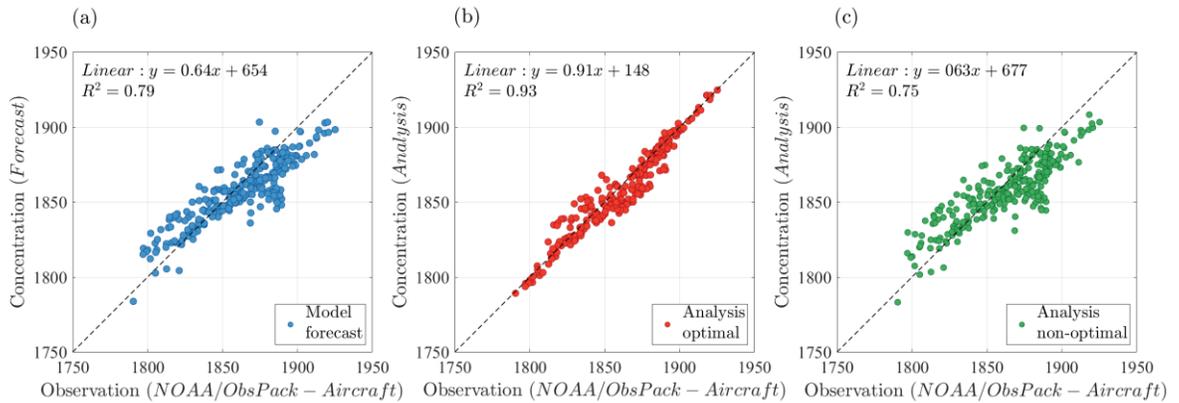
## Appendix C4. Methane Chemical Reaction Effect (Chapter 5)



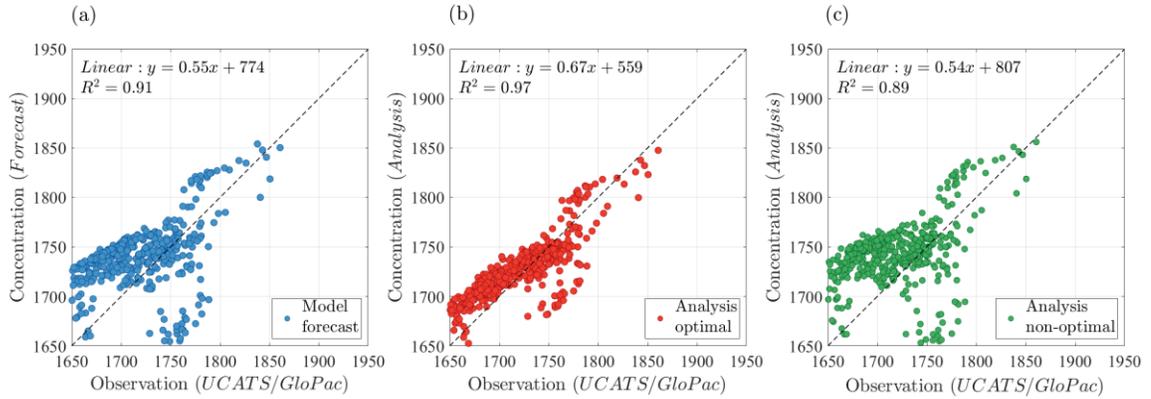
**Figure C.3. Hemispheric spatial distribution of the relative methane concentration loss due to chemical reactions. It shows daily average values after two weeks of the model forecast.**

## Appendix D. Evaluation Against Independent Observations (Chapter 6)

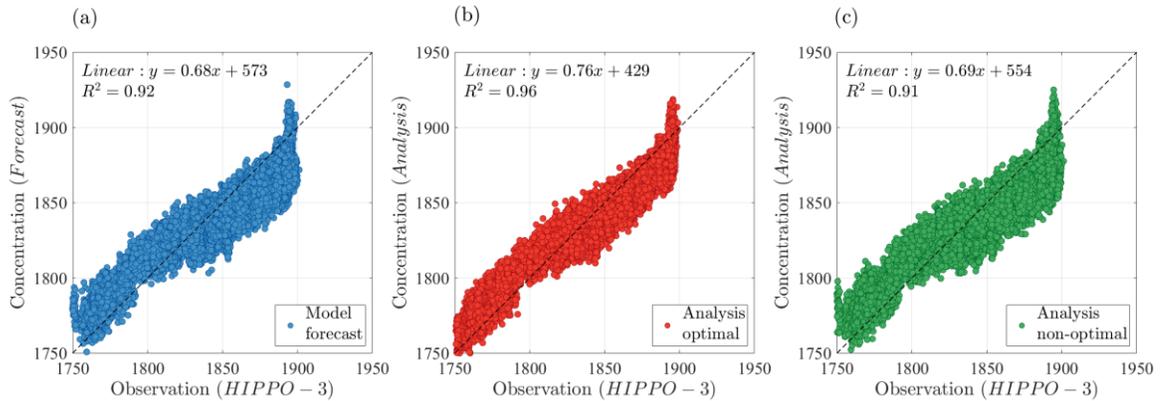
In Figure D.1-Figure D.3, optimal error covariance parameters of the analysis include  $f^o = 0.5$ ,  $f^i = 0.45$ ,  $f^q = 0.018$ ,  $L_h = 350$  km,  $L_v = 7\sigma$  while the non-optimal parameters are assumed to be  $f^o = 1.2$ ,  $f^i = 0.45$ ,  $f^q = 0$ ,  $L_h = 600$  km,  $L_v = 1\sigma$ . All comparisons are shown for April 2010.



**Figure D.1. Comparison of (a) model forecast (blue circles), (b) analysis with optimized parameters (red circles), and (c) analysis with non-optimized parameters (green circles) against independent NOAA/ObsPack aircraft observations.**

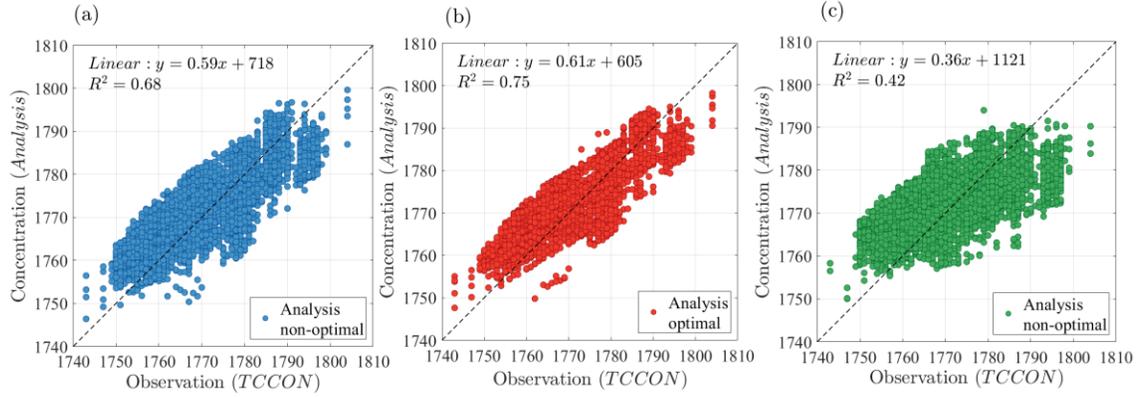


**Figure D.2.** Comparison of (a) model forecast (blue circles), (b) analysis with optimized parameters (red circles), and (c) analysis with non-optimized parameters (green circles) against independent UCATS/GloPac observations.

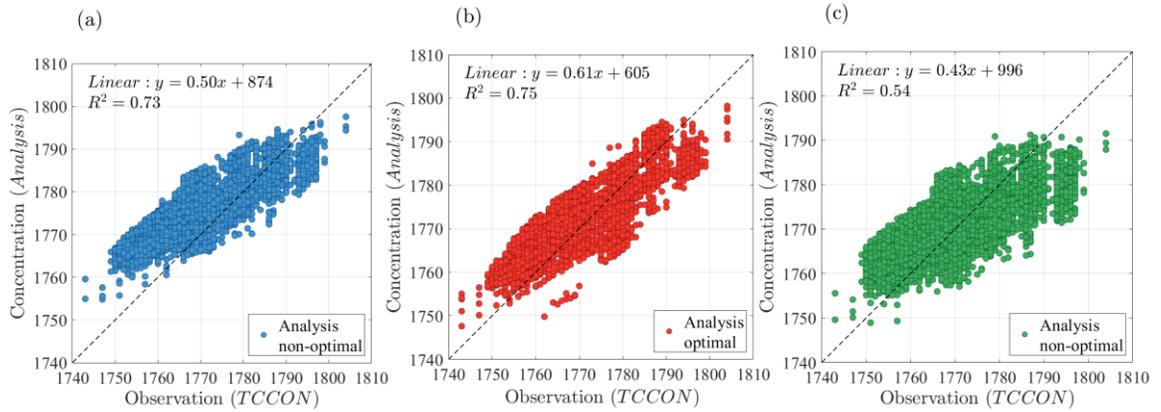


**Figure D.3.** Comparison of (a) model forecast (blue circles), (b) analysis with optimized parameters (red circles), and (c) analysis with non-optimized parameters (green circles) against independent HIPPO-3 observations.

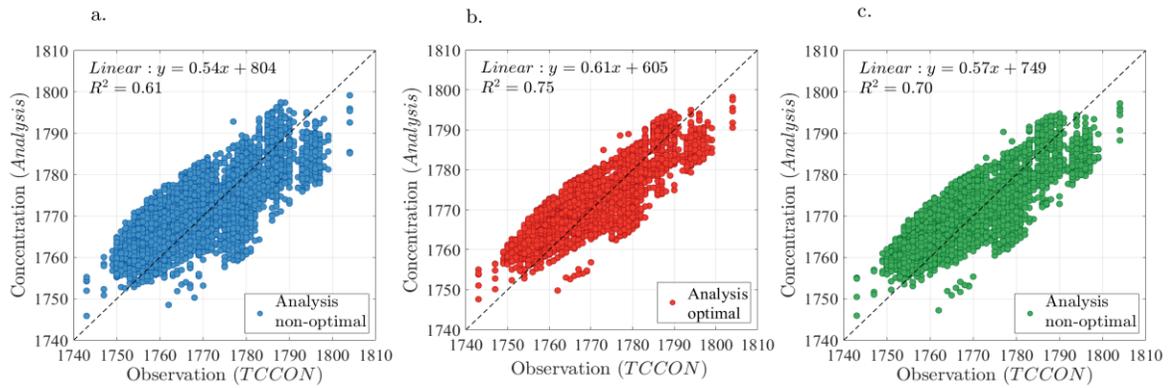
In Figure D.4-Figure D.6, the optimal analysis (red circles) with the error covariance parameters of  $f^o = 0.5, f^i = 0.45, f^q = 0.018, L_h = 350 \text{ km}, L_v = 7\sigma$  are compared with two non-optimal analyses, for which only a single covariance parameter is altered.



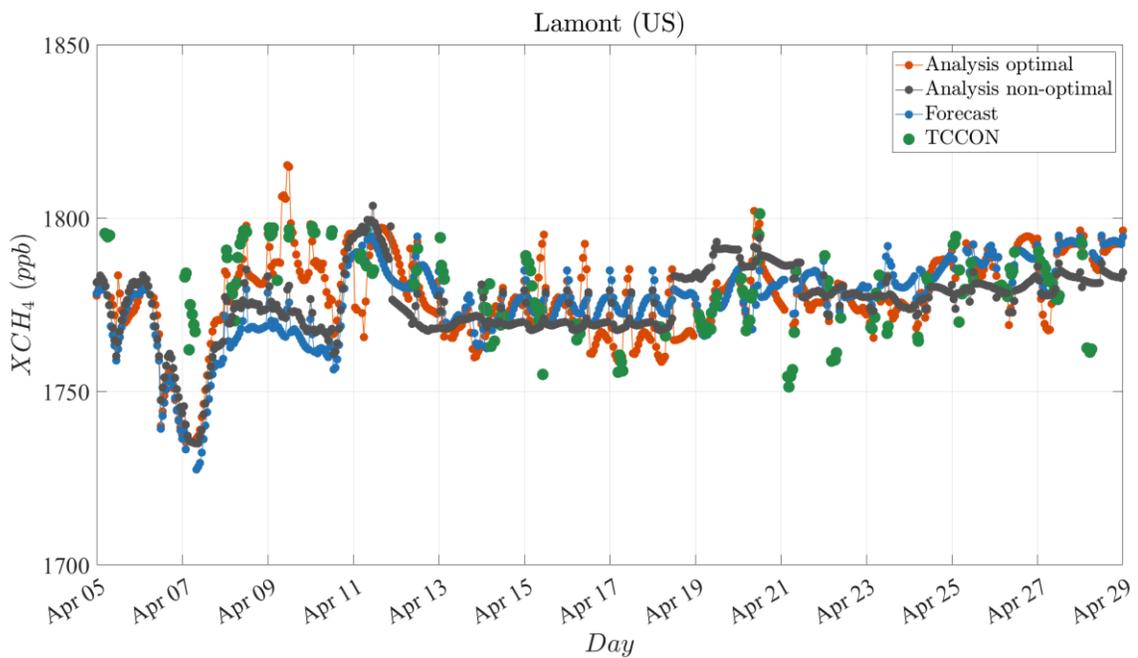
**Figure D.4.** Comparison of (a) non-optimal analysis due to only correlation lengths larger than the optimal value (e.g.,  $L_h = 600$  km,  $L_v = 10\sigma$ ), (b) optimal analysis with optimal correlation lengths (i.e.,  $L_h = 350$  km,  $L_v = 7\sigma$ ), and (c) non-optimal analysis due to only correlation lengths smaller than the optimal value (e.g.,  $L_h = 200$  km,  $L_v = 1\sigma$ ).



**Figure D.5.** Comparison of (a) non-optimal analysis due to only observation error smaller than the optimal value (e.g.,  $f^o = 0.2$ ), (b) optimal analysis with optimal observation error (i.e.,  $f^o = 0.5$ ), and (c) non-optimal analysis due to only observation error larger than the optimal value (e.g.,  $f^o = 1.5$ ).



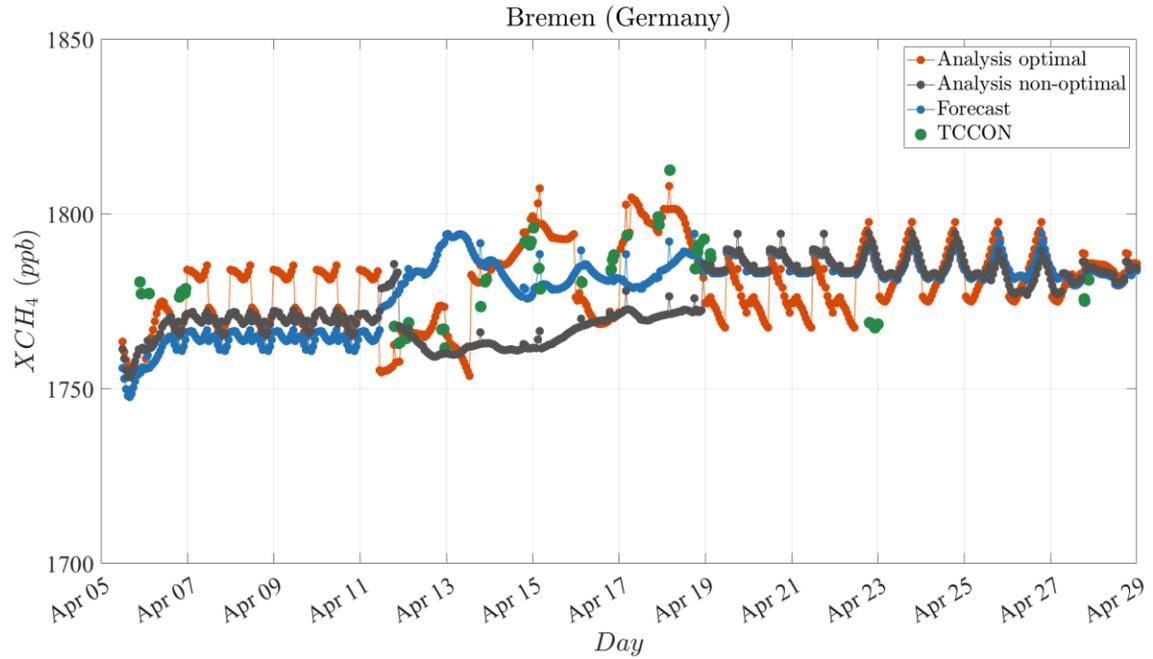
**Figure D.6.** Comparison of (a) non-optimal analysis due to only model error smaller than the optimal value (e.g.,  $f^q = 0.001$ ), (b) optimal analysis with optimal model error (i.e.,  $f^q = 0.018$ ), and (c) non-optimal analysis due to only model error larger than the optimal value (e.g.,  $f^q = 0.045$ ).



**Figure D.7.** Comparison of model forecast (blue dots), analysis with optimal parameters (red dots), and analysis with non-optimal parameters (black dots) against TCCON observations (green dots) at Lamont (36.60°N, 97.48°W). Optimal parameters of the analysis include

$f^o = 0.5, f^i = 0.45, f^q = 0.018, L_h = 350 \text{ km}, L_v = 7\sigma$  and the non-optimal parameters are assumed to be

$f^o = 1.2, f^i = 0.45, f^q = 0, L_h = 600 \text{ km}, L_v = 1\sigma$ .



**Figure D.8.** Comparison of model forecast (blue dots), analysis with optimal parameters (red dots), and analysis with non-optimal parameters (black dots) against TCCON observations (green dots) at Bremen (53.10°N, 8.85°E). Optimal parameters of the analysis include

$f^o = 0.5, f^i = 0.45, f^q = 0.018, L_h = 350 \text{ km}, L_v = 7\sigma$  and the non-optimal parameters are assumed to

be  $f^o = 1.2, f^i = 0.45, f^q = 0, L_h = 600 \text{ km}, L_v = 1\sigma$

## Appendix E1. OSSE Non-uniform Perturbation (type II) (Chapter 7)

Table E.1. Normalized mean bias (NMB), normalized mean error (NME), and Pearson's correlation coefficient (R) for variable perturbation (50% agriculture and wetland, 25% energy and waste of true emissions) and for each inversion cost functions (Equation (7.16)-(7.19))

Cost function Perturbation	Type 0:			Type 1:			Type 2:			Type 3:		
	$J_0(c_i^a, P_i^f(A_1, Q), R)$			$J_1(c_i^f, R)$			$J_2(c_i^a, R)$			$J_3(c_i^a, P_i^f(A_1), R)$		
	NMB	NME	R	NMB	NME	R	NMB	NME	R	NMB	NME	R
Case 7: All sectors/Non-uniform type II	-0.04	0.06	0.98	-0.10	0.35	0.87	-0.10	0.19	0.93	-0.04	0.07	0.96

total emissions perturbation (non-uniform , type II)

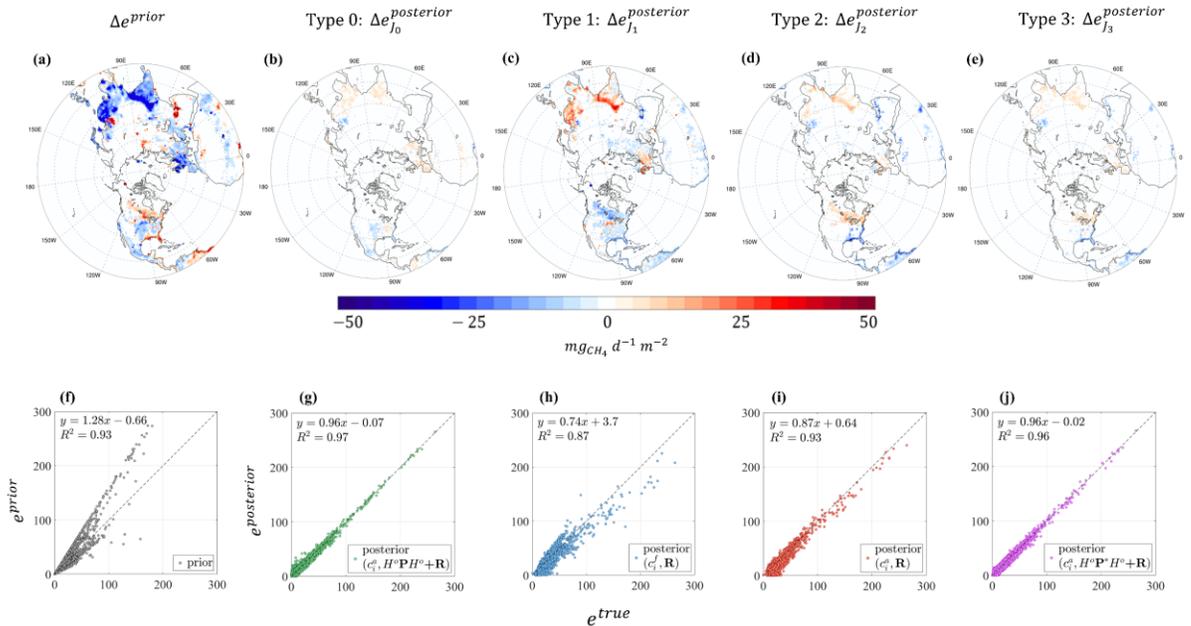


Figure E.1. (a) prior – true emissions ( $\pm 25$ -50% variable perturbation); (b) posterior – true emissions in Type 0 inversion using analysis initial ( $c_i^a$ ) and both observation R and model propagated analysis

error covariance  $H^o P_t^f(A_1, Q) H^{o T}$ ; (c) posterior – true emissions in Type 1 inversion using forecast initial ( $c_1^f$ ) and observation error covariance R, (d) posterior – true emissions in Type 2 inversion using analysis initial ( $c_1^a$ ) and observation error covariance R; (e) posterior – true emissions in Type 3 inversion using analysis initial ( $c_1^a$ ) and both observation R and model propagated analysis error covariance  $H^o P_t^f(A_1) H^{o T}$ , but with no model error. Statistical comparison of the (f) prior emissions and (g-j) posterior emissions of Type 0-3 inversion, respectively. x-axis and y-axis represent the true and prior/posterior emissions, respectively. In (f-j),  $P^f(A_1, Q)$  is shown as P, and  $P^f(A_1)$  is shown as P\*. Synthetic observations are generated using the nature run initialized by the analysis, and a 2-week spin-up is used for the initialization.

## Appendix E2. Determination of regularization parameter $\gamma$ (Chapter 7)

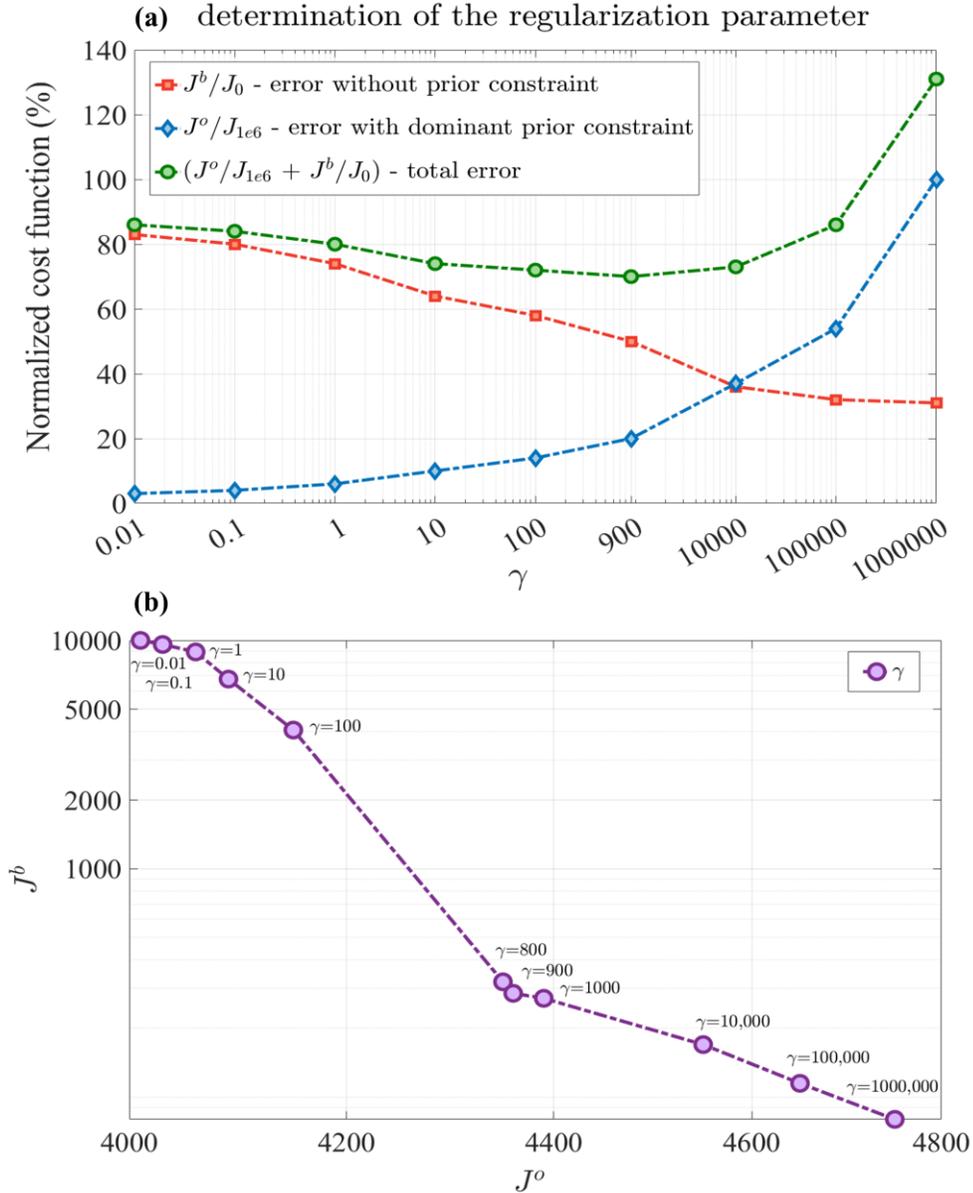


Figure E2. Given the cost function of Equation (7.15) in the main manuscript, the prior and observation

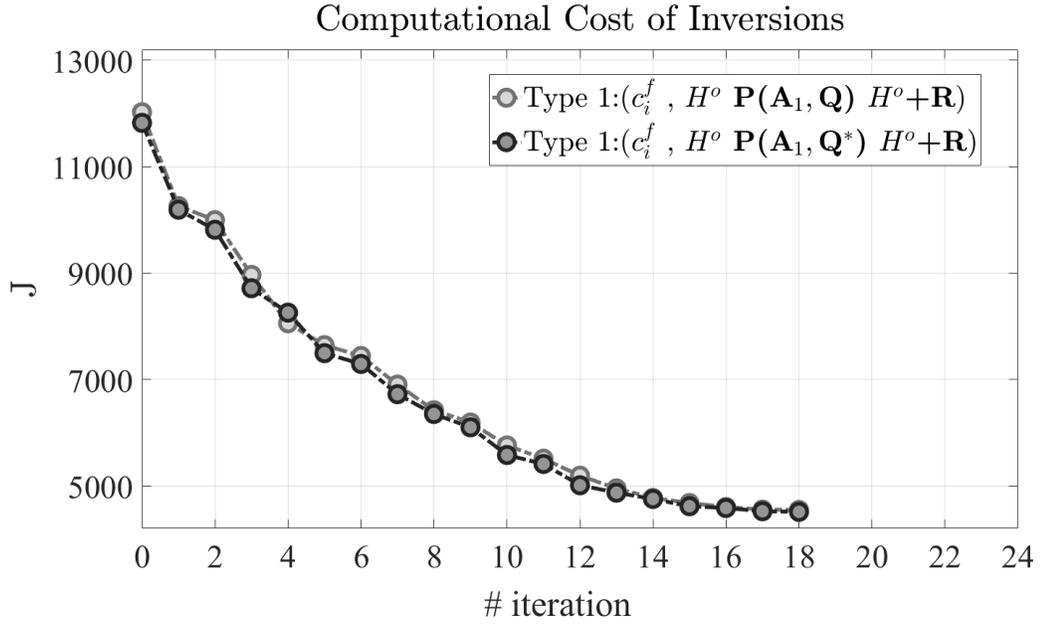
term of the cost function has the form  $J^b = \frac{1}{2}(x - x_b)^T \mathbf{B}^{-1}(x - x_b)$  and

$J^o = \sum_{i=0}^n \frac{1}{2}(y_i^o - H_i(c_1^a, x))^T (H^o \mathbf{P}_i^f(\mathbf{A}_1, \mathbf{Q})H^{oT} + \mathbf{R}_i)^{-1}(y_i^o - H_i(c_1^a, x))$ , respectively. (a) Show a traditional

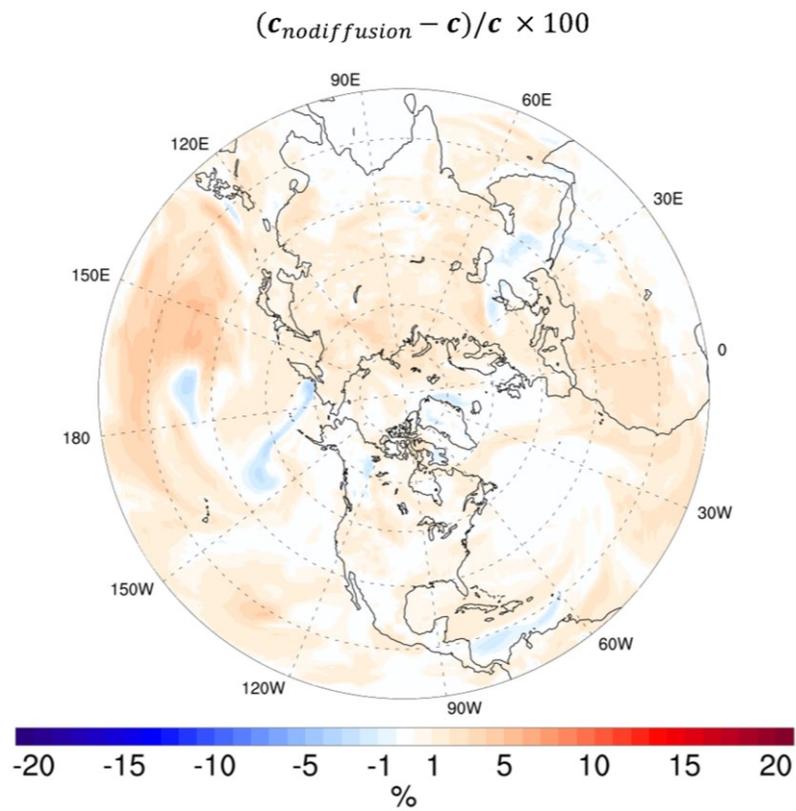
method of estimating  $\gamma$  that minimizes the sum of a normalized cost function [67].  $J_0$  is the magnitude

of the total cost function ( $J = \gamma J^b + J^o$ ) once  $\gamma = 0$ , indicating an optimization without prior constraint, and  $J_{1e6}$  is the magnitude of the total cost function at  $\gamma = 10^6$ , showing an optimization with a dominant prior constraint. In this method, we aim at a  $\gamma$  among a few selected values that minimize the total normalized error. It shows that  $\gamma = 900$  is the appropriate choice, although, for a wider range of this parameter (e.g., 500-2000), the choice of  $\gamma$  has little impact on the overall optimization (inversion) solution. (b) The L-curve method for the determination of the regularization parameter shows a comparison between the prior term of the cost function ( $J^b$ ) in the y-axis and the observation term of the cost function ( $J^o$ ) in the x-axis for different choices of  $\gamma$ . According to the method of Hansen (1999) [70],  $\gamma = 900$  is an optimal (balanced) choice for the regularization parameter. In principle, the optimal  $\gamma$  is obtained when the solution tends to change in nature from being dominated by the prior cost (or perturbation error, where a small variation of  $\gamma$  causes rapid changes in  $J^b$ ) to being dominated by the observation cost (or regularization/smoothing error where a large variation of  $\gamma$  makes a slow improvement in  $J^b$ ).

**Appendix E3. Diffusion effect of modelling transport error (Chapter 7)**



**Figure E3: Comparison between the computational cost of two inversions in which only model transport error is different in the cost function.  $Q^*$  is the updated form of model transport error ( $Q^* = Q + Q_{diffusion}$ ).  $Q$  is the estimated model error during the PvKF assimilation (or  $Q_{PvKF}$ ) and  $Q_{diffusion}$  is approximated model error due to neglecting propagation of error correlations by diffusion.**



**Figure E4: Normalized difference of concentrations between two cases where in the first one, the model diffusion scheme is deactivated, and in the second one, it is activated. It shows the distribution at the model's first layer after one month of simulation. Except for this difference, all other inputs and configurations between the two cases are the same.**